$\mathcal{I}f\mathcal{D}$ — Information for Discrimination

Di Cai

Department of Computing Science
University of Glasgow

A thesis submitted to the University of Glasgow
for the degree of Doctor of Philosophy
March 2004

To Ai

Abstract

The ability to automatically measure the power of discrimination of terms, or the amount of information in terms, is a fundamental issue in intelligent *information retrieval* (IR). It is widely acknowledged that the issue must be faced by almost all textual retrieval systems. The discrimination of terms has been a significant subject of interest among IR researchers since the early sixties. Since the publication of Van Rijsbergen's book in the late seventies it has moved into the mainstream of theory-oriented study and analysis. Many discrimination methods have been successively developed.

Nevertheless, there is no widely recognized formal definition of what should characterize term information. Typically, studies in related literature are accompanied by discussions of the circumstance in which the discrimination of terms is essential. Such discussions are argued by concrete examples and appeals to intuition, or by some empirical formulae. While these informal discussions might be sufficient to convey some of the ideas that discrimination encompasses, however, they are inadequate for any more formal analysis. In fact, the formal interpretation of term information for discrimination is not simple. This thesis introduces new techniques for defining term information as one, or more discrimination measure(s).

The problem of term mismatch and ambiguity has long been serious and outstanding in IR. The problem can result in the system formulating an incomplete and imprecise query representation, leading to a failure of retrieval. Many query reformulation methods have been proposed to address the problem. These methods employ term classes which are considered as related to individual query terms. They are hindered by the computational cost of term classification, and by the fact that the terms in some class are generally related to some specific query term belonging to the class rather than relevant to the context of the query.

In this thesis we propose a series of methods for *automatic query reformulation* (AQR). The methods constitute a formal model called $\mathcal{I}f\mathcal{D}$, standing for *Information for Discrimination*. In $\mathcal{I}f\mathcal{D}$, each discrimination measure is modelled as information contained in terms supporting one of two opposite hypotheses. The extent of association of terms with the query can thus be defined based directly on the discrimination. The strength of association of candidate terms with the query can then be computed, and good terms can be selected to enhance the query.

Justifications for $\mathcal{I}f\mathcal{D}$ are presented from several aspects: formal interpretations of information for discrimination are introduced to show its soundness; criteria are put forward to show its rationality; properties of discrimination measures are analysed to show its appropriateness; examples are examined to show its usability; extension is discussed to show its potential; implementation is described to show its feasibility; comparisons with other methods are made to show its flexibility; improvements in retrieval performance are exhibited to show its powerful capability. Our conclusion is that the advantage and promise of $\mathcal{I}f\mathcal{D}$ should make it an indispensable methodology for AQR, which we believe can be an effective technique for improvement in retrieval performance.

Declaration of Authorship

This thesis was composed by myself. The original research described herein is my own. The work has not been submitted in consideration for any other degree.

Di Cai

Acknowledgements

I am first indebted to my supervisor, Keith van Rijsbergen, for his support and guidance throughout my PhD student life. Keith's book [207] influenced my thoughts on many aspects when I was writing this thesis. I am very grateful for his knowledgeable and meticulous scrutiny, and his sound judgement.

Many thanks are extended to my second supervisor, Joemon M. Jose, for his encouragement.

The enthusiastic interest of my colleagues, in particular, Iadh Ounis and Ian Ruthven, has been a perpetual source of encouragement. To all of them I am most grateful.

This thesis was written when I was a PhD student at the University of Glasgow, Department of Computing Science. The preparation of this thesis was made possible by the joint support of EPSRC and the Department while I was a PhD student. I am grateful to Ray C. Welland, Head of the Department during the period of my research, for his encouragement and support of my study during this time.

Drafts of the thesis were reviewed by Paul C. Philbrow. His feedback, and constructive criticism, contributed to the improvement of the manuscript. I offer my sincerest thanks to him. My sincere thanks to Juliet van Rijsbergen for her comments on earlier drafts of the thesis.

My parents have generously helped in many ways. Without their help my thesis would not have been possible. I can never thank them sufficiently.

Finally and foremost, I am deeply grateful to my husband for his love and support. I am deeply grateful to my little daughter for her love, her bright eyes, her brilliant smiles and her understanding well beyond her age.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Information retrieval (IR) is concerned with the processes involved in the representation, storage, searching and finding of information relevant to a query for information desired by a human user [89]. The study of information retrieval is the study of the optimal relationship between the input and output of the information retrieval system [115].

The intention of an IR system is to identify latent useful information in response to user information needs. The objective of a textual IR system is to retrieve all relevant documents, and at the same time to retrieve as few of the non-relevant ones as possible with respect to a user query. Many important issues of IR have been studied for this primary objective, for example, [167, 207]. Some good formal models have been developed, for instance, [5, 60, 62, 110, 136, 152, 167, 178, 201, 206, 207, 208, 209, 221, 223, 228].

Generally, textual retrieval strategies depend mainly on (1) document representation; (2) query representation; and (3) decision function. These three central issues are very model dependent. They are briefly described in the following three sections.

1.1 Document Representation

In a textual information retrieval system, the *objects* we deal with are documents and queries. The system does not deal directly with the objects themselves but their surrogates. In order to design a formalism having predictive capability of relevance, we first need to know the explicit representations of the objects. That is, we have to design a reasonable scheme to generate the surrogate of each object. Thus, the document representation is the first central issue for the development of a quantitative textual retrieval model.

In information retrieval, each document is characterized by a set of *concepts*. Hitherto, the simplest way of describing each concept involved in a document is to use *index terms* that appear in the document. Usually, a single index term might contain one piece of information. However, there exist complex semantic relationships between index terms. Each document may therefore contain large amounts of information. No assumptions are made regarding the structure of the information, although in practice structured subdivisions of documents may be accommodated.

Generally, there exists a weight transformation, called a *document term weight function*, which maps each index term to a numerical quantity related to a given document. The

quantity, known as the *weight* of index term, is considered to 'indicate' the importance of the index term concerning the document.

Thus, each document can be approximately represented by means of the corresponding weights of a set of index terms, which is usually referred to as a *representation* of the document (we will return to this topic in Section 3.1). With such a representation, the relationships between documents, and between documents and the query, becomes transparent when dealing with a specific quantitative retrieval model.

In an ideal retrieval environment, the document representation would be independent of individual retrieval models. It is desirable that there exists a unified formalism which can effectively represent documents. However, a feasible scheme for accurately computing the importance of terms is not available. The document representation has to depend on a specific model itself, and it is frequently consistent with the statistical nature of the indexing procedure.

It should be pointed out that to arrive at a precise representation of a document by means of weights of a set of index terms is a difficult task. This is because it is very hard to obtain sufficient statistical data for the estimation of the importance of index terms (we will return to this topic in Section 3.7). It is also very hard to explicate the complicated semantic relations between index terms. In information retrieval, the problem of how to properly represent documents has not been satisfactorily resolved.

1.2 Query Representation

In like manner, each query is characterized by a set of concepts, and index terms are used to describe the concepts involved in the query. In information retrieval, query representation is the second central issue. It is one of the main obstacles to be faced in developing an effective quantitative retrieval system. In order to make a thorough investigation into query representation, we need to define some important notions: query formulation, term mismatch, term ambiguity and query reformulation.

1.2.1 Query Formulation

An original query is a description of the information need typically expressed in natural language. The process of the original query description is complex and depends on particular attributes of the user, such as his knowledge of the contents of the database, of the indexing and searching procedure of the system, his familiarity with the topic matter being searched, his personal preferences as to vocabulary and style, and so on. Indeed, the original query can hardly include all the aspects of the need [198]; the probability of the user being able to describe a query which will retrieve all of the documents satisfying his information need is very small [158].

In a quantitative textual retrieval system, *query formulation* is a process whereby the original query is initially transformed into a numerical representation. The weight transformation is called a *query term weight function*; the quantity representation is in this thesis called an *original query representation*. In practice, the original queries are usually inadequate, imprecise, or incomplete descriptions of users' information needs, and a retrieval system cannot be expected to produce ideal retrieval results by using the original query representation.

1.2.2 Term Mismatch and Term Ambiguity

In a practical IR environment, extremely large collections are routinely processed, and documents that match the query are displayed in real time. The retrieval efficiency is attributable to the use of conventional inverted index file technology, and documents that do not have any term matching the query (i.e., that do not have at least one term in common with the query) are immediately discarded. Users of an IR system that employs term matching as a basis for retrieval are faced with the challenge of describing their queries with terms in the vocabulary of the documents they wish to retrieve. This difficulty is especially severe in extremely large, full-text databases containing many different term descriptions of the same concept [214].

Term mismatch is a phenomenon whereby the terms used to describe one concept characterizing a query are different from the terms used to describe the same concept that characterizes documents. *Term ambiguity* is a phenomenon in which the terms used to describe one concept characterizing a query can also be used to describe other concept(s) that characterize(s) documents. These two notions refer to two opposite aspects. Speaking popularly, term mismatch addresses the problem that many terms (such as, *synonymous* terms) can be used to describe one concept; whereas term ambiguity talks about the problem that one term (such as, *polysemous* term) is used to describe several concepts. In information retrieval, the problems of term mismatch and term ambiguity have long been serious and outstanding.

Some research, [63] for instance, has shown that people use a surprisingly large variety of terms to refer to the same thing in everyday life. The probability of two people - author and user - choosing identical terms is less than 0.20. A retrieval system may only be able to make a successful retrieval if users enter terms coinciding with the ones assigned to documents they desire. This means that users may fail to retrieve documents on 80 to 90 percent of their attempts.

We may consider the problems of term mismatch and term ambiguity from the several points of view given below.

Linguistic Aspect

- Semantic relationships between terms are a type of term mismatch.
  - ➟ If a user describes his information need as *Aviation School*, then relevant information indexed by terms *Aeronautical Engineering Institute* might meet with retrieval failure. Term mismatch arises from the synonymous terms.
  - ➟ A user types in a term *cow*, he might really be interested in information about *mammal*. Term mismatch arises from the speciality of terms.
  - ➟ A user enters a term *planet*, he might be thinking of something like *Mercury*, or *Venus*. Term mismatch arises from the generality of terms.
  - ➟ A user tries terms *crime*, *violate* and *murder* when she desires to find some thrillers. Term mismatch arises from the related terms.
- Two different terms referring to the same thing but used in different (specific) situations are a type of term mismatch.
  - ➟ Different authors have different vocabulary.

DNA profile is frequently used by journalists reporting crime events such as murder, whereas *DNA sequence* is usually used by the authors of scientific work in biology and medicine.

➟ Authors and users have different vocabulary.

If a user describes her query with a term *shingles*, then a relevant document indexed by terms *herpes zoster* (used by medical authors) might not be retrieved.

➟ The difference between British English (BrE) and American English (AmE).

BrE	AmE
autumn	*fall*
axe	*ax*
judgement	*judgment*
labour	*labor*
laptop	*notebook*
pavement	*sidewalk*

➟ Abbreviation (Abbr.).

Abbr.	Instead of
AST	*Atlantic Standard Time*
maths	*mathematics*
MS-DOS	*Microsoft Disk Operating System*
MSG	*Monosodium Glutamate*
U.K.	*United Kingdom*
TV	*Television*

❑ Morphological variation, that is, the structure and form of terms (including inflection, derivation, and the formation of compounds), is a type of term mismatch.

➟ In a retrieval based on term *sun*, the user might be also interested in the documents containing term *sunspot*, or term *sundog*. Also, in a retrieval based on term *constellation*, the user is usually interested in the documents indexed by terms *constellations* and *constellatory*.

❑ Polysemous terms (multiple-meaning terms) constitute a type of term ambiguity.

➟ In a retrieval based on term *phoenix*, a system can only guess whether to return documents related to

'the capital and largest city of Arizona', or

'a bird in Egyptian mythology', or

'a constellation in the Southern Hemisphere near Tucana and Sculptor'.

User Aspect

❑ Whether a retrieval is successful or not depends on the quality of the original query. A high quality of the query should consist of proper terms that can precisely describe the

concepts involved in the information need, and also should cover all aspects of the need. However, it is very hard for users to form high quality queries owing to their limited perception of the information they desire and limited understanding of the retrieval system. The use of improper terms and incomplete coverage of the subject of interest can produce an unsatisfactory output outcome, even failure.

- Domain knowledge is particularly important for users to be able to choose proper terms for describing concepts involved in his query. Unfortunately, that knowledge is exactly what users seek. For example, in a library search, a student may seek information on a concept *vector*. However, he does not understand that the concept is usually defined and explained in linear algebra. A query described by a term *vector* would not retrieve some relevant books like, for instance, '*Introduction to Linear Algebra*'.

- The problem is more pronounced for a short query consisting of just a few terms related to the subject of interest: it can best be illustrated through the scenario of information search on the Web where users' queries are usually very short [91]. As the query becomes shorter, there may be less chance for some important terms to co-occur in both relevant documents the query. It is hard for a short query itself to produce reliable predications for the information needs. With the advent of the Internet, short queries have become increasingly more common.

- It is likely that a query description for the same information need would vary from user to user, and, contrariwise, the information sought would vary with the user's different perspectives even though the same query is described.

System Aspect

- An information retrieval system is usually complex, and an information resource can be extremely large and wide-ranging. Moreover, the processes of retrieval, in most systems, are not transparent to users. It can therefore be very difficult for users to precisely describe queries according to their information needs.

- In order to achieve an effective search, retrieval systems should be able to recognize terms that are tried spontaneously by users. This means that the system vocabulary should be considerable. The problems of term mismatch and term ambiguity therefore become conspicuous and severe.

- It is also questionable whether retrieval systems can effectively formulate queries. This is because representing queries is as difficult as representing documents.

1.2.3 Query Reformulation

One motivation for the studies presented in this thesis is to meet the challenge of the increasing demand for high precision from realistic retrieval systems. In particular, with the increasing use of the Internet, the user tends to view only the top few documents retrieved. A retrieval system with high precision returns would therefore be more desirable than one with a high recall but low precision.

Because terms used to describe concepts involved in the query are frequently not the same as terms used to describe the concepts involved in the documents, because the inherent nature

of polysemous terms, because the number of terms tried spontaneously by users for describing their information needs may be very few, and so forth, the original query representation usually lacks adequate, precise and complete information to match with the representations of the potentially relevant documents. Given such a scenario, a retrieval system can never be sure it has correctly inferred the user's referent, and thus cannot be expected to accurately distinguish relevant documents from non-relevant ones. As a result, a retrieval may achieve low precision even if it achieves high recall.

We need good ways of matching and disambiguating terms. *Query reformulation* is a process that revises the original query representation by strengthening, or intensifying, some concepts so as to more precisely describe the information need. It produces an *enhanced (modified, refined) query representation.*

The most common method for query reformulation is the technique of *query expansion.* In particular, when expansion terms are drawn from a sample set of relevant documents, query expansion can be thought of as a technique that adds terms describing the concepts involved in the relevant documents into the original query which describes the same concepts involved in the query. The technique of query expansion counteracts the problem of term mismatch, whereas choice of relevant sample documents counteracts the problem of term ambiguity.

Query expansion is usually considered to be a recall-enhancing device. This is because all documents retrieved against the original query are also retrieved against the expanded query; some new documents against the expanded query are also retrieved when they contain the expansion terms. Precision, however, may decline if the expansion results in non-relevant documents being ranked above relevant ones. Past experimental investigations, reviewed in [74, 81, 164], showed that query expansion, while it improves recall, may reduce precision. Ways of effectively improving retrieval performance by query expansion, in particular, focusing on improving precision of the top-ranked documents, have been extensively studied [27, 29, 82, 96, 130].

In this thesis we propose a series of formal methods for automatic query expansion. In the presentation of the proposed methods, we will use 'automatic query reformulation (AQR)' and 'automatic query expansion (AQE)', interchangeably.

1.3 Decision Function

We have described two central IR issues — the representation of documents and of queries. The third central issue involves relevance classification. The criterion for classifying documents into the different relevance classes with respect to a given query is called a *relevance decision function* (or *similarity measure*). The decision function determines the degree to which a document is relevant to a query, that is, it is a mathematical method for predicting relevance. To be successful, the classification should be performed in such a way that the resultant prediction and the actual outcome are, on the average, in close agreement. In this thesis, the dichotomous relevance classification (relevant or, non-relevant) will be adopted.

Textual retrieval methods may be divided into exact and partial match methods [8]. The former are usually Boolean retrieval methods, the latter consist of several methods, of which the most prevalent are perhaps the linear algebraic retrieval methods and probabilistic retrieval methods.

Boolean retrieval models are perhaps the best understood by virtue of their simplicity and sound theoretical basis. In Boolean retrieval models, the relevance is interpreted as logical implication. A document is considered to be relevant to a query only when it logically implies the query. In this case, documents either exactly match the query or not, and all retrieved documents are treated to have equal relevance. Such an interpretation might be inappropriate for information retrieval. In fact, relevance often cannot be determined by strict inference (logical implication). Some disadvantages have been pointed out by some researchers [8, 101, 221]: it may miss some relevant documents whose representations match the query representation partially; it cannot take into account the importance of terms concerning a document or query; it cannot provide ranked output as all documents are considered equally important; it is apt to retrieve either too many or too few documents; it requires a complicated logical formulation of the query.

In contrast, weighting terms is a main feature of partial match retrieval methods. In partial match methods, all those documents that contain at least one query term will be retrieved and ranked according to their presumed relevance. The degree of the relevance may be calculated based on weights of document terms, and possibly also on weights of query terms. The weights are usually derived by using the statistical data available, such as, the frequencies of terms within the individual documents and the document frequencies of terms concerning the collection as a whole [162, 185]. Thus, it is clear that the partial match methods take into consideration of the importance of terms concerning individual documents and the query. Also, the query can usually be described in natural language or even as a set of terms. Thus, the partial match methods may remedy the problems of Boolean retrieval methods [101]. The evidence accumulated so far indicates that the use of term weighting provides an effective means to improve retrieval performance [164, 167, 207, 224].

In the linear algebraic retrieval model (also called the vector space model) [161, 167], the query can be directly formulated by system or user. Both documents and queries are represented as n'-dimensional (numerical) vectors in a concept space spanned by a chosen set of orthonormal base vectors (where $n' \leq n = |V|$ is the number of concepts involved in the collection). Thus, the decision function is simply the *scalar product* between a document vector and a query vector. If the document vector and the query vector are normalized, then the scalar product is the usual cosine similarity measure. A critical analysis of this model was given in study [142], and the study pointed out that one of the main problems was the assumption of term pairwise orthogonality. Some attempts [227, 228] were made to remove such a strict assumption, and a formal method, called the generalized vector space model, for computing term correlation was proposed. However, the establishment of the set of orthonormal base vectors remains an open problem.

In the conventional probabilistic retrieval model [42, 152, 206, 207, 235], the query is not directly formulated by system or user. Instead, the decision function, which represents the query, is derived by the system automatically through a relevance feedback procedure. It is well known that if an assumption of pairwise probabilistic independence among terms is made, then a linear decision function can immediately obtained. If pairwise probabilistic dependence is taken into account among terms nevertheless, then the decision function becomes quadratic in the components of document vector.

In this thesis, relevance and usefulness are not treated as equivalent concepts. The degree of relevance is estimated by the decision function designed into the system, and the estimation is objective. The choice of the decision function is essential for an effective retrieval. The notion of relevance in IR has been studied by many researchers, such as, [12, 35, 79, 173, 174].

In contrast, the degree of usefulness is subjectively assessed by the user. A document may or may not be useful to a given query depending on many factors. If a document is assessed by the user to be of interest, it is useful; it is useless otherwise. Since many factors (such as, user's knowledge, search intention, etc.) determine the interest in a complex way, it is unlikely that the system can precisely retrieve only and all useful documents for the user. Instead of this, the system normally adopts some decision function that facilitate the ranking of documents in the order of their estimated relevance to the query.

1.4 About This Thesis

Discuss **judgement** methods of 'good' terms.
Are any *judgement* methods more fundamental than **association** functions?
Are any *association* functions more fundamental than **discrimination** measures?
Should not the fundamental **theory** be about these more fundamental *discrimination* measures?

Term information for discrimination and its application to AQE are central themes of this thesis. The main objective of this thesis is to establish a formal model for studying effective methods of AQE. We call this formal model $\mathcal{I}f\mathcal{D}$, *Information for Discrimination.*

It has not been easy to interpret the meaning of the amount of information contained in a given term rationally and explicitly within the context of IR. It has not been simple to introduce the technique of query reformulation meaningfully and successfully into scientific discussions. This thesis is an attempt to do this.

Before we can talk about this thesis in more detail, let us first look at some examples and think about the corresponding questions below.

1.4.1 Examples and Questions

Example 1.4.1 A query contains terms *power plant.* Because term *plant* has multiple meanings, a query expansion technique may add terms *garden*, *tree*, *vegetable*; or add terms *amphibious plant*, *herbaceous plant*, *monoecious plant*; or add terms *cement plant*, *chemical plant*, *milling plant*, and so on.

Question-1: How can we avoid expanding with terms related to incorrect meanings of query terms?

Example 1.4.2 A query contains terms *gun control and crime.* A good query expansion technique should add terms *blood*, *dead*, *kidnap*, *murder* and *robbery*, rather than terms like *bribe*, *fraud*, *steal* and *tax evasion.* All these terms are obviously narrower terms for term *crime.*

Question-2: How can we expand terms which will precisely describe the content of the query?

Example 1.4.3 A query contains a phrase *South Africa*. It seems that it does not make sense to treat the phrase as two terms and expand them independently. The argument is that some of terms, such as, *South Pole*, *the South Seas*, *the South Downs*, *the south of Europe*, *the hot south wind*, and so on, would be incorrectly added to the query.

Question-3: How can we deal with phrases in the query in an appropriate way?

Example 1.4.4 A query is: *What is tomorrow's computer?* It seems that term *computer* would not provide any discrimination information for the relevance classification for a collection catalogued as computing science. Yet term *computer* is central to the query.

Question-4: Why does a term central to the query not provide any discrimination information about relevance?

Example 1.4.5 A query contains terms *gun control and crime*. Terms, such as, *kidnap*, *murder*, *robbery*, *bribe*, *fraud*, *steal*, may appear in some top-ranked documents. Obviously, these terms are more or less statistically associated with some of query terms.

Question-5: Which of these terms should be strongly associated with the context of the query?

Example 1.4.6 A query is: *DNA testing in trials of criminal cases*. Two phrases *DNA profile* and *DNA sequence* are nearly synonymous (both are related to query term *DNA*). Interestingly, *DNA profile* is frequently used in relevant documents, but *DNA sequence* is not. This is because journalists reporting crime events, such as murder cases, tend to use phrase *DNA profile*, while the authors of scientific work in biology and medicine tend to use phrase *DNA sequence*.

Question-6: How can we explain such a phenomenon with the Association Hypothesis given in [207]?

1.4.2 Main Ideas of the Thesis

A fundamental issue in any kind of textual retrieval model is how we can measure the power of discrimination of terms, or the amount of information in terms. A central subject in any method to query expansion is how we can judge whether a term is a good one with respect to a given query. In order to achieve an effective measure and judgement, this thesis is devoted to a theory-oriented study and analysis. This has three aspects: (a) the measurement of the discrimination information of terms; (b) the definition of association of terms with a query; (c) the construction of an association score function. The study and analysis applies basic concepts of information measures introduced in information theory and is supported by a retrieval system, $\mathcal{I}f\mathcal{D}$, developed in this thesis. The mathematical interpretations of the basic concepts are fully centred around the three aspects. We now survey the main ideas of the study, and analysis and elaborate them in the following chapters.

Discrimination Information of Terms

In this thesis, we view the measure of term information as a fundamental issue of IR. This is because the knowledge concerning the amount of information contained in a term, or the

amount of mutual information contained in a term pair, can be of great benefit. By using the knowledge, we can measure the importance of a term and decide whether to use it to index a document; we can measure the extent of the association of a term with a query and decide whether to use it to expand the query; a retrieval system can measure the degree of the relevance of documents to a query, and determines which documents would be of interest to the user; the performance of a retrieval system may be examined by measuring the information contained in documents retrieved; and so forth. Acquisition of the knowledge is therefore at the very core of central issues of IR.

The concept of information is too broad to be captured completely by a single definition. It is hence very difficult to measure term information by only one mathematical formula. In this thesis, we concentrate on the study of discrimination information of terms. *Discrimination information* of a term is used in this thesis to refer to the amount of information contained in the term in favour of one of two opposite hypotheses. In particular, when the hypotheses involve relevance, discrimination information refers to the amount of information conveyed by the term for distinguishing relevant documents from non-relevant ones. A series of discrimination measures are discussed, interpreted, and analysed. The discrimination measures are designed based on five basic concepts borrowed in information theory: *directed divergence* (also called *information measure*), *divergence*, *information radius*, *Jensen difference* (also called *entropy increase*), and *expected mutual information*. The discrimination measures form a basis for formal methods proposed in this thesis for query expansion.

How can we judge all potential good terms for query expansion? In point of fact, we cannot. However, if we can predict the expected amount of discrimination information then we will be in a strong position to judge good terms. Thus we want to predict the expected amount. The five basic concepts of information theory provide powerful tools to estimate the expected amount. We can hence measure the extents of the contributions made by individual terms to the expected amounts. The formulae used to measure the extents are called the *discrimination measures*.

We point out that the discrimination measures may or may not be query-related. Thus, a term may be very informative, i.e., it may possess higher power of discrimination, but it may not be associated with a given query. For instance, term *Himalayas* would be very informative, however, it should not be associated with the query given in Example 1.4.6. Conversely, a term may be strongly associated with some query, but not be informative, for instance, term *computer* in Example 1.4.4. These interesting phenomena will be explained in this thesis.

Association of Terms with the Context of a Query

The concept of the association is also hard to capture in only one mathematical formula. In this thesis, we focus on investigating the association of terms with a given query. Thus, each association function is defined as query-related, or more precisely, query-context-related. It should refer to the statistical association of terms with all query terms that appear in the relevant sample documents. In this way, we may effectively avoid the increase of 'query ambiguity', caused by the ambiguity of individual query terms, because it combinatorially considers all possible information contained in the query. Particularly, when expansion terms are drawn from the relevant sample documents, these terms can provide sufficient context to clear up confusion, and may have the potential power of discrimination on relevance with respect to the query.

We especially emphasize that a term strongly associated with the context of the query

is a necessary condition for the term to be a good one. Some experimental evidence has shown that expanding each query term independently and ignoring the specific context of the query completely might cause a problem that expansion terms would be related to the incorrect meanings of the query terms [123, 230]. For instance, in Example 1.4.1, if we expand query terms *power plant* independently, then some terms, such as, *garden*, *tree*, *amphibious*, *herbaceous*, *cement*, *milling*, would be added to the query, which is not desirable. Similarly, in Example 1.4.3, if the query contains a phrase *South Africa*, then terms, such as *pole*, *sea*, *europe*, *east*, *hot* and *wind*, are likely to be added to the query, which is also not desirable.

Notice that, the methods proposed in this thesis are able to accommodate the polysemous term problem. This is because, for a given term, our methods take a comprehensive consideration of association of the term with the query based on the statistical data of the co-occurrence of the term with all query terms that appear in the relevant sample documents. Consequently, when a term, such as *garden* in Example 1.4.1, is associated with query term *plant*, it would be weakly associated with query term *power* (and perhaps also weakly associated with other query terms). Thus, the (total) association of term *garden* with the query would be rather weak. In the end, term *garden* would be eliminated by the query expansion procedure. In other words, term *power* may help avoid selecting term *garden* for expansion. The consideration of the context of the query can effectively prevent some undesirable matches.

Common terms (except stop words) drawn from the relevant sample documents tend to co-occur more frequently with most query terms than uncommon ones. Thus, the common terms would have a higher chance to be selected as expansion terms. Some studies, [20] for instance, show that adding common terms from the (relevant) sample documents would achieve significant improvements in retrieval performance. Therefore, we may find good terms among the common terms drawn from the relevant sample documents.

Correlation among terms has long been an important issue, and many IR researchers have objected to the methods of automatic classification of terms on the grounds that the correlations among terms are ignored. The derivation of term correlations may be achieved by taking into account statistical correlation information on which some statistical methods of successful term classification depend. However, it should be pointed out that the concept of association of terms with the query, in this thesis, is not the same as the concept of correlation of terms. The concept of association requires terms to have both high power of discrimination and strong correlation with all query terms that appear in the relevant sample documents.

We will see that all the definitions of the association functions are given in this thesis in more formal forms. Each of the association functions depends only on its three arguments: two probability distributions related to two opposite hypotheses and the original query representation.

Association Score Functions

In this thesis, the set of candidate terms for query expansion consist of all (relevance or pseudo-relevance) feedback terms. In order to select potentially good terms with respect to the query, an association score function is designed for assigning a score to each feedback terms and then sorts them for comparison and selection.

The construction of a score function is rather intuitive and simple: it is based entirely on a certain association function. In fact, in our formal methods, the association functions are abstract forms of the score functions, whereas the score functions are specializations of the

abstract forms. In other words, each score function is an embodiment of some association function by providing particular mathematical expressions of the three association function arguments. Once these three expressions are given, the score function is uniquely determined.

Thus, from our point of view, the issue of the construction of the score function is the issue of estimating the three arguments. Evidently, the estimations are crucial for effectively distinguishing the potentially good terms from many others. In this thesis, the estimations are treated as an important subject, and detailedly discussed and carefully analysed from general to specific. Particularly, a unified framework is established to support a systematic investigation into effective estimation.

Obviously, the different association functions and a variety of ways of estimating the three arguments would generate a number of score functions. However, it is remarkable that all the score functions given in this thesis involve only three essential factors: (i) the significance of a term concerning the query; (ii) the importance of a term concerning the relevant sample set; (iii) the discrimination information of a term concerning two opposite hypotheses.

Experimentation

IR experimentation should illuminate and help to develop theories and models which, in turn, should guide the design of good systems [151]. Thus, after some formal discussions, we investigate to what extent each score function contributes to improvement in retrieval performance. We evaluate the average retrieval performances of the expanded queries obtained from our methods, and compare the performances with that of the original queries without query expansion, and with that of the expanded queries obtained from the Rocchio formula [158]. In addition, we propose a new reweighting function for weighting the expanded query terms, which emphasizes both importance of query terms and association of expansion terms with the context of the query. We experimentally demonstrate:

- Our methods are both precision-enhancing and recall-enhancing devices, particularly, their use can greatly increase precision at-5 and at-10, which indicates that they are effective in improving retrieval performance.

- Our score functions are more suitable for shorter queries on relevance feedback, whereas they are more effective for longer queries on pseudo-relevance feedback.

- treating the discrimination information of terms as an important factor in weighting expanded query terms may help increase retrieval performance.

- Weighting expanded query terms using our reweighting function works better than using the Rocchio formula on pseudo-relevance feedback, and significantly better than using the Rocchio formula on relevance feedback;

- Our query expansion methods are insensitive to the size of the sample set and to the number of expansion terms.

1.4.3 Outline of the Thesis

To make it coherent and clear how the parts of this thesis are related, we will focus especially on the fundamental issue of the discrimination information of terms and its applications to

automatic query expansion. The elaborate discussions on the strategies of document representations and the designs of decision functions can be found elsewhere in, for instance, the monographs by Van Rijsbergen [207] and Salton & McGill [167]. Thus, the organization of this thesis is outlined as follows.

In Chapter 2, we review the most popular statistical methods to query expansion proposed in the past. We want to show how a query can be enhanced and then be used to improve the retrieval performance effectively. We also discuss several existing problems of query expansion, which are closely related to the studies described in this thesis.

In Chapter 3, we study the application of the concept of directed divergence for AQE. The rationality of using a logarithmic measure of information to measure the amount of information contained in a given term is interpreted. The estimations of the term probability distributions are elaborated.

In Chapter 4, we concentrate on the application based on the concept of divergence for AQE. Because the condition of absolute continuity between two distributions may not be satisfied in a practical context of IR, this chapter is devoted to a formal analysis and mathematical discussion on the feasibility of applying divergence to AQE.

In Chapter 5, we intend to give an easily understood account of the concept of information radius and its application for AQE. Some interesting properties of the discrimination measure are discussed.

In Chapter 6, we concern ourselves with the application of the concept of entropy increase for AQE by introducing a more general concept, the Jensen difference. Three typical entropy functions are considered, and the appropriateness of applying them as a divergence measure is investigated.

In Chapter 7, we focus on the application of the concept of expected mutual information for AQE. The notion of amount of mutual information contained in a term pair is interpreted. Some important properties of the discrimination measures are discussed. The estimations of the term state distributions is studied, and a general framework for the estimations is established.

In Chapter 8, we present a series of experimental results of applying the formal methods proposed in this thesis to AQE. A function for reweighting terms of expanded queries is introduced. The retrieval performances are compared and analysed.

In Chapter 9, we summarize the contributions of studies given in this thesis, and discuss some points that are worth further investigation in the future.

Finally, in Chapter 10, we deal with some mathematical details, which are mentioned in the earlier chapters.

The formal discussions given from Chapters 3 to 7 are inevitably mathematical in tone. Readers less interested in mathematical details should still be able to follow the analyses and, particularly, the descriptions of the practical discrimination procedures we propose.

Chapter 2

Historical Review

Query reformulation, an important component in a retrieval system, has long been an effective technique to improve retrieval performance [27, 29, 47, 54, 72, 111, 130, 158, 165, 210, 231]. Before detailing our formal model $\mathcal{IfD}$, we review some methods of query reformulation that have appeared in the literature. Other good reviews of query reformulation methods can be found in [29, 53, 65].

There is a large literature on query reformulation, and we will not survey it exhaustively. Instead, we concentrate on some popular statistical methods related to the subject of this thesis. Our review will also allow us to introduce notation and concepts for the discussions given in subsequent chapters. In Section 2.1 we establish a consistent notation for describing the concepts and the formal methods proposed in this thesis. In Section 2.2 we discuss AQR by only reweighting terms of the original query without query expansion. Section 2.3 focuses on AQR by selecting good terms by means of score functions and is closest in spirit to the studies in this thesis. Other related reformulation methods are described in Section 2.4.

2.1 Notation

Let $D = \{d_1, d_2, \cdots, d_N\}$ be a document *collection*, and a finite ordered tuple $V = \{t_1, t_2, \cdots, t_n\}$ the *vocabulary* of terms used to index at least one document in D. Let q be a query.

Let x be an object representing $x = d \in D$ or $x = q$. In this thesis, we will denote V^x as the set of terms that appear in object x, and $|V^x|$ the *size* of V^x (i.e., the number of *distinct* terms appearing in x), where $|\cdot|$ is a counting measure for a set. We will denote $f_x(t)$ as the *occurrence frequency* of term t in object x (i.e., the number of postings of term t in x), and $||x|| = \sum_{t \in V^x} f_x(t)$ the *length* of x. Obviously, $|V^x| \leq ||x||$ always holds. In this thesis, we will always assume that $2 \leq |V^d| \leq |V| = n$, i.e., each document has at least two distinct terms, and we will see that such an assumption is necessary in the estimations of the probability distributions. We will denote $max_{f_x} = \max\{f_x(t) |\ t \in V^x\}$ is the maximum frequency of the occurrence frequencies of terms in x.

Let $X \subseteq D$ be the set of documents in D. We will denote V^X as the sub-vocabulary consisting of those terms that appear in at least one document in set X, and $|X|$ the size of X (the number of documents in X). Particularly, when $X = D$, it has $V^D = V$ and $|D| = N$. We will denote $f_X(t) = \sum_{d \in X} f_d(t)$ as the occurrence frequency of term t in X (i.e., the total number of postings of term t in X), and $||X|| = \sum_{t \in V^X} f_X(t)$ as the length of X (also, we can write $||X|| = \sum_{d \in X} ||d||$, i.e., it is the sum of the lengths of individual documents

in X). Obviously, $|V^X| \leq ||X||$ always holds. We will denote $ave(D) = \frac{1}{|D|}\sum_{t\in V} f_D(t) = \frac{1}{|D|}\sum_{d\in D} ||d||$ is the average length of documents in D.

We will denote $F_X(t)$ as the *frequency of documents* in X in which t occurs. Clearly, it has $F_X(t) \leq F_D(t)$ since $X \subseteq D$. It is important to understand the difference between $F_X(t)$ and $f_X(t)$. Similarly, we will denote $F_X(t_i, t_j)$ as the frequency of documents in X in which t_i and t_j co-occur.

We will denote q' as an *expanded* (*modified*) query of an *original* query q, and S^q a set consisting of *selected* terms which are judged to be good terms. Thus, $V^{q'} = S^q \cup V^q$, and, generally, $S^q \cap V^q \neq \emptyset$., i.e., selected terms can be query terms. We call terms $t \in E^q = S^q - V^q$ (usually, $E^q \subset S^q$) *expansion terms.*

Finally, for a given query q, we will denote R as the set of all relevant documents in D, and $\bar{R} = D - R$ the set of all non-relevant documents in D, with respect to q. We will denote Ξ as a sample set obtained from an initial retrieval iteration, $\Xi^+ = \Xi \cap R$ a set of all relevant sample documents in Ξ, and $\Xi^- = \Xi \cap \bar{R} = \Xi - \Xi^+$ a set of all non-relevant sample documents in Ξ. Thus, we have $\Xi^+ \cup \Xi^- = \Xi$ and $\Xi^+ \cap \Xi^- = \emptyset$.

2.2 AQR by Reweighting Query Terms

Relevance feedback, introduced in the mid-1960s, is an automatic process for reformulating the original query based on relevance assessment provided by the user. Specifically, an initial retrieval is performed using the original query, and a small number of documents with high-similarity are presented to the user for relevance assessment. The assessment is then returned to the system and used to automatically modify the original query in such a way that terms appearing in previously retrieved relevant documents are emphasized, whereas terms in previously retrieved non-relevant documents are de-emphasized. Such a query reformulation process is expected to produce an enhanced query which has greater similarity with the relevant documents and greater dissimilarity with the non-relevant ones, and so retrieve more relevant documents while at the same time fewer non-relevant ones [161, 165].

The basic assumptions underlying relevance feedback are: (i) query terms are generally good at distinguishing relevant documents from non-relevant ones; (ii) terms co-occurring frequently with query terms might be likely to be good discriminators, and should be added to the original query; (iii) terms co-occurring frequently in some documents (with low document frequencies) may relate to the same topic. Thus, term co-occurrence statistics can be used to reveal some semantic relations inherent in terms. Extensive study of relevance feedback has been made within the frameworks of Boolean, linear algebraic, probabilistic, and language modelling methods [49, 87, 152, 153, 158, 160, 161, 164, 166, 167, 168, 207, 235].

In an operational situation where no relevance information is available in advance, the feedback process is called *pseudo-relevance feedback.* In this case, all documents in the sample set obtained from the initial retrieval are treated as relevant. Pseudo-relevance feedback aims to minimize the intellectual effort of human users at the query reformulation stage. It allows the reformulation to be performed by completely automatic means, in accordance with a term discrimination measure. The intellectual effort of users is shifted and concentrated at the time the system is set up. An effective discrimination measure is required to select good terms for enhancing the original query. Much of the research on query reformulation using

pseudo-relevance feedback has made significant performance improvements [74, 130, 165, 188].

2.2.1 Linear Algebraic Methods

In the linear algebraic (vector space) method, an ideal query is defined by Rocchio [158] as one which induces a ranking over the collection such that all relevant documents are ranked higher than non-relevant ones. Since relevance is a subjective attribute determined by individual users, there is no certainty that an ideal query in fact exists for a given query. Rocchio thus suggested computing an optimal query. An optimal query, corresponding to a given relevant document set R, under a weighting function, is the one which maximizes the difference between the mean of the similarities of relevant documents in R and the mean of the similarities of non-relevant documents in $\bar{R}$.

The Rocchio method cannot be used when sets R and $\bar{R}$ are unknown, in advance of the search being carried out. However, his method can help in generating a feedback query when relevance assessment is available for documents previously retrieved in answer to the query. In this case, all relevant or non-relevant documents used in his method are replaced by known relevant documents in $\Xi^+ \subseteq R$ or non-relevant documents in $\Xi^- \subseteq \bar{R}$.

Experience shows that the original query terms should be preserved by the feedback iteration process. Therefore, the formula (reweighting function) actually used by Rocchio to construct a new query from the original query q was:

$$rew_1(t) = \alpha\, w_q(t) + \frac{\beta}{|\Xi^+|} \sum_{d \in \Xi^+} \frac{w_d(t)}{\sqrt{\sum_{t \in V} w_d^2(t)}} - \frac{\gamma}{|\Xi^-|} \sum_{d \in \Xi^-} \frac{w_d(t)}{\sqrt{\sum_{t \in V} w_d^2(t)}}, \tag{2.1}$$

where $\alpha, \beta, \gamma \geq 0$ are constants, and $w_x(t)$ is weight of terms in $x = d \in D$ or $x = q$. Rocchio investigated relevance feedback using this formula, and found that it does improve retrieval results [158].

Ide [87] extended Rocchio's work by presenting three variations:

$$rew_{1,1}(t) = \alpha\, w_q(t) + \beta \sum_{d \in \Xi^+} w_d(t) - \sum_{d \in \Xi^-} w_d(t),$$

$$rew_{1,2}(t) = \alpha\, w_q(t) + \beta \sum_{d \in \Xi^+} w_d(t) - w_{d_s}(t),$$

$$rew_{1,3}(t) = \alpha\, w_q(t) + \beta \sum_{d \in \Xi^+} w_d(t).$$

In the second variation, $d_s \in \Xi^-$ is the first non-relevant document retrieved in the ranking list. Notice that, unlike Rocchio's formula, these three variations do not normalize the vector length.

Salton & Buckley [165] experimentally investigated and compared relevance feedback methods across six different test collections. From their results, they concluded: (i) $rew_{1,2}(t)$ is the best, whilst being computationally very efficient; (ii) for $rew_1(t)$, relatively higher weights should be given to terms obtained from the relevant documents than to those extracted from the non-relevant ones; (iii) expansion using all terms from the known relevant documents (i.e., without term selection, which is generally very expensive) is preferable to expansion using only the most common terms, but the performance difference is modest; (iv) expansion using the highest weighted terms is inferior.

Salton *et al.* [18, 165] modified the Rocchio formula: initially they considered non-relevant documents to be those that had been seen by the user and judged non-relevant (i.e., $d \in \Xi^-$). However, in the modified formula, they made an assumption that all unseen documents are non-relevant (i.e., $d \in D - \Xi$).

2.2.2 Binary Independence Probabilistic Methods

Over the past 40 years there has been a great deal of interest in using probabilistic methods for textual retrieval systems. The papers of [5, 13, 14, 37, 60, 62, 77, 78, 124, 136, 152, 153, 170, 199, 202, 206, 207, 223] are representative.

Most probabilistic retrieval methods are based on the so-called 'probability ranking principle' [147]. The principle asserts that, for optimum performance with a given query, a retrieval system should rank documents in order of their probability of relevance to the query, according to the information available to the system. Some counterexamples to this assertion can be given; however, it underlies much of the research in IR that exploits probability theory in a non-superficial way [36].

The binary independence probabilistic (BIP) method [77, 152, 153, 206, 235], an alternative typical relevance feedback method, may be the most well-known. A clear mathematical account of this formal method, outlined here, was presented in [61, 207].

Assume that terms are independently assigned to relevant and non-relevant documents of a collection, and that binary term weights restricted to 0 and 1 are assigned to documents. Under these assumptions, optimum performance can be achieved [146, 150, 210] by using a similarity measure (q' is a modified query):

$$sim(d, q') = \sum_{t \in V} w_d(t) \cdot rew_{q'}(t) = \sum_{t \in V^d \cap V^q} rew_{q'}(t),$$

and, for a given term $t \in V^{q'} = V^q$, a reweighting function is

$$rew_2(t) = \log \frac{p_t(1 - q_t)}{q_t(1 - p_t)} = rew_{q'}(t),$$

where p_t expresses the probability that term t is assigned to a certain relevant document, and q_t equivalently for a certain non-relevant document. This reweighting function can also be found in Bayes' decision theory [51, 132].

Notice that $rew_2(t)$ only modifies the weights of query terms, but no new terms are added to the query (i.e., there is no query expansion). Notice also that $rew_2(t)$ cannot be used in practice without knowing probabilities p_t and q_t.

Harper & Van Rijsbergen [77] pointed out that the best way of estimating probabilities is to estimate p_t from documents known to be relevant, and to estimate q_t from all documents not known to be relevant. Thus, the probabilities can be estimated by

$$p_t = \frac{F_{\Xi^+}(t)}{|\Xi^+|} \quad \text{and} \quad q_t = \frac{F_D(t) - F_{\Xi^+}(t)}{|D| - |\Xi^+|}.$$

Problems occur for the following cases:

1) $|\Xi^+| = 0$ (in this case $F_{\Xi^+}(t) = 0$),
2) $|\Xi^+| > 0$ but $F_{\Xi^+}(t) = 0$,
3) $|\Xi^+| > 0$ but $F_{\Xi^+}(t) = |\Xi^+|$.

This is because, in these cases, the logarithmic expression in $rew_2(t)$ is meaningless. That is, these three cases respectively correspond to three indeterminate expressions:

$$\log \frac{\frac{0}{0}}{a} \qquad \log \frac{\frac{0}{b}}{a} \qquad \log \frac{\frac{b}{0}}{a}$$

where $a, b > 0$. In practice, to solve these problems, an adjustment factor 0.5 is adopted in the conventional BIP method for estimating probabilities [152, 153]:

$$p_t = \frac{F_{\Xi^+}(t) + 0.5}{|\Xi^+| + 1} \quad \text{and} \quad q_t = \frac{F_D(t) - F_{\Xi^+}(t) + 0.5}{|D| - |\Xi^+| + 1}. \tag{2.2}$$

Experiments with this modified version consistently showed high retrieval performance [74, 152, 153].

From $rew_2(t)$, it is clear that a term assigned a high weight, implies that the term is prevalent among relevant documents in the collection; the reverse obtains for a term that occurs mostly among the non-relevant documents [170]. Thus, $rew_2(t)$ succeeds in emphasizing terms concentrated in the relevant documents. Terms with high weights may also be considered to have relatively low document frequencies, because terms with high document frequencies tend to appear indistinguishably both in relevant and non-relevant documents. Consequently, $rew_2(t)$ can be viewed as a measure of discrimination on relevance, and may be expected to be capable of enhancing retrieval precision.

Sparck Jones [187, 188, 189] performed a number of experiments, and found that the use of $rew_2(t)$, with only a few relevant documents, would result in significant performance improvements over weighting terms using an alternative reweighting function

$$rew_3(t) = idf_D(t) = \log \frac{|D|}{F_D(t)},$$

which is normally called the weight of the inverse document frequency of terms. Early studies, [169, 170, 171, 172, 185, 207] for instance, showed that document frequencies, $F_D(t)$, of terms are directly related to the power of discrimination of terms on relevance. Hence document frequency can be considered as a discrimination measure. We will return to this function in Section 2.4.

Wu & Salton [229] showed that both $rew_2(t)$ and $rew_3(t)$ are closely related over a wide spectrum of document frequencies. Particularly, for the medium frequency terms, both weights $rew_2(t)$ and $rew_3(t)$ are rather similar. They pointed out, since most query terms used in practice may be expected to fall in the medium frequency range, that reweighting terms with $rew_2(t)$ would not produce retrieval results that are substantially better than with $rew_3(t)$.

The BIP method has been criticized for a variety of reasons. Firstly, the adjustment factor 0.5 may provide very poor estimation when $F_{\Xi^+}(t) = 0$ [40]. In fact, it overestimates the probabilities involved [181, 210], and hence, this solution to the estimation problem is not ideal [233]. Some different adjustment factors have successively been proposed to estimate probabilities p_t and q_t [22, 149, 165, 233]. Secondly, documents can only be represented as binary vectors, although attempts have been made to remove such a restriction, [234] for instance. Finally, it is often difficult to justify the probabilistic independence assumption.

2.2.3 The EMIM Methods

In order to overcome the estimation problems posed by the BIP method, Harper & Van Rijsbergen [77] proposed the following term reweighting function:

$$rew_4(t_i) = \sum_{\delta_i,\gamma_q=1,0} \Delta_{iq} P(\delta_i,\gamma_q) \log \frac{P(\delta_i,\gamma_q)}{P(\delta_i)P(\gamma_q)},$$

where $\Delta_{iq} = 1$ if term t_i occurs (or does not occur) in a relevant (or non-relevant) document with respect to query q; $\Delta_{iq} = -1$ otherwise. They used the factor Δ_{iq} as an indicator of how good a term t_i is as a relevance discriminator. The presence of a term in a relevant document, and its absence in a non-relevant document, implies that the term is a positive discriminator of relevance. Function $rew_4(t)$ is in fact the expected mutual information measure incorporating factor Δ_{iq}.

Further, Harper & Van Rijsbergen [77] proposed a way to estimate the probability distribution involved in $rew_4(t)$. Thus, we have the following reweighting function:

$$\begin{aligned} rew_{4,1}(t) = {} & F_{\Xi^+}(t) \log \left(\frac{F_{\Xi^+}(t)}{F_D(t)} \frac{|D|}{|\Xi^+|}\right) \\ & - \left(F_D(t) - F_{\Xi^+}(t)\right) \log \left(\frac{F_D(t) - F_{\Xi^+}(t)}{F_D(t)} \frac{|D|}{|D| - |\Xi^+|}\right) \\ & - \left(|\Xi^+| - F_{\Xi^+}(t)\right) \log \left(\frac{|\Xi^+| - F_{\Xi^+}(t)}{|D| - F_D(t)} \frac{|D|}{|\Xi^+|}\right) \\ & + \left(|D| - |\Xi^+| - F_D(t) + F_{\Xi^+}(t)\right) \log \left(\frac{|D| - |\Xi^+| - F_D(t) + F_{\Xi^+}(t)}{|D| - F_D(t)} \frac{|D|}{|D| - |\Xi^+|}\right). \end{aligned}$$

They performed a set of experiments with the Cranfield collection, using complete relevance information. The results showed that query terms reweighted using $rew_{4,1}(t)$ give much better performance than those using $rew_2(t)$.

Sparck Jones & Webster [195] conducted some experiments using $rew_2(t)$ and $rew_{4,1}(t)$ and concluded that, when a fair amount of relevance information is available, query reformulation may be positively advantageous compared with the original query.

Smeaton & Van Rijsbergen [181] used $rew_2(t)$ and $rew_{4,1}(t)$ to carry out a set of experiments using the NPL collection. The results showed that the performance obtained from $rew_2(t)$ appears better than that of $rew_{4,1}(t)$, but there was no significant difference between them.

Biru *et al.* [10] attempted to analyse the relationship between the power of discrimination of terms on relevance and the document frequencies of terms by means of an alternative reweighting function:

$$\begin{aligned} rew_{4,2}(t) = {} & F_{\Xi^+}(t) \log \left(\frac{F_{\Xi^+}(t)}{F_D(t)} \frac{1}{|\Xi^+|}\right) \\ & - \left(F_D(t) - F_{\Xi^+}(t)\right) \log \left(\frac{F_D(t) - F_{\Xi^+}(t)}{F_D(t)} \frac{1}{|D| - |\Xi^+|}\right) \\ & - \left(|\Xi^+| - F_{\Xi^+}(t)\right) \log \left(\frac{|\Xi^+| - F_{\Xi^+}(t)}{|D| - F_D(t)} \frac{1}{|\Xi^+|}\right) \\ & + \left(|D| - |\Xi^+| - F_D(t) + F_{\Xi^+}(t)\right) \log \left(\frac{|D| - |\Xi^+| - F_D(t) + F_{\Xi^+}(t)}{|D| - F_D(t)} \frac{1}{|D| - |\Xi^+|}\right). \end{aligned}$$

It is interesting to notice that this reweighting function is rather different from $rew_{4,1}(t)$, and the difference is not only a constant.

In previous experiments, it had been found that $rew_{4,1}(t)$ outperformed $rew_2(t)$. It was pointed out [40] that the only disadvantage of using $rew_{4,1}(t)$ is that some relevant documents far down in the initial ranking may be pushed even further down after feedback, and that this demotion happens because terms that do not occur in the relevant retrieved documents may be given lower weights.

Harper & Van Rijsbergen [77] wrote that there may be theoretical justification for exploiting term dependence in this particular way. Smeaton & Van Rijsbergen [181] stated that $rew_2(t)$ can be shown to be optimal (under the independence assumption), and that $rew_{4,1}(t)$ has an *ad hoc* basis and is suboptimal. Similarly, Croft [40] stated that $rew_{4,1}(t)$ is heuristic in nature, whereas $rew_2(t)$ is theoretically superior. Van Rijsbergen, Harper & Porter [210] also wrote that there is no theoretical justification for the dependence weight $rew_4(t)$, but it outperforms the independence weight $rew_2(t)$. It may be that, in the light of $rew_{4,1}(t)$'s robustness and effectiveness, some theoretical justification will be found. We will discuss this very interesting issue and justify $rew_{4,1}(t)$ in Chapter 7.

The mutual information measure has been used widely in many applications of IR, [9, 34, 64, 96, 102, 108, 230], for instance. We will discuss this measure in depth in Chapter 7.

2.2.4 Adaptive Linear Methods

The concept of user preference is closely related to the concept of relevance. A user preference can be formally expressed by a binary relation $\succ$ on collection D, which reflects the qualitative judgement of preference from the user point of view.

It has been shown [222] that if relation $\succ$ on D satisfies some additional conditions (a mathematical discussion about the conditions can be found in [56]), then we can express such a relation by a linear decision function. It has also been shown [134] that there exists one necessary and sufficient condition for the existence of a linear decision function based on measurement theory within the framework of user preference. More precisely, in the general vector space model, suppose that each document $d \in D$ is characterized by the term probability distribution $p_d(t)$ on V. Then, under the additional conditions, there exists a real-valued function u on V, for two arbitrary documents $d_1, d_2 \in D$, satisfying

$$\begin{aligned} \text{user prefers } d_1 \text{ to } d_2 \iff d_1 \succ d_2 &\iff E(u, p_{d_1}) > E(u, p_{d_2}) \\ &\iff \sum_{t \in V} u(t) p_{d_1}(t) > \sum_{t \in V} u(t) p_{d_2}(t), \end{aligned}$$

where $u = u(t)$ is called a utility function, which can be viewed as a measure of usefulness of a term with respect to relation $\succ$. The expected-utility $E(u, p_d)$ can be viewed as a measure of usefulness of document d with respect to query q, and used to rank documents $d \in D$.

Wong & Yao [221] suggested two simple methods for constructing such a linear decision function, namely, two methods of estimating the utility function $u(t)$. One is based on the user input query, i.e., $u(t) \approx w_q(t)$. Another is based on the BIP method, i.e., $u(t) \approx rew_2(t) = w_{q'}(t)$.

Further, Wong *et al.* [222, 225] proposed an iterative algorithm in terms of the gradient descent function [51] for constructing an appropriate query vector from the user preference relation. Thus, given the user preference judgement on a sample set, a query can be automatically generated by the iterative algorithm without introducing any specific formulae or

parameters. Their studies showed that Rocchio's optimal solution, $rew_1(t)$, is only a special case in the general gradient descent procedure.

Wong *et al.* [225] demonstrated, when the iterative algorithm is applied to a binary document representation system, that a modified query can be reformulated by the following reweighting function

$$rew_5(t) = |D||\Xi^+|\left[\frac{F_{\Xi^+}(t)}{|\Xi^+|} - \frac{F_D(t)}{|D|}\right].$$

They viewed the absolute value $\left|\frac{F_{\Xi^+}(t)}{|\Xi^+|} - \frac{F_D(t)}{|D|}\right|$ as an approximate measure of the power of term t to discriminate on relevance. Further, Wong & Yao [222] showed that the iterative algorithm may be applicable to any document representations.

Subsequently, Wong *et al.* [226] evaluated the retrieval performance of the iterative algorithm. The results from their experiments seem to be superior to $rew_{1,2}(t)$ [87], which is considered to be the best for relevance feedback [164].

The use of the iterative algorithm requires the user to provide a complete preference structure (restricted to satisfy the additional conditions) on the sample set according to his preference judgement. This implies that the user may be forced to browse (perhaps to 'read') all sample documents, and make some comparison between documents in order to determine his preference structure properly. Such a practice would likely lead to the user being in a much better position to reformulate the query himself or, perhaps, refusing to provide a preference structure simply because he feels that it is easier to reformulate the query. Also, in the context of IR, the conditions that the user preference relation must satisfy may be too stringent, and we may not be able to justify any choice of a linear decision function defined by $E(u, p_d)$ meets the user preference relation [221].

2.2.5 Term Dependence Probabilistic Methods

In fact, the term independence assumption is grossly inaccurate, and tends to be too simple and too strong. As stated by Cooper & Huizinga [37]: "*... arbitrarily adopting special independence assumptions is not a wholly desirable approach to the problem of obtaining sound probability-of-usefulness estimates in information search systems. Indeed, such assumptions are usually recognised to be crude even by those who employ them, their use being justified more or less as a desperation measure.*" Bollmann & Raghavan [133] also indicated that retrieval, based on a similarity measure, such as the cosine function used in the vector space model, is incompatible with term independence assumption.

In order to remove the artificially simplifying and unrealistic assumption of term independence, retrieval techniques based on the assumption of term dependence have successively been developed; [77, 206, 207] were early influential representatives.

Cooper & Huizinga [37] addressed the problem of how to make probability estimates, without introducing the independence assumption, using the maximum entropy principle (MEP). Kantor [97] discussed the technical issue of formulating the maximum entropy problem in a realistic IR environment. This work had almost no impact, although Kantor & Lee [98, 99] provided one 'toy calculation' showing that the MEP behaves sensibly in the situation where terms co-occur very frequently. After some experiments with the TREC collection, Kantor & Lee [99] drew the conclusion, "*the evidence cannot support the strong claim that the MEP accurately describes the distributions of terms across relevant and non-relevant documents. Nor*

does it support the weaker claim that computing according to the MEP will lead to enhanced information retrieval".

Croft [41] proposed a method which attempted to integrate Boolean and probabilistic retrieval methods. In his method, Boolean queries are interpreted as specifying term dependencies in the set of relevant sample documents. Later, Croft & Lewis [43] presented an algorithm to generate dependent term groups from their own representations. Losee & Bookstein [117] proposed a probabilistic method integrating Boolean queries in conjunctive normal form, where most of the dependencies exist between the disjunctions of terms.

Cho, Lee & Lee [31] proposed a method which incorporates term dependence into a probabilistic retrieval method by adapting the concept of Bahadur-Lazarsfeld expansion (BLE) developed in the area of pattern recognition [51]. A theoretic process in applying BLE to the probabilistic method [152] and to the state-of-the-art 2-Poisson method [154] was described.

It has been recognized that incorporation of term dependence information into term weights would improve the retrieval performance to some extent [31, 41, 43, 116, 133, 206]. However, the term dependence methods have not been shown to have consistently and significantly better performance than the methods assuming term independence of [40]. Also, a major disadvantage in using these methods is that they are computationally very expensive because information on the co-occurrence of two terms must be obtained at search time [31]. Another problem is that some of them need to decide the parameters which might be necessary for determining retrieval performance, but the parameter estimation cannot be performed in real time.

2.2.6 Language Modelling Methods

Language modelling methods were introduced to IR by Ponte and Croft [136], and further explored in [9, 44, 109, 112, 127, 182, 232, 238, 239]. These methods have recently been proposed as an alternative to the conventional vector space and probabilistic methods, and have been shown to have relatively effective performance experimentally. The basic idea of these methods is to estimate a language model for each document, and then rank documents by the likelihood of the query according to the estimated language models.

An essential problem in the language modelling methods is the smoothing of estimation. The language models for IR must be smoothed, so that non-zero probability can be assigned to query terms that do not appear in a given document [109]. The smoothing is directly related to the retrieval performance. Miller *et al.* [127] smoothed the document language model with a background model (i.e., a collection language model) using hidden Markov chains. Zhai & Lafferty [239] also studied the problem, and examined the sensitivity of retrieval performance to the smoothing parameters.

In the language modelling framework, Berger & Lafferty [9] proposed a method exploiting ideas and approaches of statistical machine translation: how a user may translate a given document into a query. The query expansion, which can be viewed as using a Markov chain method, when applied to a set of sample documents, can be regarded as a method that re-estimates an existing query language model. The Markov chain method is a very general method for expanding either a query language model or a document language model [109].

Lavrenko & Croft [112] suggested a method of estimating probabilities of terms in the relevant document class by estimating a query language model based on a set of relevant sample documents. Their method also attempts to address the important issues of synonymy and polysemy.

Lafferty & Zhai [109] developed a feedback method for estimating an expanded query language model, which may assign probability to terms that are not in the original query. The essence of their method is a Markov chain term translation model that can be computed based on a set of sample documents.

Zhai & Lafferty [238] proposed a feedback method within a language modelling framework, which incorporated the directed divergence measure into retrieval strategies [109]. They suggested two schemes for re-estimating an existing query language model: one is to estimate the query language model using the relevant sample documents based on a maximum likelihood, and the other is to estimate the query language model by minimizing the average divergence between the query language model and the relevant sample document language models.

Cronen-Townsend *et al.* [44] introduced a method for predicting query performance by computing the directed divergence of a query language model, estimated using the method given in [112], from a collection language model.

However, feedback strategies have been dealt with heuristically within the language modelling methods, and are not very compatible with the essence of these methods. As a result, the expanded query may be interpreted differently from the original query [238].

2.2.7 Some Experimental Methods

Croft & Harper [42] investigated application of the BIP method to the situation where no relevance information is available (i.e., $|\Xi^+| = 0$), the term probability distribution can be estimated by $p_t = \alpha$ (where $0 < \alpha < 1$ is a constant) and $q_t = \frac{F_D(t)}{|D|}$. The query terms can thus be weighted by

$$w_{CH}(t) = \log \frac{\alpha}{1-\alpha} + \log \frac{|D| - F_D(t)}{F_D(t)}.$$

In this weighting function, the first item is simply a constant, whereas the second item is essentially $rew_3(t)$. This function has been shown to be effective [40, 42, 76].

Croft [40] extended the work of Croft & Harper [42] by incorporating within-document term weights $w_d(t)$ into relevance feedback search. The reweighting function for modifying the query was:

$$rew_6(t) = \left[k + (1-k)w_d(t)\right] \times w_{CH}(t),$$

where $0 < k < 1$ is a constant, $w_d(t) = \frac{f_d(t)}{max_{f_d}}$ is term weight, and $max_{f_d} = \max\{f_d(t) | \, t \in V^d\}$ is the maximum frequency of terms in d. The Cranfield and NPL collections were used in his experiments. He found that constant k significantly affects retrieval performance, and that the optimum value of k is different for the different collections. This reweighting function significantly improved retrieval performance for both collections compared with the performances obtained from the original queries and from query reformulation using $rew_2(t)$.

Harman [74] proposed a reweighting function (without query expansion):

$$rew_7(t) = \frac{\log\left(f_d(t) + 1\right)}{\log(||d||)} \times rew_2(t).$$

10 top-ranked documents were used for relevance feedback (with the Cranfield collection). However, there was no difference between her reweighting function and $rew_2(t)$.

Klink *et al.* [103] presented a method which used relevance feedback information and information globally available from old queries. The original query was then expanded using

the previously learnt concepts. According to their method, each concept, c_j, of query term $t_j \in V^q$ can be built from the sum of all vectors of documents relevant to the old queries having term t_j in common. Then, each query term t_j corresponds to a concept c_j, which has been learnt from old queries. Their experimental results showed that this method is more effective than Rocchio's for some collections, but ineffective or even worse for others. Notice that their method used all terms in 'relevant' documents without a term selection stage, and that the documents are relevant to some old queries, rather than to the original query, and so it is likely that these relevant documents are unrelated to the original query, and the expanded query, by using terms appearing in non-relevant documents, will produce 'drift'.

2.3 AQR by Adding Good Terms

Robertson [150] stated that it may be appropriate to filter relevance feedback terms, and that rather than expanding the original query with all the relevance feedback terms, many of which may have low weights, only the best terms should be selected. Harman [74] pointed out, after experimental analysis, that adding only well-selected relevance feedback terms is superior to adding all relevance feedback terms. Carpineto *et al.* [29] also pointed out that query expansion using all relevance feedback terms may be only slightly better than using selected good terms, and that using a limited number of expansion terms may be important in reducing response time, especially for large collections.

Thus, in this section, we discuss some query expansion methods. Query expansion can be automatic and, in this case, the system judges good terms and adds them into the original query without reference to the user. Query expansion can also be semi-automatic and, in this case, the system identifies potential good terms and presents them to the user for possible addition. Either way, it is necessary to design a score function which can measure how good a term is as a discriminator on relevance.

2.3.1 Information Measure Based Methods

To measure the power of discrimination of terms Carpineto *et al.* [28, 30] proposed a score function using a divergence measure [107] (in pseudo-relevance feedback):

$$score_1(t) = \big(P_\Xi(t) - P_D(t)\big) \log \frac{P_\Xi(t)}{P_D(t)},$$

in which, term probability distributions were estimated by

$$P_X(t) = \frac{\sum_{d \in X} f_d(t)}{\sum_{d \in X} \big(\sum_{t \in V^d} f_d(t)\big)},$$

where $X = \Xi$ or $X = D$. They carried out a series of experiments on TREC-6, TREC-7 and TREC-8 data (with $|\Xi| = 5$ and $|E^q| = 30, 60$). Terms were considered as expansion terms if they made a marked contribution to the divergence (see a detailed account in Chapter 3). The experimental results showed that rankings with the expanded queries achieved a better performance than ranking with the original queries.

A necessary condition that must be satisfied in application of the divergence measure is that the two probability distributions are absolutely continuous with respect to one another,

otherwise the divergence is meaningless. Usually, the condition is not satisfied when we attempt to derive the probability distributions from the different sets of documents for the purpose of query expansion. It is a key issue that needs to be carefully analysed, in order to establish the rationality of applying the divergence measure to feedback. Carpineto, *et al.* [28] thoroughly discussed this problem, and, in order to solve it, suggested a scheme that attempted to find a 'discounting factor' μ $(0 < \mu < 1)$ for discounting the probability distribution of terms. In their work, however, it seemed that factor μ was not really found, and the main experiments described relied on $\mu = 1$. In fact, the theoretical problem of applying the divergence measure to query expansion remains open, and is one of the focal points of the study given in Chapter 4.

Carpineto *et al.* [29] proposed an alternative score function based on the directed divergence measure [106]:

$$score_2(t) = P_\Xi(t) \log \frac{P_\Xi(t)}{P_D(t)}.$$

Further, they experimentally compared $score_2(t)$ with four other score functions:

$$\begin{aligned}
score_3(t) &= \sum_{d \in \Xi} w_d(t), \\
score_4(t) &= \big(\sum_{d \in \Xi} w_d(t)\big) P_\Xi(t), \\
score_5(t) &= \big(P_\Xi(t) - P_D(t)\big)/P_D(t), \\
score_6(t) &= \big(P_\Xi(t) - P_D(t)\big)^2/P_D(t),
\end{aligned}$$

where, $score_3(t)$ is a variation of the Rocchio formula, $score_4(t)$ was given by Robertson [150] and $score_5(t)$ by Doszkocs [50]. From the experimental results (with $|\Xi| = 10$ and $|E^q| = 40$ on TREC-7 and TREC-8 data), they found: (i) expanded queries worked markedly better than the original queries for all five score functions; (ii) $score_2(t)$ is the only one which leads to a significant improvement over the Rocchio formula, whereas score functions 4,5,6 led to worse performances; (iii) the term scores obtained from $score_2(t)$ can be used not only for selecting expansion terms but also for reweighting them; (iv) when the Rocchio formula is used for reweighting expanded query terms, the use of more sophisticated methods, such as, score functions 4,5,6, for selecting expansion terms, does not produce any performance improvement.

Carmel *et al.* [27] presented a query expansion method, which was based on the information gain obtained by adding lexical affinities to the query. Each lexical affinity, LA, is a term pair (t_i, t_j), one of which must be a query term. LAs are used to represent the dependence between terms co-occurring in a document, and identified by looking at term pairs found in close proximity to each other [121]. For a given LA, denote Ξ_1 as a set including all documents containing the LA, $\Xi_0 = \Xi - \Xi_1$ as a set including all documents not containing the LA. A score function, for selecting good LAs, based on the entropy increase measure (see a detailed account in Chapter 6) was suggested:

$$score_{7,1}(LA) = H\big(p(\Xi)\big) - \Big[\frac{|\Xi_1|}{|\Xi|} H\big(p(\Xi_1)\big) + \frac{|\Xi_0|}{|\Xi|} H\big(p(\Xi_0)\big)\Big],$$

where $H\big(p(\cdot)\big) = -p(\cdot)\log\big(p(\cdot)\big) - \big(1 - p(\cdot)\big)\log\big(1 - p(\cdot)\big)$. The first item in $score_{7,1}(t)$ is the entropy of set Ξ before splitting, and the second is the average entropy of subsets Ξ_1

and Ξ_0 after splitting. The difference is information gained by the splitting process. They claimed that a good term is one that successfully differentiates all relevant sample documents into Ξ_1 and all non-relevant sample documents into Ξ_0; in this case, the entropy will be reduced to zero, and information gain $score_{7,1}(t)$ is maximal. In the situation where no relevance information is available, they estimated the probabilities by means of the similarities of documents to the query. Experiments were carried out with TREC-7 data, and results showed improvement in precision when adding 2-3 LAs. However, the analysis of lexical affinity is generally very expensive, and the quality of the estimation depends highly on the similarity measure.

It is interesting to notice that the following score function

$$score_{7,2}(LA) = H\big(p(\Xi)\big) - \Big[\frac{|\Xi^+|}{|\Xi|}H\big(p(\Xi^+)\big) + \frac{|\Xi^-|}{|\Xi|}H\big(p(\Xi^-)\big)\Big],$$

seems to capture their intention more clearly: the addition of LA, if it is a good discriminator, may better separate relevant sample documents from non-relevant sample ones. Their experimental results showed that the performance obtained from $score_{7,2}(LA)$ is markedly better than that obtained from $score_{7,1}(LA)$.

The directed divergence, divergence and entropy increase measures were also used in other studies, for instance, [5, 44, 48, 109, 232, 238]. We will discuss these measures in the subsequent chapters.

2.3.2 Some Experimental Methods

Smeaton & Van Rijsbergen [181] conducted some query expansion experiments with the NPL collection. The power of discrimination of the feedback terms ($t \in V^{\Xi^+} - V^q$) was measured by using the following four score functions:

$$\begin{aligned}
score_8(t) &= rew_2(t),\\
score_9(t) &= rew_{4,1}(t),\\
score_{10}(t) &= |D|F_{\Xi^+}(t) - |\Xi^+|F_D(t),\\
score_{11}(t) &= F_{\Xi^+}(t),
\end{aligned}$$

where $score_{10}(t)$ was proposed by Porter [138]; $score_{11}(t)$ was discussed by Martin [125] and Ingwersen [88]. For $score_{11}(t)$, the ties (whenever there is more than one term with the same document frequency) are ranked in alphabetical order. $\frac{|V^q|}{2}$ terms were added to the query, and the reweighting of expanded query terms used $rew_2(t)$ and $rew_{4,1}(t)$. For these four score functions and two reweighting functions, all performances showed degradation when compared with the performances obtained from reweighting only query terms (without query expansion) using $rew_2(t)$ and $rew_{4,1}(t)$. From the results obtained, they concluded that these four score functions possessed roughly the same discrimination power.

Buckley *et al.* [19] also used $score_{11}(t)$ to rank all relevance feedback terms for query expansion. However, in their experiments, the ties were broken by considering the highest average weight among the average weights $\frac{1}{|\Xi^+|}\sum_{d\in\Xi^+} w_d(t)$.

Harman [72] tested the effectiveness of query expansion by using the Cranfield collection. 10 top-ranked relevant documents (i.e., $|\Xi^+| = 10$) were used for query expansion (no reweighting of terms in the expanded queries). In her experiments, score functions were generated based on the product of factors considered to be important in measuring the power

of discrimination of terms. The score functions used to rank relevance feedback terms were $score_{11}(t)$ and:

$$score_{12}(t) = noise(t),$$
$$score_{13}(t) = noise(t) \times \log\Big(\sum_{d \in \Xi^+} f_d(t)\Big),$$
$$score_{14}(t) = noise(t) \times \log\Big(\sum_{d \in \Xi^+} f_d(t)\Big) \times F_{\Xi^+}(t),$$

in which, ranking with $noise(t)$ (which we discuss in Section 3.7) was from lowest noise to highest noise. 20 top-ranked terms were added to the query. From the experimental results she concluded: (i) $score_{12}(t)$ performs the worst, which indicates that term distribution within the collection is not a good discrimination measure; (ii) $score_{11}(t)$ works better than $score_{12}(t)$, because if a term appears in most relevant documents, this usually implies it describes a concept central to these documents; (iii) $score_{13}(t)$ is not very effective even though term frequencies are used; (iv) $score_{14}(t)$ seems to be the best and works much better than $score_{12}(t)$, because a term appearing frequently in a document is often an important term in the document and this idea can be extended to a set of documents.

For the four score functions 11,12,13,14, Harman [74] carried out a new set of experiments. This time, however, she used $rew_2(t)$ to reweight the expanded query terms. In her new experiments, the poorer function $score_{12}(t)$ was replaced by a score function

$$score_{15}(t) = rew_3(t).$$

Her experimental results showed that $score_{13}(t)$ is the best, and that there are marginal differences between score functions 11,14,15. She noticed that the major improvements come from expansion terms, although reweighting of terms contributes a further improvement.

Robertson [150] suggested that, under some assumptions (term independence and binary term weights for document representation), the power of discrimination of terms can be measured using a score function with the form

$$score_R(t) = w(t)(p_t - q_t),$$

where $w(t)$ is a weighting function, p_t expresses the probability that term t is assigned to a relevant document, and q_t equivalently for a non-relevant document. He stressed that the relevance feedback terms should then be ranked according to their scores $score_R(t)$, rather than their weights $w(t)$. $score_R(t)$ has been widely used in various systems with different weighting functions and different methods of estimating p_t and q_t [19, 29, 52, 53, 74, 111, 156, 159].

Harman [74] carried out a set of experiments with the Cranfield collection, and the addition of 20 expansion terms. Several score functions were tested in her experiments. These functions are all related to ratios or probabilities of a term occurring in relevant documents as opposed to occurring in non-relevant ones. They are $score_8(t)$ and

$$score_{16}(t) = \frac{F_D(t)}{|D|} \times N_{DR},$$
$$score_{17}(t) = \frac{F_R(t)}{|R|} - \frac{F_D(t)}{|D|},$$

$$score_{18}(t) = \log\left(\frac{|D|}{F_D(t)} + 1\right) \times \log\left(\sum_{d \in \Xi^+} f_d(t)\right),$$

$$score_{19}(t) = \log\left(\frac{|D|}{F_D(t)} + 1\right) \times (p_t - q_t),$$

$$score_{20}(t) = \log\left(\frac{p_t(1-q_t)}{q_t(1-p_t)}\right) \times (p_t - q_t),$$

$$score_{21}(t) = \log\left(\sum_{d \in \Xi^+} f_d(t)\right) \times (p_t - q_t),$$

where $score_{16}(t)$ was given by Doszkocs [50], in which, factor N_{DR} is the total number of retrieved documents; $score_{17}(t)$ was given by Porter & Galpin [139]; score functions 19,20,21 were $score_R(t)$ with $w(t)$ being the inverse document frequency weights, the BIP weights, and the total frequency of terms in the relevant sample set, respectively. The probabilities, p_t and q_t, were estimated by Eq.(2.2). In her experiments, she used $rew_7(t)$ to reweight the expanded query terms. From her results, she concluded: (i) score functions 16,17 are not very effective; (ii) $score_8(t)$ is the most effective; (iii) there is no significant difference between score functions 8,18,19,20,21.

In the experiments given by Allan [3], about 10% of the relevance judgements from TREC for disks 1-2 were used. The relevance feedback terms were first ranked with a simple score function

$$score_{22}(t) = \sum_{d \in \Xi^+} f_d(t).$$

500 top-ranked terms were then re-ranked according to the Rocchio formula, that is,

$$score_{23}(t) = rew_1(t),$$

with parameters $\alpha = 1$, $\beta = 2$, $\gamma = \frac{1}{2}$, and with weights of terms (in query $x = q$ or document $x = d$):

$$w_x(t) = 0.4 + 0.6\frac{f_x(t)}{f_x(t) + 0.5 + 1.5[||x||/ave(D)]}\frac{\log\left([|D| + 0.5]/F_D(t)\right)}{\log\left(|D| + 1\right)},$$

where $ave(D)$ is the average length of documents in collection D. 100 top-ranked terms in the second ranking list were treated as good discriminators and added to the query. His experimental results showed that the expanded queries give a performance improvement over the original queries.

In addition, Allan [3] experimentally investigated the effect of feedback using partial relevance information, and he concluded that partial relevance information can achieve almost the same precision and recall as complete relevance information, even though it includes only 10% of relevant documents (judgements provided by TREC data). This indicated that we have lost very little important information by sampling. His experimental results showed that the sampling provides a reasonable approximation to the complete information. In his experimental environment (with an average of almost 1800 relevant documents per query) 10% of relevant documents resulted in 10-30 relevant documents per query, which is a lot for an interactive setting.

Allan *et al.* [4] conducted an alternative experiment studying the effectiveness of query expansion. The complete relevance information from TREC disks 1-3 were used for feedback.

The following score function was used to rank relevance feedback terms:

$$score_{24}(t) = \frac{F_R(t)}{|R|} - \frac{F_{\Xi^-}(t)}{|\Xi^-|},$$

where Ξ^- (with $|\Xi^-| = |R|$) consists of the top-ranked non-relevant documents initially retrieved. Ξ^- was incorporated into the sample set (i.e., $\Xi = R \cup \Xi^-$) for training good terms. $score_{24}(t)$ is rather similar to $score_{17}(t)$. 50 top-ranked terms were selected. The expanded query terms were reweighted using the Rocchio formula $rew_1(t)$ with parameters $\alpha = 0$, $\beta = 2$, $\gamma = \frac{1}{2}$, and with weights of terms in document d:

$$w_d(t) = 0.4 + 0.6\Big[0.4 \min\{1, \frac{200}{max_{f_d}}\} + 0.6\frac{\log(f_d(t) + 0.5)}{\log(max_{f_d} + 1.0)}\Big]\frac{\log([F_D(t) + 0.5]/|D|)}{\log(|D| + 1)},$$

where max_{f_d} is the maximum frequency of terms in d. However, their experimental results did not show that this method is effective even though complete relevance information is used.

2.3.3 Passage-Level Search Methods

A long document comprises many different subtopics which may be related to one another and to the context in many different ways [83]. Long documents usually contain too much information, which reduces the effectiveness of feedback. Trimming long documents by choosing a good passage has a marked impact on effectiveness [2]. The same technique used to rank documents can be applied to the passages of a document, and the best-ranked passage of a document can be chosen for feedback in place of the entire document [2, 24].

Many attempts have been made to generate passages. Wilkinson [219] split documents into individual sections. Robertson *et al.* [157] used sub-documents consisting of an integral number of consecutive paragraphs. Hearst *et al.* [83] broke documents down to multi-paragraphs. Callan *et al.* [24] showed that passages based upon paragraph boundaries are less effective than passages based upon overlapping text windows. Buckley *et al.* [19] experimentally demonstrated that fixed-length text windows may be more effective than short sentences and paragraphs. Callan *et al.* [24] and Allan [2] suggested, when fixed-length passages are used, that anywhere between 200-300 words is a good choice for a variety of collections.

Overlapping passages were fixed at some length l: the first passage in a document starts at the first term matching a query term, and ends l terms after that, and subsequent passages begin at intervals of $\frac{l}{2}$ from the first starting point. For example, if $l = 200$ and the first matching term is at position 7, overlapping passages would start at positions 7, 107, 207, etc. The use of overlapping passages reduces the chance that a small block of relevant text passage is split among two passages. In a study given by Buckley *et al.* [19], the setting of text windows starts at the beginning of the document, with a fixed length 200 words.

In Allan's query expansion experiments [2], $|\Xi^+|$ passages (from the corresponding relevant sample documents) were used for feedback. The score function used in his experiments was:

$$score_{25}(t) = F_{\Xi^+}(t)\frac{\log([|D| + 0.5]/F_D(t))}{\log(|D| + 1)} \approx F_{\Xi^+}(t)\Big[1 + \frac{\log|D|}{\log F_D(t)}\Big],$$

since $|D| + 0.5 \approx |D| + 1 \approx |D|$. He claimed that this function is rather similar to function $f_d(t) \times idf_D(t)$, which is considered an effective discrimination measure [164, 171, 172]. The

top $|E^q| = \min\{3 + 2|\Xi^+|, 300\}$ terms were selected and added to the query. The expanded query terms were reweighted using a function:

$$rew_8(t) = \alpha(t)\Big[\sum_{d \in D} f_d(t)\Big] \frac{\log\left([|D| + 0.5]/F_D(t)\right)}{\log\left(|D| + 1\right)},$$

where the first factor $\alpha(t) = 1.0$ if $t \in V^q$ is a query term, and $\alpha(t) = 0.3$ if $t \in E^q$ is an expansion term; the second factor, in the square brackets, is the term frequency in the collection. His experimental results showed that the expanded queries result in an average performance improvement over the original queries.

In Xu's experiments [230], each document was broken into 300-word passages. TREC-3 and TREC-4 data and the WEST collection were used in the experiments. For each feedback term $t \in V^{\Xi} - V^q$, he used the following score function:

$$score_{26}(t) = \prod_{t_j \in V^q} \left[\delta_0 + \frac{\log\left(F_{\Xi}(t, t_j) + 1\right)}{\log(|\Xi|)} \times \delta_t\right]^{idf_D(t_j)},$$

where $\delta_t = \min\left\{1.0,\ 0.2 \log \frac{|D|}{F_D(t)}\right\}$. He treated the second item in square brackets to be a measure used to calculate the degree of dependence of terms t and t_j; the first item δ_0 is a small constant which is added to each degree in order to handle the situation where a query term does not occur in the top-ranked documents and the second item is zero. The results showed performance improvement compared with the original queries for TREC-3 and TREC-4, but was worse for the WEST collection.

Passage-level operations are costly in some systems [2]. Particularly, when the setting of text windows uses the technique that the first passage in a document starts at the first term matching a query term, passage locations vary from query to query and are very expensive.

2.3.4 Interactive Methods

In an interactive query expansion (IQE), the potential expansion terms are shown to the user for selection. The user then decides which to add and which to discard. Such a technique can be used with any source of candidate terms, particularly, with the feedback term source. One of the arguments in favour of IQE is that humans can recognize expansion terms that are semantically related to the information they are seeking [159].

Harman [72] tested the effectiveness of IQE by using the Cranfield collection. Relevance feedback was used to produce a list of 20 potential expansion terms, which were selected using score functions 11,12,13,14. The potential expansion terms were then presented to the user for further selection. The user's selection was simulated by using only terms which appeared in at least one of the relevant unretrieved documents, on average 12 out of 20. The simulated IQE produced a performance improvement compared with relevance feedback expansion. In her experiments, 20 terms was deemed an appropriate number to provide for user selection: the results for the 'best' score function reached a peak performance after only 12 terms were added to the query, with performance slowly decreasing after that. The shape of the performance revealed the effectiveness of the 'best' score function: putting the most useful terms at the top of the ranking list, and adding terms beyond 12 top-ranked terms, tend to be less useful.

Efthimiadis [52] carried out a set of experiments for studying IQE. The score functions used in his experiments were $score_8(t)$, $score_9(t)$, $score_{11}(t)$, $score_{17}(t)$, $score_{20}(t)$ and

$$score_{27}(t) = \log\left(\frac{F_{\Xi+}(t) + c_t}{|\Xi^+| - F_{\Xi+}(t) + 1 - c_t} \Big/ \frac{F_D(t) - F_{\Xi+}(t) + c_t}{|D| - |\Xi^+| - F_D(t) + F_{\Xi+}(t) + 1 - c_t}\right),$$

where $c_t = \frac{F_D(t)}{|D|}$. $score_{27}(t)$ was initially given by Robertson [149] for estimating probabilities p_t and q_t and for modifying $rew_2(t)$ in the BIP method. Efthimiadis concluded from his experimental results: (i) very similar rankings are obtained from pairs $score_8(t)$ and $score_{27}(t)$, $score_9(t)$ and $score_{20}(t)$, $score_{11}(t)$ and $score_{17}(t)$; (ii) there are major differences in the term rankings between the first pair $score_8(t)$ and $score_{27}(t)$ and the second pair $score_9(t)$ and $score_{20}(t)$, but there is no significant difference in performance; (iii) the best performances are given by $score_9(t)$ and $score_{20}(t)$, followed by $score_{17}(t)$, $score_{27}(t)$, $score_8(t)$ and $score_{11}(t)$.

From the study of the behaviour of these score functions and the inspection of the term rankings obtained from his experimental results, Efthimiadis suggested a simpler ranking algorithm [52] for selecting good terms, called the *r_lohi* algorithm: rank terms according to their document frequencies concerning the sample set, i.e., $F_{\Xi+}(t)$; the ties are ranked according to their document frequencies concerning the collection, i.e., $F_D(t)$, from low-to-high frequency. Efthimiadis [53] carried out an additional set of experiments, and found: (i) the *r_lohi* algorithm and $score_{17}(t)$ have similar performances to that of $score_9(t)$ and $score_{20}(t)$; (ii) the concentration of user-preferred terms at the top of the ranking list obtained from these four score functions is rather high.

Magennis & Van Rijsbergen [122] performed a set of experiments using real users to determine the effectiveness of IQE with TREC collection WSJ. 20 top-ranked documents were displayed to users for relevance assessment, $score_8(t)$ was used to rank relevance feedback terms, and 20 terms were selected for query expansion. They concluded: (i) AQE using relevance feedback offers a marked improvement in retrieval performance; (ii) IQE by an experienced user offers a small further improvement over AQE; (iii) inexperienced users of IQE do not make a proper selection and fail to do better than AQE. Hancock-Beaulieu *et al.* [69] also experimentally studied IQE using real users, and found that it fails to show any significant improvement over AQE.

Ruthven [159] conducted a series of experiments to compare the retrieval effectiveness of IQE versus AQE. From his results he concluded that IQE has the potential to be an effective technique compared with AQE. He further pointed out, however, that the potential benefits of IQE may not be easy to achieve: users cannot identify good terms for effective query expansion; users cannot identify semantic relationships between the information needs and the possible expansion terms; users cannot identify which semantic relationships are going to attract more relevant documents; users cannot identify the effect of individual expansion terms on further retrieval. A similar conclusion was also drawn by Blocks *et al.* [11]: for IQE, users usually require more background information on how and why results are retrieved and, hence, it is difficult for users to use semantic relationships even when the system supports the recognition of the relationships. Also, Magennis & Van Rijsbergen [122] pointed out that IQE requires extra effort from users involving careful reasoning (decision making) and good strategy, and may not be appropriate for inexperienced users.

2.4 Other Methods

The issue of the power of discrimination of terms and the issue of term association (term dependence) are part of the core of IR and are strongly interrelated. We look at them from a variety of angles.

2.4.1 Discrimination Values of Terms

A well-know term discrimination method was developed by Salton *et al.* [169, 170, 171, 172, 236]. It can be briefly described as follows.

In the linear algebra methods, documents are represented by vectors in a vector space. In order to achieve a maximum possible separation between the individual document vectors, an average similarity between documents over the collection is considered. The similarity can be regarded as a measure of space density. If the average similarity is small, documents are widely separated in the space. Contrariwise, if the average similarity is large, the documents exhibit close proximity to one another.

The discrimination value of a term, in Salton *et al.*'s work, is a measure of the difference of the space densities before and after assignment of the term. If the term is a good discriminator, then the space after its removal will be more compact. If the term is a poor discriminator, then the removal results in a decrease in space density. A large number of terms are indifferent discriminators, that is, their assignments would not essentially affect the space density.

The calculation of discrimination values is normally very expensive. Many IR researchers have investigated how best to calculate discrimination values. A series of algorithms for the calculation and for improving execution times were successively proposed [10, 26, 39, 45, 55, 220]. However, this term discrimination method has been criticized because it does not exhibit well-substantiated theoretical properties [164].

In addition, it is worth mentioning that if relevant and non-relevant documents are not well-separated by their representations, our chances of increasing the separation by expanded queries are low, however the queries are represented. This has been investigated by Van Rijsbergen [204] and Sparck Jones [186]. For instance, the use of document clusters results in lower performances for the Inspect and Keen collections than for the Cranfield collection [93, 205, 211]. Since the objective of this thesis concentrates only on the study of effectiveness of query representation (reformulation), document representation will not be discussed further.

2.4.2 Document Frequencies of Terms

Salton *et al.* [169, 170, 171, 172] considered the relationship between the discrimination values of terms (according to their term discrimination method) and the document frequencies of terms. They study revealed a interesting fact that discrimination values of terms are closely related to the document frequencies of terms. The relationship can be summarized as follows:

- Terms with high document frequencies, i.e., $\frac{|D|}{10} < F_D(t)$, are poor discriminators. These terms are usually too general in nature, and their use would produce an unacceptable precision loss. The retrieval performance can be improved by including these poor terms in appropriate phrases.

- Terms with neither too high nor too low document frequencies, i.e., $\frac{|D|}{100} \leq F_D(t) \leq \frac{|D|}{10}$ are good discriminators. These terms can be used directly as indexing terms.

- Terms with low document frequencies, i.e., $F_D(t) < \frac{|D|}{100}$, are indifferent discriminators. These terms are so specific that they cannot retrieve an acceptable proportion of the relevant documents, and their use would depress the recall performance. The majority of terms are indifferent discriminators. Retrieval performance can be improved by incorporating these indifferent terms into appropriate thesaurus classes.

Thus, we can see that the term discrimination method gives criteria to the automatic indexing strategies, and that document frequencies of terms may be used as an approximation of the discrimination values of terms.

The relationship between term document frequencies and term discrimination values was also remarked by Sparck Jones [184]:

- Terms with high document frequencies are not very useful.

- Terms with medium document frequencies are quite useful.

- Terms with low document frequencies are likely to be useful but not as much as terms with medium document frequencies.

- Terms with very low document frequencies are useful in the sense that they are good indicators of relevance when they do appear.

Also, the study given by Biru *et al.* [10] suggested that:

- Terms with high-frequencies of occurrence only in relevant or only in non-relevant documents are good discriminators.

- Medium-frequency terms are not necessarily the best discriminators when relevance information is available.

- Low-frequency terms can have the greatest power of discrimination on relevance.

The findings were somewhat at variance with the findings presented by Salton *et al.*. In practice, a term with a very low document frequency may have a low discrimination value (simply because it does not occur in enough documents), but nevertheless can be a good relevance discriminator for the rare documents in which it does occur [153].

Salton & Buckley [164] pointed out, when considering term discrimination values, that the best terms for document representation should be those which are able to distinguish certain individual documents from the remainder of the collection. This implies that the best term should have a high term frequency but low document frequency, i.e., $f_d(t) \times idf_D(t)$, because it allows a term to be weighted according to not only its importance within the individual documents but also to its importance within the collection as a whole [171, 172].

2.4.3 Co-occurrence Frequencies of Terms

Since the late 1950s there has been a great deal of interest in statistically-oriented retrieval methods. Luhn [119, 120] was the first to suggest that the frequencies of occurrence of terms might be used to represent documents and queries. He pointed out that automatic

retrieval systems should be based on comparison of such representations: "*The more two representations agreed in given elements* (concepts, terms, etc.) *and their distributions, the higher would be the probability of their representing similar information.*" Luhn's idea inspired many IR researchers to devote their studies along this line.

Statistical association of terms, derived from the frequencies of co-occurrence of terms, has long been a major area of interest for IR researchers. The association has been widely applied to automatic thesaurus construction and query reformulation.

Maron & Kuhns [124] followed up on Luhn's idea with an investigation into indexing and searching using term co-occurrence frequencies. Several association measures were defined. Stiles [197] carried Maron & Kuhns' work further. He showed it is possible to successfully retrieve relevant documents using expansion terms selected using an association measure.

Lesk [113] progressed some of Stiles' work. Term association was computed by an association measure. The terms assumed to be associated were clustered into one class, and all the terms in the same class (called class-related terms) were added to the query if the class contained at least one query term. However, he had little success. He found that there is little consistency between a human-produced thesaurus and statistical term classifications, and that manually constructed thesauri were superior to automatically constructed ones. He explained that the classifications capture only terms whose association is purely local (to a specific collection) and do not reflect their general meanings. This is easily understood as the association indicates statistical relations of terms, rather than general meanings of terms. Statistical relations depend completely on the statistics of the collection.

The most extensive study of statistical term classifications was conducted by Sparck Jones [184, 186, 192, 193]. She clustered term classes based on term co-occurrence frequencies, and showed that the use of automatically generated term classifications can achieve a better retrieval performance than that obtained with unclassified terms alone. She explored experimentally many different classification strategies, and found their effects relatively similar. She claimed: (i) term classes represent topic clusters rather than synonym sets; (ii) term classification limits ambiguity, thus a match on two terms from the same class is very suggestive of which term meaning is present. However, her subsequent experimental results were not optimistic: retrieval performance was improved by term classification only in one small collection.

Sparck Jones [183, 194] also experimented with many query expansion methods by adding class-related terms to the original query. She concluded that a better retrieval performance can be obtained by means of automatic term classifications. In order to improve retrieval performance by query expansion, Sparck Jones [184, 192] suggested: (i) high-frequency terms are not clustered; (ii) low-frequency terms are clustered; (iii) strongly associated terms are also clustered.

Minker, Wilson & Zimmerman [129] evaluated retrieval performance obtained from the expanded queries by adding class-related terms to the original queries. They found however that the expanded queries are marginally useful and generally produce worse performance than the original queries. Their work did not confirm the findings of Sparck Jones. They showed that term classifications can be detrimental to retrieval effectiveness.

Usually, phrases are less ambiguous than single terms. Thus, a statistical term classification method may be used in conjunction with a syntactic phrase formation method. Lewis & Croft [114] tried such a combined method with the CACM collection, and found a small performance improvement using the expanded queries.

In the early 1970s, studies on query expansion were concentrated on single term classifi-

cations before the user submitted a query. Queries were expanded by adding all class-related terms. In order to calculate the term association, a document collection was represented by a term-term association matrix [167]. The association measure was then used to cluster terms by setting a threshold. Terms with association values greater than the threshold were clustered in the same class, equivalent to a thesaurus class.

However, some studies [47, 135] argued that clustering terms into classes and treating terms of the same class as equivalent is too naive to be useful. Indeed, as Sparck Jones [190] commented, "It was depressing that, after ten years' effort, we had not been able to get anything from classification." Further, as automatic term classification is very expensive, it is unsatisfactory to use it to construct a classification which ultimately will not work [186].

2.4.4 The Maximum Spanning Tree

The maximum spanning tree (MST) method is a term-term association structure, generated using an association measure of terms. The structure is simply a tree where each term is connected to at least one other term considered to be the most associated with it. The MST method was elaborated and discussed by Van Rijsbergen in [206, 207]. The proof of the optimization procedure for generating the MST can be found in [32]. One of the effective algorithms for generating the MST from an association measure can be found in [218].

Obviously, different association measures generate different MSTs. The association measures can be used to score and rank terms for selection. For each query term t_j, some studies [207, 210] suggested considering the following association measures:

$$
\begin{aligned}
association_1(t_i, t_j) &= p(\delta_i = 1, \delta_j = 1) - p(\delta_i = 1)p(\delta_j = 1), \\
association_2(t_i, t_j) &= \log \frac{p(\delta_i = 1, \delta_j = 1)}{p(\delta_i = 1)p(\delta_j = 1)}, \\
association_3(t_i, t_j) &= \frac{p(\delta_i = 1, \delta_j = 1)}{\left[p(\delta_i = 1)p(\delta_j = 1)\right]^{\frac{1}{2}}}, \\
association_4(t_i, t_j) &= \frac{p(\delta_i = 1, \delta_j = 1)}{\frac{1}{2}p(\delta_i = 1) + \frac{1}{2}p(\delta_j = 1)}, \\
association_5(t_i, t_j) &= \sum_{\delta_i, \delta_j = 1,0} p(\delta_i, \delta_j) \log \frac{p(\delta_i, \delta_j)}{p(\delta_i)p(\delta_j)}, \\
association_6(t_i, t_j) &= -\sum_{\delta_i, \delta_j = 1,0} p(\delta_i, \delta_j) \log p(\delta_i, \delta_j), \\
association_7(t_i, t_j) &= \frac{association_5(t_i, t_j)}{association_6(t_i, t_j)},
\end{aligned}
$$

where $association_1(t_i, t_j)$ was used by Maron & Kuhns [124] and $association_2(t_i, t_j)$ was given by Ivie [90]; both of which were employed for handling different situations. The estimation of $association_5(t_i, t_j)$, called EMIM, was described in detail by Van Rijsbergen [206, 207]. Also, studies given in [77, 181, 210] showed how an MST can be generated from the distribution of co-occurrences of terms in the collection, and how the MST can be used to expand a query.

Harper & Van Rijsbergen [77] performed a set of experiments with the Cranfield collection, using complete relevance information. The experiments were designed for the following cases:

(1) query expansion using the MST generated from $association_5(t_i, t_j)$, and reweighting of terms in the expanded queries using $rew_4(t)$;

(2) query terms reweighted using $rew_4(t)$ without query expansion;

(3) query terms reweighted using $rew_2(t)$ without query expansion.

The results showed that the performance of case-1 was much better than that of case-2, which in turn was much better than that of case-3.

Van Rijsbergen, Harper & Porter [210] carried out a set of experiments with three collections (Cranfield, UKCIS I and II). The size of the sample set, $|\Xi|$, was set to 10,20, and $rew_4(t)$ was used to reweight terms in the expanded queries. Query expansion was done with MSTs, each of which was generated from the association measures 1,3,4,5,6 listed above. Their experimental results showed that the MSTs give similar retrieval performances, even though the MST generated from $association_5(t_i, t_j)$ on the whole shows a slightly better performance than others.

Van Rijsbergen, Harper & Porter [210] carried out a further set of experiments, which were designed for the following cases:

(1) query expansion using the MST generated from $association_5(t_i, t_j)$, and reweighting of terms in the expanded queries using $rew_4(t)$;

(2) query expansion using $score_8(t)$, and reweighting of terms in the expanded queries using the same function, i.e., $rew_2(t)$.

The results showed the superiority of the performance of case-1 over case-2 on all three collections.

Smeaton & Van Rijsbergen [181] gave a set of experiments with collection NPL and with $|\Xi| = 10$. Their experiments were designed for the following cases:

(1) query expansion using the MST generated from $association_5(t_i, t_j)$, and reweighting of terms in the expanded queries using $rew_4(t)$;

(2) query expansion using the MST generated from $association_5(t_i, t_j)$, and reweighting of terms in the expanded queries using $rew_2(t)$;

(3) query terms reweighted using $rew_4(t)$ without query expansion.

The results showed that the performance of case-1 was better than that of case-2, and marginally better than that of case-3.

From the above series of experiments it can be seen that the association measures derived from information measures, such as $association_5(t_i, t_j)$, produced better retrieval performance than the others.

2.4.5 Association Measures

Peat & Willett [135] considered the limitations of using term co-occurrence data for query expansion. Their analysis was based on the following three association measures [167, 207]:

$$association_8(t_i, t_j) = \frac{F_D(t_i, t_j)}{\sqrt{F_D(t_i) \times F_D(t_j)}},$$

$$association_9(t_i, t_j) = \frac{2 \times F_D(t_i, t_j)}{F_D(t_i) + F_D(t_j)},$$

$$association_{10}(t_i, t_j) = \frac{F_D(t_i, t_j)}{F_D(t_i) + F_D(t_j) - F_D(t_i, t_j)}.$$

They claimed that query expansion based on term co-occurrence data is unlikely to bring about substantial improvement in retrieval performance. The basis for this claim is that the association measures 8,9,10 may have their maxima when $F_D(t_i) = F_D(t_j)$. They argued: (i) a term is likely to be strongly associated with those terms that have comparable document frequencies; (ii) query terms tend to have substantially higher document frequencies than other terms, thus, terms strongly associated with query terms are also likely to have high document frequencies; (iii) terms with high document frequencies tend to be poor at distinguishing relevant documents from non-relevant ones, hence, terms strongly associated with query terms are unlikely to be good discriminators.

We would dispute their claim from two points. First, obviously, the association measures 8,9,10 may reach their maxima if $F_D(t_i, t_j) = \min\{F_D(t_i), F_D(t_j)\}$, rather than $F_D(t_i) = F_D(t_j)$. Also, $F_D(t_i) = F_D(t_j)$ does not imply that terms t_i and t_j co-occur in the same documents, and hence cannot infer they are associated with one another. In particular, when $F_D(t_i, t_j) = 0$, the three measures equal zero, which indicates that terms t_i and t_j are not statistically associated with each other, even though $F_D(t_i) = F_D(t_j)$. Therefore, there is no link between the association of terms and their 'comparable document frequencies' given by $F_D(t_i) = F_D(t_j)$.

Second, it may not be true that the document frequencies of query terms are substantially higher than that of other terms. The way of deriving mean document frequencies of terms (for seven collections) in their work might be too crude. As we know, many terms occur only in one or two documents, and the proportion of such very infrequent terms is extremely large. This results in mean document frequencies, obtained over all terms (including very infrequent terms), being rather low. On the other hand, the user usually does not select intentionally very infrequent or very frequent terms to formulate his query: he selects terms, randomly in some sense, according to his judgement that these terms best describe his information needs. The difference in mean document frequencies between the query terms and other terms, given in their work, cannot indicate that query terms tend to have substantially higher document frequencies than other terms. In our own studies, we found the vast majority of query terms possess document frequencies less than $5\%|D|$, which fall into the range of medium-frequency terms. Thus, according to the studies given by Salton *et al.* [169, 170, 171, 172], most query terms should be viewed as good terms. Consequently, any terms closely associated with the context of the query tend to co-occur with most query terms, and hence are likely to be good discriminators.

Chung *et al.* [33] studied the application of association measures to term classification. They considered association measures 8,10 and three further association measures:

$$association_{11}(t_i, t_j) = \log\left(\frac{F_D(t_i, t_j)}{|D|} \Big/ \frac{F_D(t_i)}{|D|}\frac{F_D(t_j)}{|D|}\right),$$

$$association_{12}(t_i, t_j) = \frac{\sqrt{F_D(t_i, t_j)F_D(\bar{t}_i, \bar{t}_j)} - \sqrt{F_D(t_i, \bar{t}_j)F_D(t_i, \bar{t}_j)}}{\sqrt{F_D(t_i, t_j)F_D(\bar{t}_i, \bar{t}_j)} + \sqrt{F_D(t_i, \bar{t}_j)F_D(t_i, \bar{t}_j)}},$$

$$association_{13}(t_i, t_j) = \frac{|D|\left[F_D(t_i, t_j)F_D(\bar{t}_i, \bar{t}_j) - F_D(t_i, \bar{t}_j)F_D(\bar{t}_i, t_j)\right]^2}{F_D(t_i)F_D(\bar{t}_i)F_D(t_j)F_D(\bar{t}_j)},$$

where $association_{11}(t_i, t_j)$ is given in [106, 207], $association_{12}(t_i, t_j)$ can be found in [237] and $association_{13}(t_i, t_j)$ can be found in [85]. They analysed the relationships and evaluated the similarities of these measures. They concluded: (i) the most similar measures are $association_{11}(t_i, t_j)$ and $association_{12}(t_i, t_j)$, whereas others show quite similar behaviour only for terms with high document frequencies; (ii) the least affected by document frequencies is measure $association_{13}(t_i, t_j)$. They stated that it is necessary to select an association measure most appropriate for the application of query expansion because different measures may emphasize terms in a different range of the document frequencies. Kageura [95] also examined and evaluated the characteristics and performance of these association measures in the morphological analysis of Japanese kanji sequences.

Kim *et al.* [102] compared experimentally five association measures for query expansion (with 100 expansion terms). They considered association measures 8,9,10 and two further association measures:

$$\begin{aligned} association_{14}(t_i, t_j) &= \frac{1}{\log |D|} \log \frac{P_D(t_i, t_j)}{P_D(t_i) P_D(t_j)} = \frac{1}{\log |D|} \log \frac{F_D(t_i, t_j)/|D|}{\big(F_D(t_i)/|D|\big)\big(F_D(t_j)/|D|\big)} \\ &= \frac{1}{\log |D|} \log \frac{F_D(t_i, t_j)}{F_D(t_i) F_D(t_j)} |D|, \\ association_{15}(t_i, t_j) &= \frac{1}{2}\Big(P_D(t_i|t_j) + P_D(t_j|t_i)\Big) = \frac{1}{2}\Big(\frac{P_D(t_i, t_j)}{P_D(t_i)} + \frac{P_D(t_i, t_j)}{P_D(t_j)}\Big) \\ &= \frac{1}{2}\Big(\frac{F_D(t_i, t_j)}{F_D(t_i)} + \frac{F_D(t_i, t_j)}{F_D(t_j)}\Big). \end{aligned}$$

They called $association_{14}(t_i, t_j)$ normalized mutual information, and $association_{15}(t_i, t_j)$ average conditional probability. Their experimental results showed that performances obtained from association measures 8,9,10 are similar, and are better than those obtained from association measures 14,15.

However, some studies have shown that exploiting the frequencies of co-occurrence of terms in the collection has generally achieved little or no effect on average retrieval performance [129, 135, 181].

2.4.6 Thesaurus

The traditional thesaurus in an IR environment may be called a global thesaurus, and differs from the local thesaurus described by Attar & Fraenkel [6, 7]. A global thesaurus is constructed prior to the indexing process and is used to index both documents and queries; in contrast, a local thesaurus is constructed dynamically during query processing (using information obtained from the documents retrieved in response to a particular query) and is used to modify only that query [46].

Kristensen [104] used a manually-constructed thesaurus in a limited domain (economics and environment). Adding loosely-defined synonyms, related terms and narrower terms, resulted in a large improvement in average recall at the expense of a small drop in average precision.

Voorhees & Hou [216] used a general purpose thesaurus, WordNet [128], as a source of related terms, resulting in the improvement of some queries but the degradation of others. Voorhees [213] attempted to exploit the lexical-semantics contained within WordNet to disambiguate word senses, but retrieval performance became degraded. She [214] further investigated the effect of query expansion using WordNet. To reduce the possibility of expanding

the query with poor terms, expansion terms were selected by hand. A series of experiments were carried out with TREC collections, and results showed that the expansion is ineffective for long queries, but indeed improves performance for short queries.

The limitation of query expansion using WordNet is that most domain-specific relationships between terms are not found in WordNet [123]. Some past studies, [94, 141, 175] for instance, reported increased retrieval effectiveness from query expansion using collection-based thesauri. A collection-based thesaurus may incorporate domain-specific information as it is constructed from a set of documents relating to a specific domain.

Crouch [45, 46] proposed a method of constructing a global thesaurus based on a specific document clustering method. He supposed that the term discrimination method [169, 170, 171, 172, 236] provides a criterion for the formation of global thesauri: the thesaurus classes should consist of indifferent discriminators (i.e., terms with low document frequencies). He used document frequencies of terms as an approximation to the discrimination values of terms. A premise of his method is that terms in a thesaurus class should come from closely related documents, which implies that the document clusters themselves must be small and tight. One algorithm that produces clusters of this type is the complete link clustering algorithm. This algorithm has a stronger grouping criterion than the single link clustering algorithm [167, 207]. Once the document clusters have been established, the thesaurus classes can be constructed from the low frequency terms contained in those clusters. The strategies to generate a thesaurus class might be: (i) the intersection of all the low frequency terms in a cluster; (ii) the union of all the low frequency terms in a cluster; (iii) the top-ranked terms from the intersection of all the low frequency terms in a cluster. Experiments carried out by Crouch [46] showed that the best results are obtained by using strategy (i). Further experiments were carried out by Crouch & Yang [47], and the results indicated that his method can be produce useful thesauri, which substantially improves retrieval effectiveness. However, a major disadvantage of his method is that the construction of a thesaurus needs to be based on a document clustering method. Cluster creation and maintenance is time-consuming, especially for an effective cluster structure consisting of many small and tight document groups; when fast responses are important, as it is in modern on-line search environments, time efficiency is essential [163, 212]. Another problem with his method is that it involves many parameters to be specified by the user.

Mandala *et al.* [123] proposed a query expansion method using heterogeneous thesauri. The expansion terms were selected from three thesauri, a general purpose thesaurus (such as, WordNet), a co-occurrence-based automatically constructed thesaurus, and a predicate-argument-based automatically constructed thesaurus (i.e., term relations are gathered on the basis of linguistic relations [84]). The weighting of the expansion terms depended not only on the weights of the original terms, but also on the weights of those terms in each thesaurus. Experiments showed that use of the combined set of thesauri produces better performance than the use of only one type of thesaurus.

Generally, a thesaurus can be viewed as a recall improving device. The formation of query term (thesaurus) classes may be expected to retrieve more relevant documents because extra 'related' terms are added to the query when the thesaurus classes are assigned to the query instead of single terms. However, if terms included in a thesaurus class have high document frequencies, then the addition of these terms would be likely to lead to unacceptable losses in precision. For this reason, thesaurus classes should be formed only from those terms which have low document frequencies [236].

Many kinds of thesauri have been constructed, often tailored to specific topic areas. How-

ever, it is not easy to apply thesauri to practical IR, and there is no guarantee that a thesaurus tailored to a particular document collection can also be used with other collections. As a result, it is unlikely that reliable improvements in retrieval effectiveness over a variety of different collections can be obtained using thesauri [162].

It is, further, very difficult to use thesauri to build term classes which effectively capture semantic relationships between terms [236]. The construction of thesauri is also extremely time-consuming [135]. An alternative, possibly more practical and equally effective procedure, may be to use statistical methods. There has therefore been a great deal of interest in techniques for the automatic identification of statistical association of terms.

2.4.7 Stemming

The conflation of morphological variants of terms using stemming (suffixing) algorithms is one of the earliest techniques used in quantitative retrieval systems. The algorithms reduce different term variants to common stems (roots) that are assumed to refer to the same concepts. Two typical algorithms, the Lovins algorithm [118] and the Porter algorithm [137], have been widely employed. The Lovins algorithm simply removes the longest suffix of a term, whereas the Porter algorithm iteratively removes endings from a term according to a set of rules until no more can be removed. By reducing query and document terms to common stems, retrieval systems can achieve the effect of automatically expanding queries with morphological variants of the original query terms.

Harman [71] pointed out that query expansion by adding morphological variants of terms does not always improve retrieval effectiveness, but it could increase retrieval efficiency because the number of terms is reduced. Harman [72], after a series of experiments using the Cranfield collection, observed that the addition of term variants using the Lovins stemmer produces a significant decrement in performance. The performance decrement is somewhat smaller using the Porter algorithm. She attributed the decrement to many of the added term variants being not useful for retrieval and reducing precision. Also, Harman [73] used the Porter stemmer on the Cranfield, Medlars and CACM collections and found no significant improvement in retrieval performance.

Other experimental results were subsequently reported: Keen [100] showed that stemming can offer small average improvements in most situations but with great variation across queries, some being improved greatly, others being degraded; Krovetz [105] found that stemming is more useful for short queries and short documents; Hull's studies [86] demonstrated that stemming can produce consistent, though small, improvements in retrieval effectiveness over a large range of collections.

2.5 Summary

¶ The potential of AQR for improving retrieval performance has been extensively investigated by many IR researchers by analysing many different score and reweighting functions, and by trying a variety of sources of candidate terms. The investigations have shown that AQR is capable of producing large improvements, particularly, when AQR is performed alongside relevance feedback.

¶ The effectiveness of AQR depends on many factors. Some of the factors are: the weighting function for document terms and query terms; the similarity measure for ranking documents against the query; the features of the collection; the length of a query; the quality of the sample set; the score function for selecting expansion terms; the reweighting function for expanded query terms; the size of the sample set; the number of expansion terms.

¶ Using statistical relations for AQR is attractive since the term associations can be easily generated from the statistics of the sample documents or collection. In contrast, using lexical semantic relations as a source of related terms is normally very expensive to build and maintain, particularly for an extremely large collection.

¶ Many AQR experimental results demonstrate that it can be profitable to use an information measure as a device to construct a discrimination measure (score function) for selecting good terms. The superiority of information measure methods over other methods is apparent.

To automatically measure the power of discrimination of terms is a fundamental issue in IR. This issue has been a significant subject of interest among IR researchers since the early sixties. Many discrimination methods have successively been developed. Nevertheless, there is no widely recognized formal definition of what should characterize term discrimination information. Typically, studies in related literature are accompanied by discussions of the circumstance in which the discrimination of terms is essential. Such discussions are argued by concrete examples and appeals to intuition, or by some empirical formulae. While these informal discussions might be sufficient to convey some of the ideas that discrimination encompasses, however, they are inadequate for any more formal analysis. Indeed, the formal interpretation of term discrimination information is not simple.

A thorough investigation into the issue of the power of discrimination of terms for effective AQR is urgently needed. This thesis attempts new practices for defining term discrimination information as one, or more discrimination measures, which are derived from information measures.

Chapter 3

AQE Based on Directed Divergence

The purpose of this chapter is to study the application of the basic concept of directed divergence to automatic query expansion. In Section 3.1, we give terminology used to formulate formal methods proposed in this thesis. In Section 3.2, we intend to provide an accessible account of the meaning of information contained in a term. The rationale of applying logarithmic measure of information to measuring the amount of information in a term is interpreted. In Section 3.3, we look at the divergence measure more generally by putting forward some necessary criteria and hypotheses that underlie the methodology introduced in this thesis. In Section 3.4, we concentrate on investigating the relevance discrimination measure, which is a basis for the formal methods proposed in the thesis, based on directed divergence. In Section 3.5, we are concerned with the definition of the concept of the association of terms with the context of the query, which plays a central role in constructing a score function for query expansion. In Section 3.6, we describe the method of construction of the score function for judging good terms with respect to the query. In Section 3.7, we present a mathematical discussion on the estimation of the term probability distributions.

3.1 Terminology

First of all let us establish a consistent terminology for describing the concepts and the formal methods proposed in this thesis.

3.1.1 Representation of Objects

Let D, $|D| = N$, be a document collection. Let V, $|V| = n$, be the vocabulary of terms indexing documents of the collection. Let q be a query. A basic tool for the construction of an object space is provided by the notion of an n-tuple. An *n-tuple* $[w(t_1), w(t_2), ..., w(t_n)]$ is an array (or a $1 \times n$ matrix) of n symbols, $w(t_1), w(t_2), ..., w(t_n)$, which are called, respectively, the first component, the second component, and so on, up to the nth component of the n-tuple. The order in which the components of the n-tuple are written is of importance. The usefulness of n-tuples derives from the fact that they are convenient devices for representing documents of the collection and queries provided by users.

To represent an object $x = d \in D$ or $x = q$, an n-tuple $M_x = [w_x(t_1), w_x(t_2), ..., w_x(t_n)] = [w_x(t)]_{1 \times n}$ is used, in which $w_x(t)$ gives statistical information of term $t \in V$ concerning object

x. In IR, component $w_x(t)$ is called the *weight* of term t. Generally, weight $w_x(t)$ is considered to 'indicate' the importance of term t concerning object x. The terms with higher weights are regarded to 'contain' more information concerning object x than those with lower weights. Thus, to describe a natural language object, one needs only to state its representation M_x. With such a knowledge representation, the relationships between the objects will become clear when one deals with a specific quantitative retrieval model.

Let a statistical population $D_k \subseteq D$, where $k = 1, 2, ..., r$, be a set of documents. Similar to the representation of a document, to represent a set D_k an n-tuple $M_{D_k} = [w_{D_k}(t_1), w_{D_k}(t_2), ..., w_{D_k}(t_n)] = [w_{D_k}(t)]_{1\times n}$ is used, in which component $w_{D_k}(t)$ gives statistical information concerning term t, and is regarded to 'reflect' the importance of term t concerning set D_k. Particularly, when $D_k = \{d\}$, i.e., only one document d in set D_k, we denote $M_{D_k} = M_{\{d\}} = [w_{\{d\}}(t)]_{1\times n} = [w_d(t)]_{1\times n} = M_d$.

3.1.2 Probability Distributions

In this thesis, we confine ourselves to consider only a situation of discrete probability distributions. Let $\mathcal{P}_n$ be a *convex* set[1] of all finite multinomial (discrete) probability distributions defined on a probability space $(V, 2^V)$,

$$\mathcal{P}_n = \Big\{P = (p_1, p_2, ..., p_n) \mid p_j \geq 0 \ (j = 1, 2, ..., n) \text{ and } \sum_{j=1}^{n} p_j = 1\Big\}.$$

Each element $P \in \mathcal{P}_n$ may be considered an experiment having n possible outcomes with probabilities $p_1, p_2 ..., p_n$.

Let $P_{D_k}(t) \in \mathcal{P}_n$, derived from set D_k, where $k = 1, 2, ..., r$, be a term probability distribution over $(V, 2^V)$. Also, $P_{D_k}(t)$ can be considered as the *weight of importance* of terms $t \in V$ concerning set D_k. We will see that distribution $P_{D_k}(t)$ is a normalized form of representation M_{D_k}. Thus, we say that $P_{D_k}(t)$ defines set D_k, or say D_k is characterized by $P_{D_k}(t)$. In particular, for a given query q, we have two mutually exclusive and exhaustive events on D: $d \in D_1 = R$ characterized by $P_R(t)$ and $d \in D_2 = \bar{R}$ characterized by $P_{\bar{R}}(t)$.

For a given term $t \in V$, by saying that term t is drawn *from* set D_k we simply mean that t should always have distribution $P_{D_k}(t)$ which *defines* D_k, even though it does not occur in any document $d \in D_k$ (in this case, $P_{D_k}(t) = 0$). Thus, it is important to understand that term t should have domain V, rather than V^{D_k}, unless otherwise indicated.

Since $\log \varepsilon$ is not defined when $\varepsilon \leq 0$, it does not make sense to ask what happens to $\log \varepsilon$ as $\varepsilon \to 0^-$. However we can ask what is the situation when we have, for example, product $\varepsilon \log \varepsilon$ of ε multiplied by $\log \varepsilon$ as $\varepsilon \to 0^+$. Thus, in what follows, we shall use the following expressions:

$$0 \cdot \log 0 = \lim_{\varepsilon_1 \to 0^+} \varepsilon_1 \cdot \log \varepsilon_1 = 0,$$

$$0 \cdot \log\left(\frac{0}{0}\right) = \lim_{\varepsilon_1, \varepsilon_2 \to 0^+} \varepsilon_1 \cdot \log\left(\frac{\varepsilon_1}{\varepsilon_2}\right) = 0,$$

$$0 \cdot \log\left(\frac{0}{a}\right) = \lim_{\varepsilon \to 0^+} \varepsilon \cdot \log\left(\frac{\varepsilon}{a}\right) = 0,$$

[1] By the *convexity* of set $\mathcal{P}_n$ we mean here that $\lambda_1 P_{D_1}(t) + \lambda_2 P_{D_2}(t) + ... + \lambda_r P_{D_r}(t) = P_\Sigma(t) \in \mathcal{P}_n$ if $P_{D_k}(t) \in \mathcal{P}_n$ for $k = 1, 2, ..., r$ and $P_\lambda = \{\lambda_1, \lambda_2, ..., \lambda_r\}$ is an *a priori* probability distribution concerning r distributions $P_{D_1}(t), P_{D_2}(t), ..., P_{D_r}(t)$.

where a satisfies $0 < a < +\infty$. Also, in order to avoid meaningless expressions in the discussion in this thesis, we adopt the following notational conventions:

$$a \cdot \log\left(\frac{a}{0}\right) = \lim_{\varepsilon \to 0^+} a \cdot \log\left(\frac{a}{\varepsilon}\right) = +\infty,$$
$$(0-a) \cdot \log\left(\frac{0}{a}\right) = \lim_{\varepsilon \to 0^+} (\varepsilon - a) \cdot \log\left(\frac{\varepsilon}{a}\right) = +\infty,$$
$$(a-0) \cdot \log\left(\frac{a}{0}\right) = \lim_{\varepsilon \to 0^+} (a - \varepsilon) \cdot \log\left(\frac{a}{\varepsilon}\right) = +\infty.$$

For instance, for some $t' \in V$, if $P_{D_1}(t') = 0$ (but $P_{D_2}(t') \neq 0$), then the conventions that $\left(0 - P_{D_2}(t')\right) \log \frac{0}{P_{D_2}(t')} = +\infty$ are accepted.

A very important notion in this thesis is *absolute continuity*[2] of probability distribution $P_{D_1}(t)$ with respect to probability distribution $P_{D_2}(t)$, denoted $P_{D_1}(t) \ll P_{D_2}(t)$. It is generally necessary in applications of the divergence measure that the two probability distributions should satisfy condition(s) of absolute continuity. However, in practice, the condition(s) are usually not satisfied when we attempt to derive the probability distributions from the different sets of documents: because the above expressions may be encountered in the discrimination measures. We will discuss this problem in depth in the subsequent chapters.

3.1.3 Terms and Proposition

Although probability associated with random variables seems of more direct interest to most of us, the more fundamental idea is that of the probability of a proposition. All other types of probabilities are special cases on that basis, provided that the word proposition is taken in its most general sense. In the propositional notation, distribution $P(t)$ should be written as $P(\{\xi = t\})$, (here ξ is a discrete random variable), and $P(t)$ is best regarded as abbreviated notation. Hypothesis H will be regarded as a special type of proposition.

In this thesis, we assume that t is a proposition describing an *event*: term t occurs. Thus, proposition t is true if term t occurs; proposition t is false, or proposition $\bar{t}$ is true, if term t does not occur. Also, we assume that a term pair (t_i, t_j) expresses proposition $t_i \wedge t_j$ describing an event: terms t_i and t_j co-occur. Term pair (t_i, t_j) must satisfy the requirement that two terms t_i and t_j are distinct, i.e., $i \neq j$. Thus, proposition (t_i, t_j) is true if terms t_i and term t_j co-occur; proposition $(t_i, \bar{t}_j)$ is true if term t_i occurs but term t_j does not occur; and so forth. In what follows, we will use, for instance, 'term t occurs (or, terms t_i and t_j co-occur)' and 'proposition t is true (or, proposition (t_i, t_j) is true)', interchangeably.

In like manner, we assume that H_k is a proposition representing a *hypothesis* concerning a statistical non-empty document set D_k with some characteristic, where $k = 1, 2, ..., r$. Two types of characteristics considered in this thesis are *document relevance* and *term dependence*. We will indicate the concrete explication of hypotheses in a specific context.

3.1.4 Quantitative Aspect of Information

Information has both qualitative and quantitative aspects. Information theory is concerned only with the quantitative aspect [203].

[2]Probability distribution $P_{D_1}(t)$ is said to be *absolutely continuous* with respect to distribution $P_{D_2}(t)$, or in symbols, $P_{D_1}(t) \ll P_{D_2}(t)$, if $P_{D_1}(t) = 0$ whenever $P_{D_2}(t) = 0$.

Before introducing numbers and formulae, let us get our bearings by thinking about an interesting example. Suppose a policeman is talking to a lady. If the man says 'Your house is red', the lady may remain indifferent, unless her house is blue. If, however, he says 'Your house was robbed this morning', the lady's reaction will be very different.

According to our usual way of looking at information, the amount of information conveyed in an event should depend on the probability of the occurrence of the event. If we are told something that we already know, the probability before being told was already unity, the probability remains unity after being told. Then, the statement that the event will occur does not give much information. On the other hand, if we are told something that was almost improbable, the probability changes from a small value before being told to unity afterwards. The statement that the event will occur gives a good deal of information [66, 203].

Thus we can see that the amount of information is strongly connected to the amount of uncertainty. In fact, the information is equal to the removed uncertainty. Shannon [176] made the first consistent attempt towards the measurement of such difficult and abstract notions as information and uncertainty [68]. Shannon [176] introduces two important ideas in his mathematical theory of communication. The first idea is that information should be a statistical concept, that is, amount of information or, measure of information, should be defined in a technical sense, and it should not be confused with a semantic concept. The distribution of statistical frequency of symbols that make up a message must be considered before the notion can be discussed adequately. The second idea springs from the first one that, on the basis of the frequency distribution, there is an essentially unique function of probability distribution which measures the amount of information. The second of Shannon's ideas has been applied by Kullback and Leibler in [107]. Following them, speaking broadly, whenever we make observations, or conduct experiments, we seek information.

In the propositional notation, in probabilistic IR, the information contained in a certain term t should be interpreted as the amount of information received when we discover that proposition t is true. In other words, information in term t should be regarded as the amount gained when we observe that term t occurs in a document, or a set of documents. Similarly, the mutual information contained in a certain term pair (t_i, t_j) should be interpreted as the amount received when we discover that proposition (t_i, t_j) is true. In other words, the mutual information of terms t_i and t_j should be regarded as the amount gained when we observe that terms t_i and t_j co-occur.

Probability $P(t)$ is usually interpreted as the uncertainty concerning the occurrence of term t before an experiment is performed. The larger the probability that term t has, the less information that term t contains when it occurs. Based on the second of Shannon's ideas, the *amount of information* contained in term t, or the *gain in information* about term t, can be defined as

$$i(t) = -\log P(t),$$

which can also be considered as the uncertainty concerning the occurrence of term t before we observe that t appears in a document or a set of documents. In fact, if we decide to make the amount of information depend only on the probabilities of occurrence of terms, if we want it to be a decreasing function of $P(t)$, and if we insist on its having the additive property for probabilistically independent terms, then the expression $i(t)$ is the only possible definition [67, 176]. Any other function satisfying these properties must be proportional to $i(t)$.

The *conditional gain in information*, denoted by $i(t_j|t_i)$, is defined as the gain in infor-

mation when t_j occurs given t_i has occurred, provided that $P(t_i) > 0$, that is.

$$i(t_j|t_i) = -\log P(t_j|t_i) = -\log \frac{P(t_i, t_j)}{P(t_i)}.$$

If terms t_i and t_j are independent then $i(t_j|t_i) = i(t_j)$. That is, the occurrence of term t_i gives us no information concerning term t_j.

The amount of information may more specifically be called *an amount of probabilistic information.* This is, it is a statistical notion, rather than a semantic one.

The purpose of this thesis is to apply Shannon's two ideas to IR theory by interpretations of the notion of the amount of information contained in a given term or, term pair, (rather than in a message). Some basic concepts of information theory are introduced for constructing mathematical forms of discriminant measures for the selection of good terms for AQE. These concepts are closely related and share a number of simple properties. Thus, we will see that formal methods proposed are information-theoretic in nature, and that the measures of information of terms will be proportional to the degree of the uncertainty of the occurrence of terms.

3.2 Information Gain $I(P_R : P_{\bar{R}})$

The concept of *directed divergence*, which is what is generally called *information gain*, or in short *information*, by statisticians and communications engineers [92], is by now a familiar one for many IR researchers. A detailed account about it is given in [106], and an axiomatic characterization for it can be found in [145]. A general definition for the discrete case, in the context of IR, is written as follows.

3.2.1 Information Contained in a Term

The first step in the subject is to define what we mean by *information contained in a term.* In order to fix our ideas we will always imagine in this thesis that each term is related to two opposite hypotheses.

Let H_1 and H_2 be two opposite hypotheses (i.e., $H_2 = \bar{H}_1$, the complement of H_1) related to a certain term t. We ignore the specific meaning of H_1 and H_2 at the moment, and only know they are hypotheses concerning t.

Then, by the product axiom in probability theory,

$$P(H_k, t) = P(H_k|t)P(t) = P(t|H_k)P(H_k) \qquad (k = 1, 2).$$

A *likelihood* can be written as

$$P(t|H_k) = P(H_k|t)P(t)/P(H_k) \qquad (k = 1, 2).$$

Express them together in the *odds* form as a *likelihood ratio*

$$\frac{P(t|H_1)}{P(t|H_2)} = \frac{P(H_1|t)}{P(H_2|t)} \Big/ \frac{P(H_1)}{P(H_2)} = O(H_1|t)/O(H_1),$$

where $O(H_1|t)$ is the odds in favour of H_1 against H_2 given t, and $O(H_1)$ is the odds in favour of H_1 against H_2.

One can intuitively feel that term t may be statistically dependent on hypothesis H_k, in the sense that the probability of the occurrence of term t is affected by the knowledge that H_k was true. Denote the logarithm of the likelihood ratio by

$$i(H_1 : H_2|t) = \log \frac{P(t|H_1)}{P(t|H_2)} = \log \frac{P(H_1|t)}{P(H_2|t)} - \log \frac{P(H_1)}{P(H_2)},$$

which is a very central subject of this thesis. Let us now carefully examine $i(H_1 : H_2|t)$ to look at what insight it can give. As we know, $P(H_k)$ is the *a priori* probability of H_k, and $P(H_k|t)$ is the *a posteriori* probability of H_k given proposition t to be true, where $k = 1, 2$. Therefore, $i(H_1 : H_2|t)$ is a measure of the *difference* between the logarithm of the odds after the observation showing the occurrence of term t and that before the observation. This difference, which can be positive or negative, may be interpreted as the information gained from the observation. Consequently, $i(H_1 : H_2|t)$ measures the amount of information contained in term t in support of H_1 as opposed to H_2. The base of the logarithm in $i(H_1 : H_2|t)$ is immaterial. Throughout this thesis, logarithms are taken to base 2, unless otherwise specified. We will return to this measure and to a detailed discussion in Section 3.4.

An alternative well-known information measure is $i(H, t) = \log \frac{P(H,t)}{P(H)P(t)}$, which is the amount of information in hypothesis H concerning term t, and also the amount of information in term t concerning hypothesis H, in virtue of the symmetry in H and t, namely, it is the intersection of the information in H with the information in t. We can easily see that measure $i(H_1 : H_2|t)$ is closely related to measure $i(H, t)$. In fact, if $P(t) \neq 0$ then

$$\begin{aligned} i(H_1 : H_2|t) &= \log \frac{P(H_1|t)}{P(H_2|t)} - \log \frac{P(H_1)}{P(H_2)} = \log \frac{P(H_1|t)}{P(H_1)} - \log \frac{P(H_2|t)}{P(H_2)} \\ &= \log \frac{P(H_1, t)}{P(H_1)P(t)} - \log \frac{P(H_2, t)}{P(H_2)P(t)} = i(H_1, t) - i(H_2, t). \end{aligned}$$

From such an equality, we can give a different interpretation to $i(H_1 : H_2|t)$: it is a measure of the difference between the intersection of information in H_1 with information in t and that of information in H_2 with information in t. In other words, it is the difference in information about H_1 compared to H_2 provided by t. In applications, $i(H_1 : H_2|t)$ is a more intuitive and basic concept than the information measure $i(H, t)$ [67].

Also, we can see that measure $i(H_1 : H_2|t)$ is related to information measures $i(t|H)$:

$$i(H_1 : H_2|t) = \log \frac{P(t|H_1)}{P(t|H_2)} = \log P(t|H_1) - \log P(t|H_2) = i(t|H_2) - i(t|H_1).$$

This equality give us an alternative interpretation to $i(H_1 : H_2|t)$: it is a measure of the difference between the gain in information when t occurs given H_2 is true and of that given H_1 is true. In other words, it is the difference in information about t provided by H_2 and H_1, respectively.

3.2.2 Directed Divergence Measure

Now, for a given query q, let H_1 be the hypothesis that term t is drawn from the relevant document set R, and H_2 the hypothesis that term t is drawn from the non-relevant document set

$\bar{R}$. Assume that $P_R(t)$ and $P_{\bar{R}}(t)$ are two term probability distributions over the same probability space $(V, 2^V)$ under two opposite hypotheses H_1 and H_2, respectively. Then $P(t|H_1)$ and $P(t|H_2)$ are familiar expressions stating that term t follows distribution $P(t|H_1) = P_R(t)$ and $P(t|H_2) = P_{\bar{R}}(t)$, respectively. Thus, when H_1 is true, measure

$$i(H_1 : H_2|t) = \log \frac{P(t|H_1)}{P(t|H_2)} = \log \frac{P_R(t)}{P_{\bar{R}}(t)},$$

can be used to measure the amount of information contained in term t in accepting the relevant hypothesis H_1 rejecting the non-relevant hypothesis H_2, or more precisely, in favour of $P_R(t)$ against $P_{\bar{R}}(t)$, when t occurs. In this thesis, we will use 'in favour of H_1 against H_2' and 'in favour of $P_R(t)$ against $P_{\bar{R}}(t)$', interchangeably.

Let us now further assume that distribution $P_R(t)$ is absolutely continuous with respect to distribution $P_{\bar{R}}(t)$, i.e., $P_R(t) \ll P_{\bar{R}}(t)$. Then, the *expected information*, given H_1 was true, is defined by

$$I\big(P_R : P_{\bar{R}}\big) = \sum_{t \in V} P_R(t) \log \frac{P_R(t)}{P_{\bar{R}}(t)} = \sum_{t \in V} P(t|H_1) i(H_1 : H_2|t) = I(H_1 : H_2|H_1), \tag{3.1}$$

which can also be referred to as the expected gain in information in favour of $P_R(t)$ against $P_{\bar{R}}(t)$. Kullback and Leibler [107] regarded it as a measure of the *directed divergence*, which means that $I\big(P_R : P_{\bar{R}}\big)$ can be used to measure the *expected divergence* of distribution $P_{\bar{R}}(t)$ from distribution $P_R(t)$. In practical applications, $I\big(P_R : P_{\bar{R}}\big)$ can also be interpreted as the measure of the *expected difference* of the information contained in $P_R(t)$ and that contained in $P_{\bar{R}}(t)$ about $P_R(t)$.

One of the typical applications of directed divergence to IR theory can be found in the study described by Van Rijsbergen [206]. The study designed a term weighting method under an assumption that terms are not independently distributed with respect to each other. In his explorative study, the extent to which two terms t_i and t_j deviate from independence is measured by the directed divergence

$$\begin{aligned} I\big(\delta_i, \delta_j\big) &= I\big(P(\delta_i, \delta_j) : P(\delta_i)P(\delta_j)\big) = \sum_{\delta_i, \delta_j = 1,0} P(\delta_i, \delta_j) \log \frac{P(\delta_i, \delta_j)}{P(\delta_i)P(\delta_j)} \\ &= \sum_{\delta_i, \delta_j = 1,0} P\big((\delta_i, \delta_j)|H_1\big) i\big(H_1 : H_2|(\delta_i, \delta_j)\big), \end{aligned}$$

where variable $\delta = 1, 0$ indicates that term t occurs or, does not occur, respectively, under hypothesis H_1: δ_i and δ_j are dependent with a joint distribution $P(\delta_i, \delta_j)$, and hypothesis H_2: δ_i and δ_j are independent with the product of marginal distributions $P(\delta_i)$ and $P(\delta_j)$. Therefore, $I\big(\delta_i, \delta_j\big)$ is the expected information in a term pair (t_i, t_j) in favour of dependent hypothesis H_1 against independent hypothesis H_2.

Therefore, it is easily seen that expression $I\big(\delta_i, \delta_j\big)$, also called *expected mutual information* [67, 107, 176], is a special case of directed divergence. In the context of IR, it is usually used as a measure of the statistical dependence between terms t_i and t_j, that is, as the statistical amount of information in term t_i concerning term t_j, and vice versa. We will return to this important topic in Chapter 7.

An alternative application is the probability distribution model proposed by Wong & Yao [221]. Their study attempted to apply $I(p_q : p_d)$ as a divergence measure between $p_q(t)$ and

$p_d(t)$, where $p_q(t)$ and $p_d(t)$ were term probability distributions representing query q and document d, respectively. However, in the context of IR, one usually cannot directly use the directed divergence measure because $p_q(t)$ is not necessarily absolutely continuous with respect to $p_d(t)$. For solving such a problem, a divergence measure called *entropy increase*, closely related to directed divergence, was introduced in their work, and a similarity measure was then defined based on the entropy increase measure. We will also discuss this interesting issue in Chapters 5 and 6.

In what follows, we will concentrate on the discussion of the discrimination measure based on directed divergence, and its application to query expansion. Before doing so let us first look at *divergence measure* more generally.

3.3 On Divergence Measures

Probability distributions derived from the different document sets form a basis for the divergence measures. The divergence measures of the distributions form the basis for the derivation of the discrimination measures for the judgement of good terms.

3.3.1 Two Criteria

In a practical IR context, the first stage in measuring the power of discrimination of terms is to calculate the expected divergence, the expected information, followed by the derivation of the contributions made by individual terms to the expected divergence.

For the divergence measure to be appropriate, with respect to a number of terms, for judging potential good terms, the measure should satisfy some criteria.

Criterion 1: It should be possible to compare the *extent* to which each term contributes to the expected divergence.

For instance, term *desk* can apparently be discriminated from both terms *vegetable* and *lamp*, but it should be that *desk* differs more from *vegetable* than from *lamp*.

Suppose that, for each document set $D_k \subseteq D$, an effective method for estimating the term distribution can be devised, i.e., probability densities $P_{D_k}(t)$, for each term $t \in V$, can be obtained. Then, set D_k can be characterized and analysed by the densities. For a given term, it is likely that this term will have unequal densities related to the different sets. Therefore, the expected divergence of the distributions would be measured by means of the extent of the differences in the densities of individual terms. Simply stated, the individual pieces of divergence, each of which arises from some term, can be combined to obtain the expected divergence. Thus, an assumption that the extent to which individual terms contribute to the divergence can be measured, is needed. Under such an assumption, the contributions can be meaningfully combined to yield the expected divergence of the distributions.

It is seemingly simple but important that 'meanings' of terms and 'statistical quantities' of terms should be well distinguished in applying the divergence measure to AQE. In practice, confusion may arise from attempts to measure the divergence of the meanings of terms rather than of the statistical quantities referring to the terms. There can be no statistical comparability between term meanings. For instance, the meanings of terms *desk*, *lamp*, and *vegetable* cannot be compared by their contributions to the divergence, even though *desk* is

more different from *vegetable* than from *lamp* in meanings. In this thesis, for the sake of convenience, we will simply say 'the contributions made by terms to the divergence', but it should be understood that the discrimination is really in the sense of the statistical contributions (of the terms) to the divergence in our formal methods, rather than the meanings themselves.

Criterion 2: The *effect* of adding or removing terms unrelated to the classification should make no difference to the divergence.

By saying that terms are unrelated to a classification, we mean here they have an invariant probability density over the document sets considered. It is essential that a divergence measure should be independent of the addition or removal of terms which are unrelated to the classification.

Informally speaking, in IR, for a given query, a document collection is normally classified into two sets, R and $\bar{R}$. When a term has equal probability density over R and $\bar{R}$, it implies 'this term is not related to the relevance classification', i.e., this term does not provide any relevance information for classifying D into R or $\bar{R}$. The implication should be carefully distinguished from 'this term is not relevant to the query'. A term may be statistically closely related to the relevance classification when it is entirely non-relevant to the query, and vice versa.

For instance, let us consider a query 'What is tomorrow's computer?' in Example 1.4.4. Term *computer* may be an unrelated term in respect to the relevance classification for a collection catalogued as computing science. It is apparent that term *computer* would distribute rather uniformly over the whole collection, that it would therefore have an invariant density over any document sets, and would not provide any profitable information for the purpose of the relevance classification. However, everyone would agree that term *computer* is central to the query. This query may not be a good one from the IR point of view. It is fairly intuitive and understandable that the divergence measure should not be dependent on the addition or removal of those terms, such as, *computer*, which are unrelated to the relevance classification.

3.3.2 Two Hypotheses

Divergence has different applications in a variety of research areas, in particular, it has become a useful tool in designing discrimination measures in a probabilistic IR framework. Perhaps the usefulness of divergence can be best illustrated by the following specific situation.

In practice, it is desirable or necessary to consider the expected divergence of two distributions derived from the relevant and non-relevant document sets R and $\bar{R}$, respectively. This is because one would expect that the expected divergence, a statistical measure, may reveal some semantic relations between terms. A feasible scheme of capturing true semantic relations of complicated semantics is not yet available. But if the expected divergence in the form of the distributions can be obtained, and if the distributions can reflect statistical information concerning the sets, then one will know for sure that the expected divergence may meet one's needs. Underlying all the discussions given in this thesis is the following hypothesis.

Hypothesis 1: The expected divergence of two term probability distributions derived from the relevant and non-relevant document sets may be related to semantic relations between terms.

Many experimental results have shown that the difference between the distributions of terms in the relevant and non-relevant document sets can reflect some semantic relations between terms. One would expect, for instance, that the terms strongly associated with the query will occur more frequently in relevant documents than in non-relevant ones.

In IR, the idea that some terms are more important than others is in fact rather vague. It is almost impossible to rationally derive *a priori* weights of terms, which truly indicate the importance of terms, based on no empirical or observational information. Thus, the issue we should focus mainly on is the question: what forms of *a posteriori* weights of terms may be used in a relevance classification process for the purpose of effective retrieval? Generally, it is accepted that terms with higher power of discrimination should be considered more important. Statistically, terms which are thought of as having higher power of discrimination tend to contribute more to the expected divergence than others. The extent of the contributions that terms make may hence be used as a device for representing *a posteriori* weights to reflect the importance of terms. These statements can be formulated by the following hypothesis.

Hypothesis 2: The terms making a greater contribution to the expected divergence should be regarded as statistically conveying more valuable discrimination information, and therefore being more important than others.

According to the foregoing discussion, it appears that the terms with more concentrated distribution in one of sets R and $\bar{R}$, i.e., with greater variant probability densities within sets R and $\bar{R}$, would make more contribution to the expected divergence and, therefore, should be viewed as statistically containing more discrimination information.

In Section 3.5, we will focus mainly on a detailed account of the construction of score functions based on the measure of discrimination information for selecting good terms from the relevant sample documents. For this purpose let us give an in-depth investigation of the concept of discrimination information of terms.

3.4 Discrimination Measure $\mathbf{ifd}_I(t)$

This section concentrates on the definition of discrimination information, which is a basis for all methods proposed in this thesis. The estimation of discrimination information will be discussed in Section 3.7.

3.4.1 Definition of Discrimination Measure

As pointed out, for the selection of good terms, we have to measure the discrimination information contained in individual terms, i.e., to measure the extent of the contributions made by individual terms to the expected divergence (expected information).

Let us return to Eq.(3.1). The directed divergence can be expressed as a sum of the items,

$$I(P_R : P_{\bar{R}}) = \sum_{t \in V} \mathbf{ifd}_{I_{12}}(t),$$

where each item, in short $\mathbf{ifd}_I(t)$ when without confusion, can be written down

$$\mathbf{ifd}_I(t) = P_R(t) \log \frac{P_R(t)}{P_{\bar{R}}(t)} = P(t|H_1) i(H_1 : H_2|t).$$

It is remarkable that the likelihood ratio

$$\frac{P_R(t)}{P_{\bar{R}}(t)} = O(H_1|t)/O(H_1),$$

or, with Turing's appealing terminology, *Bayes factor*, is an intuitive and important concept in information theory. Turing introduces the expression 'Bayes factor *in favour of a hypothesis*'. Kullback [106] defined the logarithm of the Bayes factor,

$$i(H_1 : H_2|t) = \log\big(O(H_1|t)/O(H_1)\big),$$

as *the 'information for discrimination' in favour of* H_1 *against* H_2. Good [67] also gives a similar interpretation, he describes the logarithm of the Bayes factor as the *'weight of evidence' concerning* H_1 *as opposed to* H_2, provided by t (in this case, the occurrence of term t is thought of as a piece of evidence).

Consequently, the amount of information $i(H_1 : H_2|t)$ in $\mathbf{ifd}_I(t)$ can be viewed as the power of term t to discriminate two opposite relevance hypotheses H_1 and H_2. The magnitude of probability $P(t|H_1)$ in $\mathbf{ifd}_I(t)$ measures the significance of term t concerning relevant set R in determining the power of discrimination. Thus, quantity $\mathbf{ifd}_I(t)$ indicates the 'information for discrimination' for term t supporting relevant hypothesis H_1 but opposing non-relevant hypothesis H_2, and summation $I\big(P_R : P_{\bar{R}}\big)$ is the expectation of the discrimination information of terms over vocabulary V.

The above explains what we mean by the discrimination information of a given term. Thus, we can introduce a discrimination measure which computes the extent of the contributions made by individual terms to the expected discrimination information.

More generally, for two opposite hypotheses H_1 and H_2 related to a given term t drawn from sets R and $\bar{R}$, respectively, we can make the following formal definition.

Definition 3.4.1 Let $P_R(t) = P(t|H_1)$ and $P_{\bar{R}}(t) = P(t|H_2)$ be discrete probability distributions over $(V, 2^V)$, and derived from sets R and $\bar{R}$, respectively. Assume that $P_R(t) \ll P_{\bar{R}}(t)$ when $t \in V$. The information in term t for discrimination on hypotheses H_1 and H_2 is defined by

$$\mathbf{ifd}_{I_{12}}(t) = P_R(t) \log \frac{P_R(t)}{P_{\bar{R}}(t)} = P(t|H_1) \cdot i(H_1 : H_2|t) \qquad (t \in V),$$

which is referred as to the (relevance) *discrimination measure* of terms, and $i(H_1 : H_2|t)$ the (relevance) *discrimination factor* of terms.

Notice that the directed divergence is information-theoretic, its units are information units (*bits*). Also, since $P_R(t) \ll P_{\bar{R}}(t)$ for every term $t \in V$, its individual items satisfy $\mathbf{ifd}_I(t) < +\infty$, and thus the items always exist and are comparable (i.e., it satisfies Criterion 1).

Let us further consider the situation where term t occurs in both sets R and $\bar{R}$ with an equal probability $P_R(t) = P_{\bar{R}}(t) \neq 0$. As pointed out above it has $\mathbf{ifd}_I(t) = 0$, namely, the contribution made by term t to summation $I\big(P_R : P_{\bar{R}}\big)$ will be zero. Therefore, the directed divergence possesses the property that it is independent of those terms which are unrelated to the relevance classification (i.e., it satisfies Criterion 2). From this property we can see that the discriminant measure $\mathbf{ifd}_I(t)$ emphasizes the importance of the terms that have variant probabilities within sets R and $\bar{R}$, and removes the dependence on the terms that have invariant probabilities in the sets.

It is shown that $I(P_R : P_{\bar{R}}) \geq 0$, with equality if and only if $P_R(t) = P_{\bar{R}}(t)$ for all $t \in V$. This property tells us that, in the expectation, the discrimination information obtained from all terms is positive. There is no the expected discrimination information if the term distributions are identical under the hypotheses.

3.4.2 Interpretation of Discrimination Measure

The interpretation of the discrimination information depends on a specific application. For instance, when the hypotheses involve the relevance of documents to the query, then the discrimination information is explained as the relevance discrimination as given in Definition 3.4.1. In this case we have the following interpretation (notice that $\mathbf{ifd}_I(t)$ can be positive or negative):

- ☞ If $P_R(t) = P_{\bar{R}}(t)$, then the discrimination factor $i(H_1 : H_2|t) = 0$, and term t gives us no discrimination information about the relevance classification, and the corresponding quantity $\mathbf{ifd}_I(t) = 0$.
- ☞ If $P_R(t) > P_{\bar{R}}(t)$, then the discrimination factor $i(H_1 : H_2|t) > 0$, and term t contains positive information in support of the relevant hypothesis H_1; thus the discrimination measure indicates that term t contributes quantity $\mathbf{ifd}_I(t) = |\mathbf{ifd}_I(t)|$ for supporting H_1.
- ☞ If $P_R(t) < P_{\bar{R}}(t)$, then the discrimination factor $i(H_1 : H_2|t) < 0$, and term t contains negative information in support of the relevant hypothesis H_1; thus the discrimination measure indicates that term t contributes quantity $\mathbf{ifd}_I(t) = -|\mathbf{ifd}_I(t)|$ for supporting H_1.

However, when the hypotheses are concerned with the statistical dependence of terms, then the discrimination information should be interpreted as the dependence discrimination. We will discuss such an application in Chapter 7.

3.4.3 About Absolute Continuity

Notice that in order to speak of the discrimination information of terms, we must consider distributions $P_R(t)$ and $P_{\bar{R}}(t)$ to be defined on the *same* probability space V, and assume $P_R(t) \ll P_{\bar{R}}(t)$ for all terms $t \in V$ (i.e., $P_R(t) = 0$ whenever $P_{\bar{R}}(t) = 0$). Consequently, with the notational conventions: $0 \cdot \log\left(\frac{0}{a}\right) = 0$ and $0 \cdot \log\left(\frac{0}{0}\right) = 0$, measure $\mathbf{ifd}_I(t) \neq \infty$ always holds, and summation $I(P_R : P_{\bar{R}})$ exists.

For example, let $V = \{t_1, t_2, t_3\}$, and let $P_R(t_1) = P_R(t_2) = \frac{1}{2}$ and $P_{\bar{R}}(t_1) = P_{\bar{R}}(t_2) = P_{\bar{R}}(t_3) = \frac{1}{3}$. Then $P_R(t)$ is absolutely continuous with respect to $P_{\bar{R}}(t)$, that is, $P_R(t) = 0$ whenever $P_{\bar{R}}(t) = 0$. However, $P_{\bar{R}}(t)$ is not absolutely continuous with respect to $P_R(t)$ since $P_{\bar{R}}(t_3) = \frac{1}{3}$ when $P_R(t_3) = 0$. More precisely, from the definition, we have

$$
\begin{aligned}
I(P_R : P_{\bar{R}}) &= \sum_{t \in V} P_R(t) \log \frac{P_R(t)}{P_{\bar{R}}(t)} \\
&= \frac{1}{2} \log \frac{\frac{1}{2}}{\frac{1}{3}} + \frac{1}{2} \log \frac{\frac{1}{2}}{\frac{1}{3}} + 0 \log \frac{0}{\frac{1}{3}} = \log \frac{3}{2} + 0 = \log \frac{3}{2},
\end{aligned}
$$

$$I(P_{\bar{R}} : P_R) = \sum_{t \in V} P_{\bar{R}}(t) \log \frac{P_{\bar{R}}(t)}{P_R(t)}$$
$$= \frac{1}{3} \log \frac{\frac{1}{3}}{\frac{1}{2}} + \frac{1}{3} \log \frac{\frac{1}{3}}{\frac{1}{2}} + \frac{1}{3} \log \frac{\frac{1}{3}}{0} = \frac{2}{3} \log \frac{2}{3} + (+\infty) = +\infty.$$

Let us now further examine requirement $P_R(t) \ll P_{\bar{R}}(t)$ for $I(P_R : P_{\bar{R}})$ to look at what insight it can give. Consider a more general situation. Suppose that

$$P_R(t) \begin{cases} > 0 & t \in V^R \\ = 0 & t \in V - V^R \end{cases} \quad \text{and} \quad P_{\bar{R}}(t) \begin{cases} > 0 & t \in V^{\bar{R}} \\ = 0 & t \in V - V^{\bar{R}} \end{cases}$$

are two probability distributions over $(V, 2^V)$. Notice that

$$V = V^R \cup V^{\bar{R}} = (V^R - V^{\bar{R}}) \cup (V^{\bar{R}} - V^R) \cup (V^R \cap V^{\bar{R}}).$$

Thus, consider the following three cases:

case-A: when $t \in V^R$ but $t \notin V^{\bar{R}}$, it has $i(H_1 : H_2|t) = \log \frac{P_R(t)}{0} = +\infty$;

case-B: when $t \notin V^R$ but $t \in V^{\bar{R}}$, it has $i(H_1 : H_2|t) = \log \frac{0}{P_{\bar{R}}(t)} = -\infty$;

case-C: when $t \in V^R$ and $t \in V^{\bar{R}}$, it has $-\infty < i(H_1 : H_2|t) = \log \frac{P_R(t)}{P_{\bar{R}}(t)} < +\infty$.

It is easy to imagine that if terms occur only in some relevant documents but never in any non-relevant documents (case-A), then they should be considered to be associated with the query and added into the query. Conversely, if terms occur only in some non-relevant documents but never in any relevant documents (case-B), then they should be viewed as not associated with the query and discarded immediately. Thus, it is clear that case-A and case-B are not our main concerns.

The problem that mainly concerns us here is those terms that occur in both relevant and non-relevant documents (case-C), i.e., $t \in V^R \cap V^{\bar{R}}$. Notice that we here suppose $P_R(t) > 0$ if $t \in V^R$, and $P_{\bar{R}}(t) > 0$ if $t \in V^{\bar{R}}$. Notice also that the probability distributions $P_R(t)$ and $P_{\bar{R}}(t)$ are generally not disjoint (see a footnote given in Section 5.1), i.e., $V^R \cap V^{\bar{R}} \neq \emptyset$. Thus, if we assume $P_R(t) \ll P_{\bar{R}}(t)$ then, according to the definition of absolute continuity, it must satisfy $V^R \subseteq V^{\bar{R}}$. Consequently, the assumption of absolute continuity actually implies that the terms considered in the current discrimination method should be only those which satisfy $t \in V^R \subseteq V^{\bar{R}}$ and $P_R(t) \cdot P_{\bar{R}}(t) \neq 0$.

We proceed with our examination of applications of the discrimination information by considering some special situations. One such situation we shall discuss in the next section is to define the concept of the statistical association of a given term with the query, which plays a central role in constructing a score function for query expansion.

3.5 Association Function $atq_I(t, q)$

The discrimination measure $\mathbf{ifd}_I(t)$ corresponding to sets R and $\bar{R}$ has no direct implications for retrieval since the sets in question are the object of the retrieval. In practice, there is

no *a priori* way to obtain the term discrimination information $\mathbf{ifd}_I(t)$. This kind of circularity however suggests a strong analogy to the relevance or, pseudo-relevance, feedback process.

In order to give a systematic investigation into the derivation of the association of terms with the query through a feedback process, in this thesis, we start our discussion with some necessary assumptions.

3.5.1 Three Assumptions

It has been stated that query expansion is a technique for enhancement of queries. In other words, the technique is concerned with *good terms* with respect to the query about which we have postulated certain properties. The question immediately arises: how do we formulate postulates for good terms? This is done by introducing an assumption about good terms:

Assumption 1: In a general probabilistic IR framework, a good term should be defined as statistically informative; a good term with respect to a given query should be defined as statistically informative and strongly associated with the context of the query.

The *informativeness* of terms can generally be measured by means of some information measure(s) offered in, such as, information theory. Whereas the *association* of terms with the query may be designed through some technique of composition/combination of the information measure(s).

To simplify the discussion, an alternative assumption about good terms would be:

Assumption 2: Query terms, which occur in some relevant documents, should be regarded as good ones with respect to the query.

In other words, query terms which do not index any relevant documents should not be treated as good terms. It is interesting to notice that a term that is not considered a good one is not necessarily a poor term, but means we are unable to tell that it is a good or poor term.

An important notion in developing any feedback technique is the sample set Ξ since it is the starting point of all feedback methods. Before designing a score function to judge good terms with respect to the query through a feedback process, we must first decide the sample set. We mention here the difficulties that arise in obtaining an effective sample set through an initial retrieval. If the sample set is effective as desired, it is likely that all relevance information will be contained. To be able to talk about our subjects explicitly, we shall state an assumption below.

Assumption 3: The sample set used to establish the statistical association functions in a feedback procedure is effective if it contains all important statistical information of relevance to the query.

However, in the more practical situation, where the sample set is poorer than we might like, the problem of choice of the sample set is a pressing one. Unfortunately, solutions remain unsatisfactory even though much effort has been made. It is beyond the scope of this thesis to discuss the problem of choice of the sample set in greater detail, and will be treated as an important subject for further study. It should be emphasized that a query expansion method itself cannot compensate for a poor sample set.

3.5.2 Generalized Association Hypothesis

The Association Hypothesis due to Van Rijsbergen [207] (p.134) is an important underlying hypothesis theoretically, which we now write down as follows.

> *If an index term is good at distinguish relevant from non-relevant documents then any closely associated index term is also likely to be good at this.*

Some researchers, [230] for instance, have questioned the correctness of the Association Hypothesis. Let us now return to Example 1.4.6, an interesting example taken from [230] (p.36). Two phrases *DNA profile* and *DNA sequence* in [230] are viewed as nearly synonymous, and related to the query term *DNA*. It seems that *DNA profile* is good at distinguish relevant from non-relevant documents, but *DNA sequence* is not. A question arises: 'How can we explain such a phenomenon with the Association Hypothesis?'

Generally, in practice, it is very hard for a single term to achieve the difficult task of separating relevant documents from non-relevant ones for an extremely large collection. In fact, if there really existed such a (unique) term that could identify all relevant documents, then this term would contain complete discrimination information about relevance. Thus, we would not need query expansion, and it should be true that any term associated strongly with this term was likely to possess the ability of discrimination of relevance. However, a single term, even a group of terms (such as, all good query terms), can usually provide very limited discrimination information about relevance. That is the reason why we need to explore discrimination measures for capturing the power of discrimination of other terms so as to obtain more discrimination information for enhancing the query and improving the retrieval performance. The problem does not lie in the Association Hypothesis itself, but in it being very difficult to find a single term which satisfies the condition of the Association Hypothesis. As we know, theoretically, the conclusion(s) of a hypothesis holds only when its condition(s) can be satisfied.

Very often, two terms t_1 and t_2 refer to the same concept c but serve for different specific situations (see Section 1.2). Such situations might be American English and British English, or might be terms used by authors of literature and by authors of medicine, and so forth. In these cases, terms t_1 and t_2 are called *conditional synonyms* concerning concept c. When retrieval is performed over a collection that is based in some specific situation, the choice of term from a group of conditional synonyms for describing concept c would directly affect the retrieval performance. A consistent choice (for both relevant documents and the query) would increase the match and improve retrieval performance, otherwise, the match would decrease and the performance would decline. Clearly, conditional synonyms may not be statistically associated with one another. This is because terms t_1 and t_2 tend not to co-occur to describe concept c under the specific situation. This point is very important to understand basic concepts (such as, the Association Hypothesis) in IR, and to establish an effective IR system.

We are now ready to analyse the above example and answer the above question. Notice that *DNA profile* can be thought of as two terms, as can *DNA sequence*. Notice also that terms *profile* and *sequence* are neither synonyms, nor conditional synonyms, but they can be thought of as associated conditionally with term *DNA*. Obviously, term *DNA* itself in the above query cannot completely distinguish the relevant documents from non-relevant ones, even though it might have relatively higher power of discrimination than others. It is possible that phrase *DNA profile* enhances the power of discrimination, while phrase *DNA sequence* weakens the power of discrimination. This is because, when terms *profile* and *sequence* are

combined with term *DNA* to form phrases, they are normally used for different situations: the former is used by journalists reporting crime events and the latter is for scientific work in biology and medicine. The feature of the use determines that phrases *DNA profile* and *DNA sequence* would not be associated with one another. On the other hand, even if there really exists such a term or phrase closely associated with phrase *DNA profile*, we remain unable to ensure that it possesses the ability of discrimination on relevance since phrase *DNA profile* is weak in the ability (that is, the phrase does not satisfy the condition of the Association Hypothesis).

Also, it is clear that the Association Hypothesis is a very statistical concept. Therefore, not surprisingly, exceptions might happen in some special situations. Mathematically, we can never assert a statistical hypothesis wrong when some exception happens unless a number of observations/experiments are able to prove the assertion.

The Association Hypothesis [207] derives its importance from the fact that it is an underlying basis for the following more general hypothesis which we call the *Generalized Association Hypothesis*.

Hypothesis 3: If a group of terms combinatorially possesses high power of discrimination on relevance then any term associated closely with the group of terms as a whole is also likely to possess higher power of discrimination on relevance.

Notice that in Hypothesis 3 we do not make any restriction on what the group of terms should be. In the case of query expansion, the group of terms can refer to all good query terms.

Comparing Hypothesis 3 with the Association Hypothesis [207], it is clearly seen that there is one important different point between them: we are talking about the association of a given term closely with a group of terms as a whole, rather than with a single term. This generalization is necessary for almost all methods of query expansion. In fact, users usually describe their queries with more than one term. An expansion term should be associated with the context of the query, that is, with all good query terms, rather than only one of them. For instance, a term that is associated simultaneously with a group of terms *DNA*, *test*, *trial* and *criminal* should have a higher power of discrimination than another that is associated only with *DNA*. The former should be considered as being more strongly associated with the context of the query than the latter.

3.5.3 Association Function

Let the sample set Ξ consist of the top-ranked documents obtained from the previous search iteration. Based on the user's opinion about relevance, we may define two mutually exclusive and exhaustive events on set Ξ by assuming that $\Xi^+ \neq \emptyset$ is the set of top relevant sample documents, and that Ξ^- is the set of top non-relevant sample documents.

As we know, in practice, it is unlikely that every relevance feedback terms would contain information related closely to the query. Our aim is to judge which feedback terms are good ones with respect to the query. This may be achieved by estimating the statistical association of terms with the context of the query. That is, we assume that set V^{Ξ^+} constitutes a source of *candidate* terms, and select good terms among them (i.e., $S^q \subset V^{\Xi^+}$).

How can we directly estimate the association? We do not know. However, if the relevant sample set Ξ^+ is effective, that is, all important relevant information pertaining to the query is

contained in Ξ^+, then it is natural and reasonable for us to derive the association by drawing 'useful information' from the set.

What is the useful information? For a given term $t \in V^{\Xi^+}$, a piece of useful information would be: the amount of information contained in term t for discriminating on relevance. The discrimination measure $\mathbf{ifd}_I(t)$ can be invoked to measure the amount. In order to 'explain' why $\mathbf{ifd}_I(t)$ can provide a piece of useful information, we need only adopt an assumption (stated rather informally): The statement, 'the extent of the association of term t with the context of the query' can be restated as, 'the power of the discrimination of term t in favour of relevant hypothesis H_1 against non-relevant hypothesis H_2'.

The above discussion may already answer the question: what do we mean by the statement that term t is associated with the context of the query? The concept of the discrimination information of terms derives its importance from the fact that it provides a means to define the concept of the association, which we define formally as follows.

Definition 3.5.1 Let $P_{\Xi+}(t)$ and $P_D(t)$ be discrete probability distributions over $(V, 2^V)$, and derived from sets Ξ^+ and D, respectively. Assume that $P_{\Xi+}(t) \ll P_D(t)$ when $t \in V$. The association of term t with query q, denoted by $atq_I(t,q)$, is defined as

$$atq_I(t,q) = \mathcal{Q}(t) \cdot \mathbf{ifd}_I(t) = \mathcal{Q}(t) P_{\Xi+}(t) \log \frac{P_{\Xi+}(t)}{P_D(t)} \qquad (t \in V),$$

where $\mathcal{Q}(t) \geq 0$ measures the significance of terms $t \in V$ concerning query q.

With function $\mathcal{Q}(t)$, the statistical information contained in query terms can be effectively incorporated into the association function $atq_I(t,q)$.

We will see shortly that the association function provides a convenient way of combining miscellaneous pieces of evidence into the association score of terms. The various statistical clues — query term weights, document term weights, term importance concerning the relevant sample set, term specificity concerning the collection, the discrimination information of terms, etc. — as important factors, can be considered for constructing the score functions. Once the score functions have been constructed, the desired association values for the individual terms can be derived from each of them and terms can be sorted in accordance with the values.

3.6 Score Function $score_I(t)$

A more detailed discussion on selection of good terms is offered in this section. The discussion bases directly on the concept of the association given in Definition 3.5.1.

3.6.1 Relevance Feedback Process

The relevance feedback can be easily implemented by using information display techniques to establish communication between system and the user: a set of top-ranked documents can be graphically displayed for the user, and screen pointers can be used to designate some of the top-ranked documents as relevant to his information needs. The relevance information is then further used by the system to produce a modified feedback query.

In the $\mathcal{IFD}$ system, the sets of all relevant or non-relevant documents used in measure $\mathbf{ifd}_I(t)$ are replaced by the set of known relevant documents and the collection of all documents, respectively. That is, we use $P_{\Xi+}(t)$ instead of $P_R(t)$, and $P_D(t)$ instead of $P_{\bar{R}}(t)$

(since $\bar{R} = D - R \approx D$ because the size of R is generally very small so that R is negligible compared with that of D). This is equivalent to stating that $P_{\bar{R}}(t) \approx P_D(t)$ does not vary from query to query. Thus, H_1 may be the hypothesis that term t is drawn from set Ξ^+ defined by distribution $P_{\Xi+}(t)$, and H_2 is the hypothesis that term t is drawn from collection D defined by distribution $P_D(t)$.

Now consider the activity of $\mathcal{I}f\mathcal{D}$ as a decision procedure: whether or not a candidate term $t \in V^{\Xi+}$ should become a selected one with respect to the query. The decision depends on the sorting of the candidate terms, and ultimately, on the extent of the association of the candidate terms with the query.

Suppose that the statistical frequency data $f_{\Xi+}(t)$ and $f_D(t)$ for all terms $t \in V$ has been given. Then both sets Ξ^+ and D can be characterized by the probability distributions $P_{\Xi+}(t)$ and $P_D(t)$ which are estimated by using the frequency data. Consequently, the discrimination information of individual terms can be examined based on the estimates.

More precisely, suppose that the term probability distributions have the form (see Subsection 3.7.4):

$$P_{\Xi+}(t)\begin{cases} > 0 & t \in V^{\Xi+} \\ = 0 & t \in V - V^{\Xi+} \end{cases} \quad \text{and} \quad P_D(t)\begin{cases} > 0 & t \in V^{\Xi+} \\ \geq 0 & t \in V - V^{\Xi+}. \end{cases}$$

Obviously, $V^{\Xi+} \subseteq V$ and thus $P_{\Xi+}(t) \ll P_D(t)$ for all terms $t \in V$. Therefore, we can directly apply $I(P_{\Xi+} : P_D)$ to a query expansion procedure.

Next, for each term $t \in V$, there are two probability densities $P(t|H_1) = P_{\Xi+}(t)$ and $P(t|H_2) = P_D(t)$. In view of the interpretation of the discrimination measure $\mathbf{ifd}_I(t)$, the concept of the association of term t with the context of the query can be introduced, and then an association score function can be constructed. Consequently, each term can be assigned a score and be sorted in order to compare with others for the selection.

Obviously, it has $|S^q| < |V^{\Xi+}|$; we need not construct the score functions, otherwise. In practice, the number of selected terms is very much less than the total number of candidate terms indexing the relevant sample documents.

It is worth pointing out that we do not deal with the situation where $\Xi^+ = \emptyset$, that is, where there is no positive relevance information available and all documents in the sample set Ξ are assessed to be non-relevant. In this case, the user should be required to renew his query in order to produce an effective sample set Ξ satisfying $|\Xi^+| > 0$.

3.6.2 A General Form

After the discussion of the concept of the association, we are ready to tackle the problem of the construction of the association score function, which uses the ideas we met in previous sections. In particular, with a direct use of the Definition 3.5.1, the score function given below is rather intuitive and simple.

There may be many ways to construct the score function based on the association function by the different methods of estimating $\mathcal{Q}(t)$, $P_{\Xi+}(t)$ and $P_D(t)$. One of the methods we show in this subsection considers just the estimation of function $\mathcal{Q}(t)$. The estimation of distributions $P_{\Xi+}(t)$ and $P_D(t)$ will be discussed in the next section.

To begin with, consider a query q. Assume that q is initially represented as a matrix $M_q = [w_q(t)]_{1\times n}$, where weight $w_q(t)$ (satisfying $w_q(t) > 0$ when $t \in V^q$) indicates the

importance of term $t \in V$ in representing query q. As we pointed out, function $\mathcal{Q}(t)$ should be able to reflect the significance of terms $t \in V$ concerning query q.

For this purpose, let κ_1 and κ_2 be two constants satisfying $0 \leq \kappa_2 \leq \kappa_1 \leq \min_{t \in V^q} \{w_q(t)\}$. Define

$$\mathcal{Q}(t) = \begin{cases} w_q(t) & \text{when } t \in V \cap V^q \\ \kappa_1 & \text{when } t \in V^{\Xi+} - V^q \\ \kappa_2 & \text{when } t \in V - V^{\Xi+} - V^q. \end{cases}$$

That is,

- if term t is a query term then it is assigned a true weight $w_q(t)$;
- if term t is not a query term but appears in at least one relevant sample document then it is assigned a *stronger* 'fictitious' weight κ_1;
- if term t is not a query term and never appears in any relevant sample documents then it is assigned a *weaker* 'fictitious' weight κ_2;
- if term t is a query term but does not index any documents in the collection then it is discarded immediately.

Generally, the setting of the fictitious weights should be dependent on a specific system, and normally not greater than the minimum query term weight. Also, all terms in $V^{\Xi+} - V^q$ are treated equally (i.e., assigned an equal fictitious weight κ_1), so are all terms in $V - V^{\Xi+} - V^q$ (i.e., assigned an equal fictitious weight κ_2). In practice, we may be interested only in those terms which belong to domain $t \in V^{\Xi+}$, in this case, the fictitious weight κ_2 is set to zero (or simply ignored).

Consequently, with Definition 3.5.1, the association score function may be defined by

$$\begin{aligned} score_I(t) = atq_I(t,q) &= \mathcal{Q}(t) \cdot \mathbf{ifd}_I(t) \\ &= \begin{cases} w_q(t) \cdot \mathbf{ifd}_I(t) & \text{when } t \in V \cap V^q \\ \kappa_1 \cdot \mathbf{ifd}_I(t) & \text{when } t \in V^{\Xi+} - V^q \\ \kappa_2 \cdot \mathbf{ifd}_I(t) & \text{when } t \in V - V^{\Xi+} - V^q. \end{cases} \end{aligned}$$

If nothing special is known about the score function under consideration, we might conjecture that this score function should be one that can put good terms, that distinguish the relevant documents from non-relevant ones, near to the top of the sorting list.

It should be noticed that, in a probabilistic IR environment, two terms are considered as being 'semantically related' if they refer to similar documents which may be regarded as relevant to the same query. Thus, 'semantic relation', like the association score, is a statistical property of a term that may change from query to query. We point out that the association of a term with the query is not a property intrinsic to the term. Rather it is a term property with respect to the query. A given term may have as many association scores as the number of queries in which it appears, and its score might be high with respect to some queries and low with respect to some others.

3.6.3 Reduction of Domain

It is very important to understand that that $\mathbf{ifd}_I(t) = 0$ when $t \in V - V^{\Xi+}$ since $P_{\Xi+}(t) = 0$. That is, all contributions made by terms to summation $I(P_{\Xi+} : P_D)$ come purely from candidate terms $t \in V^{\Xi+}$, rather than from terms $t \in V - V^{\Xi+}$. Therefore, a completely equivalent score function can be written as

$$score_I(t) = \mathcal{Q}(t)P_{\Xi+}(t)\log\frac{P_{\Xi+}(t)}{P_D(t)}$$
$$= \begin{cases} w_q(t)\cdot P_{\Xi+}(t)\log\frac{P_{\Xi+}(t)}{P_D(t)} & \text{when } t \in V^{\Xi+} \cap V^q \\ \kappa_1 \quad \cdot P_{\Xi+}(t)\log\frac{P_{\Xi+}(t)}{P_D(t)} & \text{when } t \in V^{\Xi+} - V^q, \end{cases}$$

which is called the association *score* of term t with query q.

If the reader traces through all the discussions given in this chapter, it should become clear that the mathematical definition of the association score embodies the intuitive meaning of the definition, which is that the score involves the product of three essential factors: significance $\mathcal{Q}(t)$ of term t concerning query q, importance $P_{\Xi+}(t)$ of term t concerning the relevant sample set Ξ^+, and the discrimination information $i(H_1 : H_2|t) = \log\left(P_{\Xi+}(t)/P_D(t)\right)$ of term t concerning two opposite relevance hypotheses H_1 and H_2.

Notice that function $score_I(t)$ also assigns scores to all query terms $t \in V^{\Xi+} \cap V^q$. In practice, as documents become longer, almost all the query terms can appear in the relevant sample documents (if they index documents in D). The function judges good terms among all candidate terms $t \in V^{\Xi+}$, even though some of them are query terms that are considered as good ones with respect to the query itself under Assumption 2. In practice, it might happen that the expansion terms $t \in E^q = S^q - V^q \subset V^{\Xi+} - V^q$ obtain higher scores than query terms $t \in V^{\Xi+} \cap V^q$. This implies that the expansion terms might be more strongly associated with the query than query terms themselves. This may happen if the original query is not good enough, and relevant sample documents provided by the user contain completely relevant information concerning his information needs.

Notice also that, when some query term, say t_j, does not appear in any relevant sample documents, i.e., $t_j \in V - V^{\Xi+}$, using function $score_I(t)$ may cause the loss of the consideration of term t_j, that is, $score_I(t_j)$ would not exist (or precisely, $score_I(t_j) = \kappa_2 \cdot 0\log\frac{0}{P_D(t_j)} = 0$) and term t_j would be ignored, but such a possibility would be generally very rare in a practical query expansion process.

3.6.4 About Positive Scores

For some candidate term $t \in V^{\Xi+}$, if $score_I(t) = \mathcal{Q}(t)P_{\Xi+}(t)i(H_1 : H_2|t) < 0$ then it must have $i(H_1 : H_2|t) < 0$ since $\mathcal{Q}(t) \geq 0$ and $P_{\Xi+}(t) > 0$, namely, term t contributes quantity $-|\mathbf{ifd}_I(t)|$ for supporting the relevant hypothesis H_1. If $score_I(t) = 0$ then it must have $i(H_1 : H_2|t) = 0$ and/or $\mathcal{Q}(t) = 0$, namely, term t is either unrelated to the relevance classification or insignificant concerning query q. Consequently, we limit ourselves to consider only those terms which obtain *positive* scores as probably selected terms. This implies that, in effect, the highest priority of the judgement of terms is given to the discrimination factor $i(H_1 : H_2|t)$ in the score function: if the candidate term t is asserted in favour of H_1 negatively, it is immediately discarded even though it might be 'significant' (i.e., it has a greater value $\mathcal{Q}(t)$) and/or 'importance' (i.e., it has a greater value $P_{\Xi+}(t)$), even if it is a query term.

Fortunately, in the practical context of IR, we always have $P_{\Xi^+}(t) > P_D(t)$ for $t \in V^{-\Xi^+}$. This is because the size of set Ξ^+ (set to $|\Xi^+| \leq |\Xi| = 10$ in our experimental designs, for instance) is negligible compared with the size of the extremely large collection D. Therefore, densities $P_{\Xi^+}(t)$ are relatively much greater than densities $P_D(t)$ for all terms $t \in V^{-\Xi^+}$: $\frac{P_{\Xi^+}(t)}{P_D(t)} >> 1$ (here '$x >> y$' denotes that number x is much greater than number y). $P_{\Xi^+}(t) > P_D(t)$ ensures that the discrimination factor $i(H_1 : H_2|t) > 0$, and that term t conveys the positive discrimination information in favour of H_1 against H_2. Thus, when $t \in V^{\Xi^+}$, we always have $score_I(t) > 0$ which makes our experimental consideration become very simple.

To sum up, let us say that we consider a term t whenever we find that it belongs to V^{Ξ^+}. The problem can then be stated as that of computing term scores with function $score_I(t)$ over set V^{Ξ^+}. The terms can be sorted in decreasing order of their scores, which can be regarded as the extent of the association of terms with the context of the query. The terms with the highest positive scores should be given a high priority as selected terms $t \in S^q$ because these terms make the greatest contribution to summation $I(P_{\Xi^+} : P_D)$ among terms $t \in V$. These selected terms should be regarded as strongly associated with query q, and as good discriminators to distinguish relevant documents from many non-relevant ones.

3.6.5 Pseudo-Relevance Feedback Process

In an operational situation where no relevance information is available in advance, we would proceed as follows. Let the sample set Ξ be the top-ranked documents obtained from an initial (a previous) retrieval iteration. All documents $d \in \Xi$ are treated as relevant, and V^{Ξ} constitutes a source of candidate terms. We can also invoke the method proposed in this section to construct the score functions, with Ξ instead of Ξ^+, as discussed in the case of the relevance feedback process.

Since terms in the top-ranked documents are more likely to be relevant to the context of the query than many other documents, it may be reasonable for us to consider the judgement of good terms from the top-ranked documents.

However, in a pseudo-relevance feedback procedure, if the initial retrieval returns a low precision, the estimates of the discrimination measure or, the estimate of the probability distributions (see Section 3.7), may be poor due to limited and noisy training samples providing insufficient and unreliable relevance information. In this case, the expanded query cannot be expected to produce any further improvement in retrieval performance, and may even hurt the original query.

Insofar as query expansion is the technique of enhancement of queries, it cannot give rules for construction of the sample set. The consideration by which we choose an effective sample set is a part of the art of designing a feasible and effective query expansion method.

3.6.6 Examples for Estimating $\mathcal{Q}(t)$

In order to actually consider a more practical form of function $score_I(t)$, we here provide two examples of estimating $\mathcal{Q}(t)$ based on the different considerations on the importance of query terms.

Example 3.6.1 As we know, determining the importance of each query term is generally rather difficult, and not all query terms have the same discrimination capability. For instance, it is generally assumed that query terms that are assigned to many documents are not very

useful in distinguishing the relevant documents from the non-relevant ones; whereas terms that occur in a few documents have a good chance of occurring in the relevant documents. Such an assumption, while not necessarily true for every term in every query, may be expected to hold for the vast majority of query terms.

Basically, the factors for measuring the importance of terms are the occurrence frequencies of terms in the query, and probably the document frequencies of terms. Thus, assume that query q is initially represented by $M_q = \left[w_q(t)\right]_{1\times n}$, we can give the following initial weights for query terms:

$$w_q(t) = p_q(t) \times idf_D(t),$$

where

$$p_q(t) = \frac{f_q(t)}{\sum_{t'\in V^q} f_q(t')} = \begin{cases} \frac{f_q(t)}{||q||} & \text{when } t \in V^q \\ 0 & \text{when } t \in V - V^q \end{cases}$$

is the *a priori* probability of term t being true in query q.

Suppose that $F_D(t) \leq 0.1|D|$ for all terms $t \in V$ holds[3]. Thus, we can set $\kappa_1 = \frac{1}{||q||} \times \log 10$. It has $\kappa_1 \leq min_{t\in V^q}\{w_q(t)\}$ since $p_q(t) = \frac{f_q(t)}{||q||} \geq \frac{1}{||q||}$ and $idf_D(t) = \log \frac{|D|}{F_D(t)} \geq \log \frac{|D|}{0.1|D|} = \log 10$. Then, for each candidate term $t \in V^{\Xi+}$, we have

$$\mathcal{Q}(t) = \max\{w_q(t), \kappa_1\} = \max\{\frac{f_q(t)idf_D(t)}{||q||}, \frac{\log 10}{||q||}\}.$$

That is, if term $t \in V^{\Xi+} \cap V^q$ then it is assigned a true weight $\frac{f_q(t)idf_D(t)}{||q||}$; if term $t \in V^{\Xi+} - V^q$ then it is assigned a fictitious weight $\frac{\log 10}{||q||}$. Notice that $||q||$ is basically just a scale factor normalizing query q, and is independent of all terms $t \in V^{\Xi+}$. By eliminating the scale factor, we obtain the following (equivalent) score function:

$$score_I(t) = \max\{f_q(t)idf_D(t),\ \log 10\}P_{\Xi+}(t)\log\frac{P_{\Xi+}(t)}{P_D(t)} \qquad (t \in V^{\Xi+}).$$

From this, we can easily see that, normally, the length of the query need not be taken into account in the $\mathcal{I}f\mathcal{D}$ system since we consider only one query at a time. ♠

Example 3.6.2 Consider a special case where all query terms are regarded as being equally important for the query. In this case, we can easily take $w_q(t) = 1$ (when $t \in V^q$) and set $\kappa_1 = 1$ (when $t \in V^{\Xi+} - V^q$). Clearly, it has $\kappa_1 \leq min_{t\in V^q}\{w_q(t)\}$. Then, for each candidate term $t \in V^{\Xi+}$, we have $\mathcal{Q}(t) = 1$ and

$$score_I^*(t) = P_{\Xi+}(t)\log\frac{P_{\Xi+}(t)}{P_D(t)} \qquad (t \in V^{\Xi+}).$$

Thus, all candidate terms are scored by considering only their discrimination information. ♠

[3]In the $\mathcal{I}f\mathcal{D}$ model, terms with the document frequencies $F_D(t) > 0.1|D|$ are removed (see Subsection 8.2.2). Here number $0.1|D|$ is immaterial, and depends on a specific model itself.

3.7 Estimation of Term Distributions

The discrimination measure is the main component of the score function. Thus, the estimation of the discrimination measure is crucial for effectively identifying the potentially good terms from many others. Therefore, before finishing this chapter, we will discuss this important estimation issue.

It is clear that the discrimination measure $\mathbf{ifd}_I(t)$ in Definition 3.4.1 is uniquely determined by its two arguments $P_R(t)$ and $P_{\bar{R}}(t)$. Thus, the issue of the estimation is centred round the estimation of its arguments. Based on Shannon's basic ideas that the probability distributions should be established before the discrimination information of terms can be considered. Thus, we now concentrate on mathematical discussions of these distributions. A general framework for the estimation is established. Some estimation schemes are, as examples, elaborated to embody the arguments in the discrimination measure.

The term probability distribution concerning a certain document set, D_k, can generally be estimated based on the representation, M_{D_k}, of the set, which is in turn estimated based on the representation, M_d, of document d in the set. Set D_k may consist of just a few sample documents, or of all of documents in the collection; whatever the set may be, depends on a specific application. The representation of the document set plays an essential role in determining retrieval effectiveness. As we know, the accuracy and validity of effective representations for individual documents and document sets has long been a crucial and open problem. This is because, as mentioned before, it is very difficult to obtain sufficient statistics for the estimation of the amount of information contained in terms, and for the indication of the semantic relations between terms. Almost all existing probabilistic methods suffer from the same problem. This thesis does not give rules for the representations, which will be regarded as a significant subject for further study. Instead, we only show some simple estimation methods, and the corresponding experimental investigations on the effectiveness of the methods will be discussed in Chapter 8.

3.7.1 Estimation of M_{D_k}

In this thesis, the general form of the representation, $M_{D_k} = [w_{D_k}(t)]_{1\times n}$, of document set D_k is defined as

$$w_{D_k}(t) = \sum_{d\in D_k} \chi_{D_k}(d) w_d(t),$$

where $\chi_{D_k}(d)$ is a function used to reflect the importance of document d concerning set D_k. For instance, for each $d \in D_k$, we may set

$$\chi_{D_k}(d) = 1 \quad \text{and} \quad \chi_{D_k}(d) = sim(d, q).$$

Function $\chi_{D_k}(d) = 1$ is the most common one, which indicates that all documents in D_k is treated as equally important concerning D_k. This function is needed when one has no particular reason to emphasize any one of documents in D_k. A typical example for using this function is the situation where we consider documents in the whole collection D.

In order to design an effective discriminant measure through the relevance feedback, the relevance information obtained from the initial retrieval iteration may be taken into account

by incorporating the information into representation M_{D_k} to estimate distribution $P_{D_k}(t)$. Thus, function $\chi_{D_k}(d) = sim(d,q)$ is introduced to depict the importance of documents by their similarity with query q (the choice of the similarity measure $sim(d,q)$ depends on a specific model itself). An example for this is the case where the relevant sample set Ξ^+ is considered, and we say that one document $d \in \Xi^+$ is more important if it obtains higher similarity than others.

Thus, we can see that the component, $w_{D_k}(t)$, of representation M_{D_k} is the summation of weights $w_d(t)$, multiplied by the importance $\chi_{D_k}(d)$ of the corresponding document d, of term $t \in V^d$ over $d \in D_k$. It is clear that term weight $w_d(t)$ concerning a given document d is essential in component $w_{D_k}(t)$, which we discuss below.

3.7.2 Estimation of M_d

The different *recall*[4] and *precision*[5] requirements may favour the different combination of factors, that contain recall- and/or precision-enhancing components, for weighting terms of documents. Some such factors and their combinations have been studied experimentally for representing the statistical importance of a term concerning the individual documents, [70, 164] for instance. We here discuss some common important factors.

Frequencies of Terms

In practice, what we have are only observations, i.e., the statistical frequency of terms within documents. Thus, the factor affecting the importance of a term t concerning a given document d is, first and foremost, its frequency of occurrence within the document:

$$f_d(t), \qquad \log f_d(t), \qquad p_d(t) = \frac{f_d(t)}{||d||}.$$

It is clear that terms frequently appearing in individual documents are useful as a recall-enhancing device. The term frequency $f_d(t)$ considered as a component of weighting functions of terms has been used for many years in automatic indexing environments [119, 161, 167, 207].

Function $\log f_d(t)$ is a variety of frequency $f_d(t)$, and might be necessary to accord with other functions in scale. In particular, $\log f_d(t)$ is needed when it is incorporated into other functions with log.

Function $p_d(t)$ is the normalization of frequency $f_d(t)$, i.e., it considers the length, $||d|| = \sum_{t \in V^d} f_d(t)$, of document d. Generally, the importance of a term has two aspects:

- the importance in representing a specific document.

  For instance, $d_1 = \{t_1, t_2, t_2, t_3\}$,

  thus we may say that term t_2 is more important than others for document d_1.

- the importance in representing the different documents.

  For instance, $d_1 = \{t_1, t_2, t_2, t_3\}$, and $p_{d_1}(t_2) = \frac{f_{d_1}(t_2)}{||d_1||} = \frac{2}{4}$,

  $d_2 = \{t_1, t_2, t_2, t_2, t_3, t_4, t_5, t_6, t_7, t_8, t_9, t_{10}\}$, and $p_{d_2}(t_2) = \frac{f_{d_2}(t_2)}{||d_2||} = \frac{3}{12}$.

[4] *Recall*—the proportion of relevant documents actually retrieved in answer to a query.

[5] *Precision*—the proportion of retrieved documents actually relevant to the query.

thus we may say that term t_2 is more important for document d_1 than document d_2.

Normalization of the term frequency in some sense moderates the effect of high-frequency terms. This point is rather important because a term t_2, for instance, with $f_{d_2}(t_2) = 3 > 2 = f_{d_1}(t_2)$ should not be viewed as more important for document d_2 than for document d_1. In other words, the importance of a term in a specific document would depend highly on the 'percentage' that this term possesses over the total sum of frequency of terms in the document (i.e., $||d||$). The higher the percentage, the more important the term would be for the document.

On the other hand, when some longer document involves a large number of terms, the chance of term matches between the document and the query would be rather high, and hence the document has a better chance of being retrieved than other short ones. Thus, the use of the normalization factor can make all documents be treated equally for retrieval purposes.

Inverse Document Frequency of Term $idf_D(t)$

Besides the functions of term frequency given above, another factor that can also affect the importance of terms concerning a given document, and thus should be incorporated into the term weights, is the information of the specificity of a term concerning the whole collection. A well-known one is the *inverse document frequency*:

$$idf_D(t) = -\log \frac{F_D(t)}{|D|},$$

which states that the specificity of term t is inversely proportional to the document frequency, $F_D(t)$, of the term.

Thus, the more documents in which the term occurs, the less specificity the term has. This formula was implemented in some experiments [187] as

$$idf^*_D(t) = -\log \frac{F_D(t)}{max_{F_D}},$$

where $max_{F_D} = \max\{F_D(t); t \in V\}$ was the maximum document frequency among all terms.

We can hence see that the inverse document frequency provides a measure of the specificity of a term. The measure assigns higher values to more specific terms that tend to be capable of isolating the few relevant documents from the many non-relevant ones. Thus, the inverse document frequency can be viewed as a precision-enhancing device.

The inverse document frequency weighting was initially proposed by Sparck Jones [185]. She justified this weighting scheme based on the observation that term distribution had the similar Zipf shape [240]. Robertson [148] pointed out however that Zipf's argument is not intended as a theoretical justification for the weighting function, and the only justification suggested is retrieval performance.

The use of document frequency for representing term importance is a simplifying device which is generally, but not absolutely, valid. A more accurate indication of term importance may be obtained by using the distribution of the document frequency of term across all documents of the collection, which is discussed below.

Inverse Noise of Term $int_D(t)$

An alternative function, *noise* of term [167, 171], can be used to capture the unspecificity of a term concerning collection D:

$$noise_D(t) = H\big(P_D(d|t)\big) = -\sum_{d\in D} P_D(d|t)\log P_D(d|t),$$

where

$$P_D(d|t) = \frac{f_d(t)}{\sum_{d\in D} f_d(t)} = \frac{f_d(t)}{f_D(t)}.$$

That is, $noise_D(t)$ is the entropy of conditional probability distribution $P_D(d|t)$. From the properties of the entropy function, we have the following conclusions.

- If term t occurs in only one document, that is, for some fixed k $(1 \le k \le N)$, it has

  $$P_D(d_k|t) = 1, \quad P_D(d_1|t) = ... = P_D(d_{k-1}|t) = P_D(d_{k+1}|t) = ... = P_D(d_N|t) = 0,$$

  then the noise of term t will receive zero.

- If term t is uniformly distributed over a certain document set $D_t \subseteq D$ (D_t is the set of documents in which term t appears, obviously, $F_D(t) = |D_t|$), that is,

  $$P_D(d|t) = \begin{cases} \frac{f_d(t)}{\sum_{d\in D_t} f_d(t)} = \frac{f_d(t)}{|D_t| f_d(t)} = \frac{1}{F_D(t)} & \text{when } d \in D_t \\ 0 & \text{when } d \in D - D_t, \end{cases}$$

  then the noise of term t will be:

  $$\begin{aligned} H\big(P_D(d|t)\big) &= -\sum_{d\in D} P_D(d|t)\log P_D(d|t) = -\sum_{d\in D_t} \frac{1}{F_D(t)}\log\frac{1}{F_D(t)} \\ &= -|D_t|\frac{1}{F_D(t)}\log\frac{1}{F_D(t)} = \log F_D(t). \end{aligned}$$

- Particularly, if term t is uniformly distributed over the whole collection (i.e., $D_t = D$), we have,

  $$P_D(d_1|t) = P_D(d_2|t) = ... = P_D(d_N|t) = \frac{1}{|D|}$$

  and the noise of term t will arrive at the maximum $\log|D|$.

It is very clear that entropy $H\big(P_D(d|t)\big)$ gives the degree of uncertainty of term t when it is used to index collection D. Namely, measure $noise_D(t)$ offers the extent of the lack of concentration of the occurrence of term t, and thus it emphasizes the uselessness of those terms that are in agreement with probability $P_D(d|t)$ for individual documents in D.

It is worth noticing that the specificity of term t is in inverse relation to its noise. Thus, the specificity of term t may be computed by [70]:

$$int_D(t) = noise_{max} - noise_D(t),$$

where $noise_{max} = \max\{noise_D(t)|t \in V\}$, i.e., the maximum noise among all terms. The measure of $int_D(t)$ is called *inverse noise* of term t. Because measure $int_D(t)$ assigns low values to those terms that are not concentrated in a few particular documents, but instead are prevalent in the whole collection, it should be an appropriate measure of term specificity, and hence, as measure $idf_D(t)$, be thought of as a precision-enhancing device.

Example 3.7.1 Let $D = \{d_1, d_2, d_3, d_4\}$, suppose $t_1 = computer$, $t_2 = information$, $t_3 = divergence$, and so on.

D	$f(t_1)$	$f(t_2)$	$f(t_3)$	$f(t_4)$	...	$f(t_n)$
d_1	1	2	0	0	...	0
d_2	3	0	0	0	...	0
d_3	1	1	1	0	...	0
d_4	2	1	0	0	...	0

$$
\begin{aligned}
noise_D(t_1) &= -\frac{1}{7}\log\frac{1}{7} - \frac{3}{7}\log\frac{3}{7}\frac{1}{7}\log\frac{1}{7} - \frac{2}{7}\log\frac{2}{7} \\
&= -0.1429\log 0.1429 - 0.4286\log 0.4286 - 0.1429\log 0.1429 - 0.2857\log 0.2857 \\
&= 0.1429 \times 1.9456 + 0.4286 \times 0.8472 + 0.1429 \times 1.9456 + 0.2857 \times 1.2528 \\
&= 0.2780 + 0.3631 + 0.2780 + 0.3579 = 1.2770, \\
noise_D(t_2) &= -\frac{2}{4}\log\frac{2}{4} - \frac{0}{4}\log\frac{0}{4} - \frac{1}{4}\log\frac{1}{4} - \frac{1}{4}\log\frac{1}{4} \\
&= -0.5\log 0.5 - 0.5\log 0.25 = 0.5 \times 0.6931 + 0.5 \times 1.3863 \\
&= 0.3466 + 0.6932 = 1.0398, \\
noise_D(t_3) &= -\frac{0}{1}\log\frac{0}{1} - \frac{0}{1}\log\frac{0}{1} - \frac{1}{1}\log\frac{1}{1} - \frac{0}{1}\log\frac{0}{1} = 0.0000.
\end{aligned}
$$

Thus, if we take $noise_{max} = 1.2770$, then the specificity of terms might be computed by

$$
\begin{aligned}
int_D(t_1) &= noise_{max} - noise_D(t_1) = 1.2770 - 1.2770 = 0.0000, \\
int_D(t_2) &= noise_{max} - noise_D(t_2) = 1.2770 - 1.0398 = 0.2379, \\
int_D(t_3) &= noise_{max} - noise_D(t_3) = 1.2770 - 0.0000 = 1.2770.
\end{aligned}
$$

Obviously, the occurrence of term *divergence* is more important for the subject of the document than the occurrence of term *computer*. ♠

The Relationship Between $idf_D(t)$ and $int_D(t)$

A common basic idea used in both measures $idf_D(t)$ and $int_D(t)$ is that if term t has a skewed document frequency distribution over D, then term t can be expected to be a good discriminator for distinguishing one document from many others [167]. However, the point of these two measures is different: measure $idf_D(t)$ ignores the consideration of term frequencies within documents, and thus terms with the same document frequency will be treated equally by assigning the same weights. In contrast, measure $int_D(t)$ takes into account both term and document frequencies, and hence it is very likely that terms with the same document frequency are given different weights.

In order to further investigate the relationship between measures $idf_D(t)$ and $int_D(t)$, let us consider an alternative measure, very similar to $int_D(t)$, suggested by Wong & Yao [224]:

$$int^{\star}_D(t) = 1 - \frac{noise_D(t)}{H_{max}},$$

where H_{max} is denoted as the maximum entropy. According to the properties of the entropy function, the value of H_{max} for entropy function $H\big(P_D(d|t)\big)$ should be $\log|D|$ for a collection of $|D|$ documents. Thus, we can write down

$$int^{\star}_D(t) = 1 - \frac{noise_D(t)}{\log|D|} = \frac{1}{\log|D|}\big(\log|D| - noise_D(t)\big),$$

which is completely equivalent to

$$int^{*}_D(t) = \log|D| - noise_D(t)$$

since factor $\frac{1}{\log|D|}$ is a positive constant independent of any particular term. We can state that both measures $idf_D(t)$ and $int_D(t)$ are special cases of measure $int^{*}_D(t)$.

On one hand, in practice, maximum $H_{max} = \log|D|$ would never be attained simply because, in practice, we usually remove all stop-words (even some general terms with very high document frequencies, e.g., $F_D(t) > 0.1|D|$ in our experimental setting), which implies that there is no term that would appear in all documents. It is clear that generally $H_{max} > noise_{max}$ for every term $t \in V$. When the maximum entropy H_{max} is reasonably substituted by the maximum noise $noise_{max}$, we obtain $int^{*}_D(t) = int_D(t)$.

On the other hand, assume now that documents are represented by the binary weights of terms. In this case, term frequency information would be ignored, and each term $t \in V$ will correspond to a distribution,

$$P_D(d|t) = \begin{cases} \frac{1}{F_D(t)} & \text{when } d \in D_t \\ 0 & \text{when } d \in D - D_t, \end{cases}$$

which is the same as the distribution derived from the case that term t is uniformly distributed over D_t. From the above discussion on the properties of the entropy function, it can readily be seen that $int^{*}_D(t) = \log|D| - \log F_D(t) = idf_D(t)$. More generally, as pointed out by Wong & Yao [224], it is explicitly shown that $idf_D(t)$ can be derived from $int^{*}_D(t)$ by assuming that the document frequency of a term is uniform within the corresponding set D_t.

We can now interpret measure $idf_D(t)$ in the information theoretic sense. Notice that, with the binary representation, it has $\log F_D(t) = H\big(P_D(d|t)\big)$. Thus, as mentioned above in the context of the entropy function, $\log F_D(t)$ measures the degree of uncertainty of term t when it is used to index collection D. Clearly, the larger the number of documents in which the term t appears, the larger the uncertainty that term t causes. Further, $\log \frac{F_D(t)}{|D|}$ can be thought of as relative uncertainty, and $idf_D(t) = \log \frac{|D|}{F_D(t)}$ can be regarded as a measure of certainty.

In addition, it may be worth mentioning the following formula proposed by Salton & McGill [167]:

$$int^{**}_D(t) = \log\Big(\sum_{d \in D} f_d(t)\Big) - noise_D(t) = \log f_D(t) - noise_D(t).$$

It was shown that this measure produces inferior performance compared with the $idf_D(t)$ measure [167]. We point out that this measure may have a problem for the application in a practical IR context, and attempt to explain the reason why an inferior performance was caused by this measure by an example.

Consider Example 3.7.1, we can easily compute

$$int^{**}_D(t_1) = f_D(t_1) - noise_D(t_1) = \log 7 - 1.2770 = 1.9460 - 1.2770 = 0.6690,$$
$$int^{**}_D(t_2) = f_D(t_2) - noise_D(t_2) = \log 4 - 1.0398 = 1.3863 - 1.0398 = 0.3465,$$
$$int^{**}_D(t_3) = f_D(t_3) - noise_D(t_3) = \log 1 - 0.0000 = 0 - 0.0000 = 0.$$

These results are very different from the corresponding results obtained in Example 3.7.1 in which measure $int_D(t)$ was applied. For this measure, a term with a larger total frequency $f_D(t)$ will obtain a higher term weight. For instance, term t_1 appearing in all documents receives the greatest weight, whereas term t_3 occurring in only one document (so it should be the most specific) gains least weight zero. Such results are not acceptable for an effective term weighting scheme. This problem is absent in the more general measure $int^*_D(t)$. Thus, this may explain the reason why the $idf_D(t)$ weighting method, as a special case of $int^*_D(t)$, may produce a superior performance than that achieved by the $int^{**}_D(t)$ weighting method.

3.7.3 Combination Schemes

All functions discussed above can be combinatorially considered to form term weights for representing documents and the document sets considered. Generally, the term weights should be designed to be able to highlight the natures of the sets themselves.

Once term weights $w_d(t)$ for the individual documents in corresponding sets Ξ^+ and D, and functions $\chi_{\Xi+}(d)$ and $\chi_D(d)$, are given, components $w_{\Xi+}(t)$ and $w_D(t)$ of representations $M_{\Xi+}$ and M_D can be easily obtained, respectively. The different combinations will produce the different estimations of $P_{\Xi+}(t)$ and $P_D(t)$, and then generate the various score functions.

For instance, consider set Ξ^+, if we take $\chi_{\Xi+}(d) = sim(d,q)$ and $w_d(t) = \log f_d(t)$, then the component of representation $M_{\Xi+}$ can be written down

$$w_{\Xi+}(t) = \sum_{d\in\Xi+} \chi_{\Xi+}(d) w_d(t) = \sum_{d\in\Xi+} sim(d,q)\big(\log f_d(t)\big),$$

which indicates the importance of term $t \in V$ concerning set Ξ^+.

Also, consider collection D, if we take $\chi_D(d) = 1$ and $w_d(t) = \big(\log f_d(t)\big) idf_D(t)$, then we have component of representation M_D as

$$w_D(t) = \sum_{d\in D} \chi_D(d) w_d(t) = \sum_{d\in D} \big(\log f_d(t)\big) \log \frac{|D|}{F_D(t)},$$

which indicates the importance of term $t \in V$ concerning collection D.

Salton & Buckley [164] in their studies pointed out that the 'best' terms for document representation are those which can distinguish certain individual documents from the remainder of the collection, and in this case, should have high term frequencies but low inverse document frequencies. Thus, a reasonable measure of term importance may be obtained by product $f_d(t) idf_D(t)$ [171, 172]. Notice that, for $d \in D$, we may adopt $w_d(t) = \big(\log f_d(t)\big) idf_D(t)$ instead of $w_d(t) = f_d(t) idf_D(t)$ since it may be necessary for $f_d(t)$ to accord with $\log \frac{|D|}{F_D(t)}$ in scale.

Schemes	$\chi_{\Xi^+}(d) = sim(d,q)$	$\chi_D(d) = 1$
scheme-a	$w_d(t) = \big(\log f_d(t)\big) int_D(t)$	$w_d(t) = \big(\log f_d(t)\big) int_D(t)$
scheme-b	$w_d(t) = \log f_d(t)$	$w_d(t) = \big(\log f_d(t)\big) int_D(t)$
scheme-c	$w_d(t) = f_d(t) int_D(t)$	$w_d(t) = f_d(t) int_D(t)$
scheme-d	$w_d(t) = f_d(t)$	$w_d(t) = f_d(t) int_D(t)$
scheme-e	$w_d(t) = \big(\log f_d(t)\big) idf_D(t)$	$w_d(t) = \big(\log f_d(t)\big) idf_D(t)$
scheme-f	$w_d(t) = \log f_d(t)$	$w_d(t) = \big(\log f_d(t)\big) idf_D(t)$
scheme-g	$w_d(t) = f_d(t) idf_D(t)$	$w_d(t) = f_d(t) idf_D(t)$
scheme-h	$w_d(t) = f_d(t)$	$w_d(t) = f_d(t) idf_D(t)$

Schemes	$\chi_{\Xi^+}(d) = 1$	$\chi_D(d) = 1$
scheme-i	$w_d(t) = p_d(t)$	$w_d(t) = f_d(t) int_D(t)$
scheme-j	$w_d(t) = p_d(t)$	$w_d(t) = f_d(t) idf_D(t)$
scheme-k	$w_d(t) = f_d(t)$	$w_d(t) = f_d(t) int_D(t)$
scheme-l	$w_d(t) = f_d(t)$	$w_d(t) = f_d(t) idf_D(t)$
scheme-m	$w_d(t) = f_d(t)$	$w_d(t) = f_d(t)$

Some simple examples of combination schemes of the representations $M_{\Xi^+} = \big[w_{\Xi^+}(t)\big]_{1\times n}$ and $M_D = \big[w_D(t)\big]_{1\times n}$ are listed in the tables above.

3.7.4 Estimation of $P_{\Xi^+}(t)$ and $P_D(t)$

Assume that, for a given document set D_k, representation $M_{D_k} = \big[w_{D_k}(t)\big]_{1\times n}$ has been obtained. Once we have quantities $w_{D_k}(t)$ for all terms $t \in V$, it will be simple to express for distribution $P_{D_k}(t)$:

$$P_{D_k}(t) = \begin{cases} \frac{w_{D_k}(t)}{\sum_{t\in V^{D_k}} w_{D_k}(t)} & \text{when } t \in V^{D_k} \\ 0 & \text{when } t \in V - V^{D_k}, \end{cases} \tag{3.2}$$

which is clearly a probability distribution over V. It is easily seen that $P_{D_k}(t) > 0$ for every $t \in V^{D_k}$. Obviously, $V^{D_1} \subseteq V^{D_2}$ if $D_1 \subseteq D_2$, and $P_{D_1}(t) \ll P_{D_2}(t)$ when $t \in V$ (since $P_{D_1}(t) = 0$ whenever $P_{D_2}(t) = 0$). This ensures that summation $I\big(P_{D_1} : P_{D_2}\big)$ always exists.

Particularly, for $D_1 = \Xi^+$ and $D_2 = D$, we can write representations $M_{\Xi^+} = \big[w_{\Xi^+}(t)\big]_{1\times n}$ and $M_D = \big[w_D(t)\big]_{1\times n}$ according to the foregoing discussions, and estimate distributions $P_{\Xi^+}(t)$ and $P_D(t)$ in terms of expression (3.2). That is,

$$P_{\Xi^+}(t) = \begin{cases} \frac{w_{\Xi^+}(t)}{\sum_{t\in V^{\Xi^+}} w_{\Xi^+}(t)} & t \in V^{\Xi^+} \\ 0 & t \in V - V^{\Xi^+}, \end{cases} \qquad P_D(t) = \frac{w_D(t)}{\sum_{t\in V} w_D(t)} \quad t \in V. \tag{3.3}$$

Clearly, it has $\Xi^+ \subseteq D$, and so summation $I(P_{\Xi^+} : P_D)$ exists.

Therefore, we can estimate $P_{\Xi^+}(t)$, for instance, for scheme-f, by

$$P_{\Xi^+}(t) = \frac{w_{\Xi^+}(t)}{\sum_{t\in V^{\Xi^+}} w_{\Xi^+}(t)} = \frac{\sum_{d\in\Xi^+} sim(d,q)\big(\log f_d(t)\big)}{\sum_{t\in V^{\Xi^+}} \big[\sum_{d\in\Xi^+} sim(d,q)\big(\log f_d(t)\big)\big]},$$

$$P_D(t) = \frac{w_D(t)}{\sum_{t\in V} w_D(t)} = \frac{\sum_{d\in D} \big(\log f_d(t)\big) idf_D(t)}{\sum_{t\in V} \big[\sum_{d\in D} \big(\log f_d(t)\big) idf_D(t)\big]},$$

which satisfies that $P_{\Xi+}(t) > 0$ for every $t \in V^{\Xi+}$ and $P_D(t) > 0$ for every $t \in V$.

Notice that in the $\mathcal{I}f\mathcal{D}$ model the *postulates* of $P_{\Xi+}(t) > 0$ and $P_D(t) > 0$ for every $t \in V^{\Xi+} \subseteq V$ are not excessive. They are necessary and the least of conditions for applying the directed divergence $I(P_{\Xi+} : P_D)$ to construct the score function $score_I(t)$. Because vocabulary V is a finite tuple, these postulates are not infeasible and are practical in a realistic IR context.

Finally, we would like to point out that the discrimination information gained from the probability distributions estimated from the (relevant) sample set may not be correctly measured for each term. This fact should not be taken as a major criticism of any formal model because almost all feedback methods would suffer from the same problem of sampling [60, 226].

3.8 Summary

This chapter describes the application of the basic concept of directed divergence to the technique of automatic query expansion. The rationale of applying logarithmic measure of information to measuring the discrimination information contained in terms is interpreted. Some important points of this chapter are now summarized as follows.

¶ It is essential and important for any divergence measure to satisfy two criterion which were given in Section 3.3. Under these two criteria, the extent to which terms contribute to the expected divergence can be measured, and the divergence measure can be independent of the addition or removal of terms which are unrelated to the relevance classification.

¶ Generally, it is accepted that terms with higher power of discrimination should be considered as more important. Statistically, terms which are thought of as having higher power of discrimination tend to contribute more to the expected divergence than others. It appears that the terms with more concentrated distribution in a certain document set, i.e., with greater variant probabilities within the different document sets, would make a greater contribution to the expected divergence and, therefore, should be viewed as statistically containing more discriminant information.

¶ The information of terms for discrimination is a fundamental issue in IR. The discrimination factor $i(H_1 : H_2|t)$ is carefully examined, and is regarded as a measure of the amount of the information contained in term t for discrimination in favour of H_1 against H_2. The discrimination measure $\mathbf{ifd}_I(t)$, a basis for the methods proposed in this thesis, is formally introduced.

¶ The concept of the association of terms with a query plays a central role in the construction of the score functions for query expansion. The association function is formally defined. We pointed out that the Association Hypothesis due to Van Rijsbergen [207] is an important underlying hypothesis theoretically in IR. A more general hypothesis, called the Generalized Association Hypothesis, is introduced based on the Association Hypothesis. The difference between these two hypotheses is discussed.

¶ The construction of a score function is described for judging good terms. A general form of the construction indicates that the mathematical definition of the association score

involves three essential important factors: significance $\mathcal{Q}(t)$ of term t concerning query q, importance $P_{\Xi+}(t)$ of term t concerning relevant sample set, and the discrimination information $i(H_1 : H_2|t)$ of term t concerning two opposite relevance hypotheses. There may be many ways to construct the score function by the different methods to estimate $\mathcal{Q}(t)$, $P_{\Xi+}(t)$ and $P_D(t)$.

¶ The estimation of distributions $P_{\Xi+}(t)$ and $P_D(t)$ is crucial for effectively distinguishing the potentially good terms from many others. Some estimation schemes are elaborated to embody the arguments of the discrimination measure. A preliminary study is given from general to specific. Some factors are combinatorially considered to form term weights for representing the different documents sets, thus producing different estimations of the term probability distributions, and generating the various score functions.

It should be emphasized again that, in order to speak of the discrimination information of terms, we should regard the arguments of the divergence measure, i.e., the probability distributions involved, as defined on the *same* probability space. Thus, we say that both distributions $P_{\Xi+}(t)$ and $P_D(t)$ are over V even though $P_{\Xi+}(t) = 0$ for all terms $t \in V - V^{\Xi+}$. We will see that this point is the major premise of all discussions given in the next chapter.

In addition, $I\big(P_R : P_{\bar{R}}\big)$ is *not symmetric* in arguments $P_R(t)$ and $P_{\bar{R}}(t)$. It may be desirable to have a symmetrical divergence measure which is meaningful in terms of the information gain. Symmetric divergence measures will be discussed in the following chapters.

Chapter 4

AQE Based on Divergence

This chapter is mainly concerned with discussion of a formal method, based on the basic concept of divergence, for automatic query expansion. After discussing the divergence measure concisely in Section 4.1, in Section 4.2, we introduce a relevance discrimination measure based on the concept of divergence, and discuss a severe application problem that arises: the condition of absolute continuity of probability distributions may not be satisfied; this problem must be solved if the divergence measure is to be applied to automatic query expansion. In Section 4.3, a possible way of solving the problem is suggested, and the solution is carefully discussed in general form. Then, a modified discrimination measure is formally defined. In Section 4.4, we give the concept of the association of terms with the context of the query based on the modified discrimination measure. In Section 4.5, we focus on the construction of the score function, and address the issue of the reduction of the domain of the score function. In Section 4.6, we make further mathematical discussions about the existence of the modified discrimination measure by providing concrete forms of modifying probability distributions. Two methods of modification are described.

4.1 Information Gain $J(P_R : P_{\bar{R}})$

Let H_1 and H_2 be two opposite hypotheses that term t is drawn from sets R and $\bar{R}$, respectively. Assume that $P_R(t)$ and $P_{\bar{R}}(t)$ are term probability distributions over $(V, 2^V)$ under the hypotheses. Then, *divergence* between hypotheses H_1 and H_2 due to Kullback & Leibler [107] is defined by

$$\begin{aligned} J(P_R, P_{\bar{R}}) &= I(H_1 : H_2|H_1) - I(H_1 : H_2|H_2) \\ &= I(H_1 : H_2|H_1) + I(H_2 : H_1|H_2) = I(P_R : P_{\bar{R}}) + I(P_{\bar{R}} : P_R). \end{aligned}$$

It may also be interpreted as the expected information for discrimination in favour of H_1 against H_2, given H_1, plus the expected information for discrimination in favour of H_2 against H_1, given H_2.

If $P_R(t)$ and $P_{\bar{R}}(t)$ are absolutely continuous with respect to each other, then $J(P_R, P_{\bar{R}}) < \infty$, and can be expressed as

$$J(P_R, P_{\bar{R}}) = \sum_{t \in V} (P_R(t) - P_{\bar{R}}(t)) \log \frac{P_R(t)}{P_{\bar{R}}(t)},$$

which can be used to measure the expected divergence of distribution $P_{\bar{R}}(t)$ from distribution $P_R(t)$, plus the expected divergence of distribution $P_R(t)$ from distribution $P_{\bar{R}}(t)$. In applications of IR, $J(P_R, P_{\bar{R}})$ can also be interpreted as a measure of the *difference* between the information contained in $P_R(t)$ and that contained in $P_{\bar{R}}(t)$ about $P_R(t)$, and vice versa.

It is shown that $J(P_R, P_{\bar{R}}) \geq 0$ with equality if and only if $P_R(t) = P_{\bar{R}}(t)$ for all $t \in V$. There is no expected discrimination information if the term distributions are identical.

It can be seen that $J(P_R, P_{\bar{R}})$ is symmetric with respect to $P_R(t)$ and $P_{\bar{R}}(t)$. In practical IR, it may sometimes be desirable to be consistent in measuring the difference between two distributions. Measure $J(P_R, P_{\bar{R}})$ is explored so as to produce a symmetric divergence measure when we have no particular reason to emphasize either $P_R(t)$ or $P_{\bar{R}}(t)$. Thus, it may be more natural and reasonable for us to think of the divergence as a '*distance*' measure between distributions.

4.2 Discrimination Measure $\mathbf{ifd}_J(t)$

4.2.1 Definition of Discrimination Measure

In order to measure the extent of the contributions made by individual terms to the divergence, similar to the discussion of the directed divergence, let us write the divergence as the sum of items,

$$J(P_R, P_{\bar{R}}) = \sum_{t \in V} \mathbf{ifd}_J(t),$$

in which, each item,

$$\begin{aligned} \mathbf{ifd}_J(t) &= \big(P_R(t) - P_{\bar{R}}(t)\big) \log \frac{P_R(t)}{P_{\bar{R}}(t)} \\ &= P(t|H_1) i(H_1 : H_2|t) + P(t|H_2) i(H_2 : H_1|t) \\ &= \mathbf{ifd}_{I_{12}}(t) + \mathbf{ifd}_{I_{21}}(t), \end{aligned}$$

indicates 'information for discrimination' for term t.

Recall that, for the non-symmetric direct divergence $I(P_R, P_{\bar{R}})$, each of its items, $\mathbf{ifd}_I(t)$, can be positive or negative in sign. In contrast, the items of the symmetric divergence $J(P_R, P_{\bar{R}})$ are always non-negative. The non-negativeness is implied by $P_R(t) - P_{\bar{R}}(t) > 0$ and $\log \frac{P_R(t)}{P_{\bar{R}}(t)} > 0$ if $P_R(t) > P_{\bar{R}}(t)$, and by $P_R(t) - P_{\bar{R}}(t) < 0$ and $\log \frac{P_R(t)}{P_{\bar{R}}(t)} < 0$, otherwise.

We can make a formal definition as follows.

Definition 4.2.1 Let $P_R(t) = P(t|H_1)$ and $P_{\bar{R}}(t) = P(t|H_2)$ be discrete probability distributions over $(V, 2^V)$, and derived from sets R and $\bar{R}$, respectively. Assume that both $P_R(t) \ll P_{\bar{R}}(t)$ and $P_R(t) \ll P_{\bar{R}}(t)$ hold when $t \in V$. The information in term t for discrimination on two opposite hypotheses H_1 and H_2 is defined by

$$\begin{aligned} \mathbf{ifd}_J(t) &= \big(P_R(t) - P_{\bar{R}}(t)\big) \log \frac{P_R(t)}{P_{\bar{R}}(t)} \\ &= \big(P(t|H_1) - P(t|H_2)\big) \cdot i(H_1 : H_2|t) \qquad (t \in V), \end{aligned}$$

which is referred as to the (relevance) *discrimination measure* of terms.

4.2.2 Interpretation of Discrimination Measure

It is interesting to notice that two sub-items, $\mathbf{ifd}_{I_{12}}(t)$ and $\mathbf{ifd}_{I_{21}}(t)$, of $\mathbf{ifd}_J(t)$ are always opposite in sign. We can show this through the following simple theorem.

Theorem 4.2.1 For an arbitrary term $t \in V$ satisfying $P_R(t) \cdot P_{\bar{R}}(t) > 0$, we always have

(1) $\mathbf{ifd}_{I_{12}}(t) = 0$ if and only if $P_R(t) = P_{\bar{R}}(t)$, i.e., $\mathbf{ifd}_{I_{21}}(t) = 0$;

(2) $\mathbf{ifd}_{I_{12}}(t) > 0$ if and only if $P_R(t) > P_{\bar{R}}(t)$, i.e., $\mathbf{ifd}_{I_{21}}(t) < 0$.

Proof. If $P_R(t) \neq 0$ and $P_{\bar{R}}(t) \neq 0$, then

(1) $\mathbf{ifd}_{I_{12}}(t) = 0$ if and only if $\log \frac{P_R(t)}{P_{\bar{R}}(t)} = 0$, i.e., $P_R(t) = P_{\bar{R}}(t)$, i.e., $\log \frac{P_{\bar{R}}(t)}{P_R(t)} = 0$, i.e., $\mathbf{ifd}_{I_{21}}(t) = 0$.

(2) $\mathbf{ifd}_{I_{12}}(t) > 0$ if and only if $\log \frac{P_R(t)}{P_{\bar{R}}(t)} > 0$, i.e., $P_R(t) > P_{\bar{R}}(t)$, i.e., $\log \frac{P_{\bar{R}}(t)}{P_R(t)} < 0$, i.e., $\mathbf{ifd}_{I_{21}}(t) < 0$.

The proof is complete.

Consequently, similar to the discussion given in Section 3.4, we have the following interpretations.

- ☞ If $P_R(t) = P_{\bar{R}}(t)$, then the discrimination factors $i(H_1 : H_2|t) = i(H_2 : H_1|t) = 0$, and term t gives us no discrimination information about the relevance classification, and the corresponding quantity $\mathbf{ifd}_J(t) = 0$.

- ☞ If $P_R(t) > P_{\bar{R}}(t)$, then the discrimination factor $i(H_1 : H_2|t) > 0$, term t contributes quantity $\mathbf{ifd}_{I_{12}}(t) = |\mathbf{ifd}_{I_{12}}(t)|$ for supporting the relevant hypothesis H_1. Whereas the discrimination factor $i(H_2 : H_1|t) < 0$, term t contributes quantity $\mathbf{ifd}_{I_{21}}(t) = -|\mathbf{ifd}_{I_{21}}(t)|$ for supporting the non-relevant hypothesis H_2.

  Thus, the positive quantity $\mathbf{ifd}_J(t)$, which is an algebraic sum, is dominated by its positive sub-item $\mathbf{ifd}_{I_{12}}(t)$. The algebraic sum, the difference between the information in t in favour of H_1 and the information in t in favour of H_2, indicates that term t contributes quantity $\mathbf{ifd}_J(t)$ for supporting H_1.

- ☞ If $P_R(t) < P_{\bar{R}}(t)$, then $i(H_1 : H_2|t) < 0$, term t contributes $\mathbf{ifd}_{I_{12}}(t) = -|\mathbf{ifd}_{I_{12}}(t)|$ for supporting H_1. Whereas $i(H_2 : H_1|t) > 0$, term t contributes $\mathbf{ifd}_{I_{21}}(t) = |\mathbf{ifd}_{I_{21}}(t)|$ for supporting H_2.

  Thus, the positive quantity $\mathbf{ifd}_J(t)$ is dominated by its positive sub-item $\mathbf{ifd}_{I_{21}}(t)$. The algebraic sum indicates that term t contributes $\mathbf{ifd}_J(t)$ for supporting H_2.

4.2.3 About Absolute Continuity

Notice that $\mathbf{ifd}_J(t) = 0$ when $P_R(t) = P_{\bar{R}}(t) \neq 0$ (it also has $\mathbf{ifd}_J(t) = 0 \log \frac{0}{0} = 0$ when $P_R(t) = P_{\bar{R}}(t) = 0$). We can thus see that the contribution, to the expected divergence, of terms unrelated to the relevance classification, will be zero. Thus, divergence $J(P_R, P_{\bar{R}})$ satisfies Criterion 2.

It should be emphasized especially that in order to speak of the discrimination information of terms in the sense of the divergence, one must regard distributions $P_R(t)$ and $P_{\bar{R}}(t)$ to be

absolutely continuous with respect to one another. The continuity ensures that $\mathbf{ifd}_J(t) < \infty$ for all terms $t \in V$, that summation $J(P_R, P_{\bar{R}})$ exists, and that Criterion 1 can be satisfied.

In fact, if we desire that both $P_R(t) \ll P_{\bar{R}}(t)$ and $P_{\bar{R}}(t) \ll P_R(t)$ hold simultaneously, then it must have $V^R \subseteq V^{\bar{R}}$ and $V^{\bar{R}} \subseteq V^R$, that is, $V^R = V^{\bar{R}}$. Obviously, requirement $V^R = V^{\bar{R}}$ is a rigid *restriction* that is very difficult to satisfy in realistic IR applications. It is a crucial problem that will be solved in this chapter.

4.3 Solution

Let $\Xi^+ \neq \emptyset$ be the set of relevant sample documents. Let $P_{\Xi+}(t)$ and $P_D(t)$ be the probability distributions over $(V, 2^V)$ as given in Eq.(3.3).

Obviously, $V^{\Xi^+} \subseteq V$. If $V^{\Xi^+} = V$ then $P_{\Xi+}(t) \ll P_D(t)$ and $P_D(t) \ll P_{\Xi+}(t)$ when $t \in V$, and divergence $J(P_{\Xi+}, P_D)$ is meaningful. We can therefore directly apply $J(P_{\Xi+}, P_D)$ to generate the discrimination measure

$$\mathbf{ifd}_J(t) = \big(P_{\Xi+}(t) - P_D(t)\big) \log \frac{P_{\Xi+}(t)}{P_D(t)} \qquad (t \in V),$$

and use the measure as a device to derive the association of terms with the query.

Without losing generality, let us consider $V^{\Xi^+} \subset V$, i.e., V^{Ξ^+} is a proper subset of V. In this case, $P_{\Xi+}(t) \ll P_D(t)$ but $P_D(t) \not\ll P_{\Xi+}(t)$ when $t \in V$, because $P_D(t) \neq 0$ but $P_{\Xi+}(t) = 0$ for $t \in V - V^{\Xi^+}$. Such a case results in $\mathbf{ifd}_J(t) = \big(0 - P_D(t)\big) \log \frac{0}{P_D(t)} = +\infty$ for $t \in V - V^{\Xi^+}$. On the other hand, notice that $\mathbf{ifd}_J(t)$ is meaningful, and so $\mathbf{ifd}_J(t) < +\infty$, for all $t \in V^{\Xi^+}$. Consequently, $\mathbf{ifd}_J(t_i) < \mathbf{ifd}_J(t_j) = +\infty$, where $t_i \in V^{\Xi^+}$ and $t_j \in V - V^{\Xi^+}$, and we under no circumstances say that the contributions to 'summation' $J(P_{\Xi+}, P_D)$ come mostly from terms $t_i \in V^{\Xi^+}$ mathematically. In fact, because $J(P_{\Xi+}, P_D)$ is meaningless (the summation does not exist at all), and because quantities $\mathbf{ifd}_J(t_j) = +\infty$ cannot be compared with each other for all terms $t_j \in V - V^{\Xi^+}$ (thus, Criterion 1 is not satisfied), it does not make sense to generate the discrimination measure $\mathbf{ifd}_J(t)$, and then compare the extent to which each term contributes to $J(P_{\Xi+}, P_D)$.

4.3.1 Solution of Problem

It is almost impossible to have $V^{\Xi^+} = V$, i.e., $P_{\Xi+}(t) \ll P_D(t)$ and $P_D(t) \ll P_{\Xi+}(t)$, in a practical IR context. As mentioned before, $P_{\Xi+}(t)$ and $P_D(t)$ naturally characterize set Ξ^+ and collection D, respectively. Thus one way of finding the solution to the problem might be to change the characterizations conditionally.

More precisely, a direct way to solve the problem is to modify distributions $P_{\Xi+}(t)$ and $P_D(t)$ to respective distributions $P'_{\Xi+}(t)$ and $P'_D(t)$, which are over some domain $V' \supseteq V^{\Xi^+}$ (where it may or may not have $V' = V$). The essential aim of the modification is to produce a meaningful summation $J(P'_{\Xi+}, P'_D)$ so it becomes possible for the comparison of its individual items over V' (i.e., Criterion 1 can be satisfied).

In order to make the modification profitable in the sense that the discrimination information of the candidate terms $t \in V^{\Xi^+}$ can be captured, we wish that the modified distributions characterize the candidate terms in the same way as the original distributions do. If this can

be achieved then $P'_{\Xi+}(t)$ and $P'_D(t)$ can faithfully reflect the same information as contained in $P_{\Xi+}(t)$ and $P_D(t)$, respectively, for the candidate terms.

Also, we wish that the modified distributions are capable of highlighting the contributions made by candidate terms to the modified divergence, that is, the contributions made by terms in $V' - V^{\Xi+}$ should not 'overwhelm' the ones by terms in $V^{\Xi+}$. If this point can be achieved then it is appropriate for us to give a mathematically reasonable explanation that the score function with domain $t \in V'$ can be reduced to having the domain $V^{\Xi+}$.

All of these statements on the modification imply that $P'_{\Xi+}(t)$ and $P'_D(t)$ should satisfy some *conditions*. Before we are able to give the conditions, a '*crux*' term, denoted by t_0, will play a key role throughout the discussions given below, and needs first be set. In the later stages, we will prove that term t_0 can be ignored in a practical IR environment. ('Pay attention in order to ignore'. 'Make concession for the sake of future gains'. Such thoughts may be helpful to understand the strategy we will take for the crux term.)

It is clear that $V^{\Xi+}$ is a finite set, and that $\mathbf{ifd}_J(t)$ is entirely meaningful over $V^{\Xi+}$. So $\{\mathbf{ifd}_J(t)|t \in V^{\Xi+}\}$ is a finite set, and we are able to take an *argument minimum* over $V^{\Xi+}$. Let term t_0 be such an argument minimum, that is,

$$t_0 = arg\min\{\mathbf{ifd}_J(t)|t \in V^{\Xi+}\},$$

which is a *minimal meaningful value* amongst all of the meaningful values $\mathbf{ifd}_J(t)$ for terms $t \in V^{\Xi+}$.

Now, we can give the conditions that $P'_{\Xi+}(t)$ and $P'_D(t)$ should satisfy as follows.

(C1) $P'_{\Xi+}(t)$ and $P'_D(t)$ are absolutely continuous with respect to one another when terms belong to domain $V' \supseteq V^{\Xi+}$;

(C2) Except only one term t_0, $P'_{\Xi+}(t)$ and $P_{\Xi+}(t)$ are identical in domain $V^{\Xi+} - \{t_0\}$, so are $P'_D(t)$ and $P_D(t)$;

(C3) The extents of the contributions made by term t_0 and all terms $t \in V' - V^{\Xi+}$ to divergence $J(P'_{\Xi+}, P'_D)$ can be proven to be never greater than the minimal meaningful value $\mathbf{ifd}_J(t_0)$.

Obviously, Condition (C1) is the minimum requirement enabling the modified distributions to make the corresponding divergence meaningful and so satisfy Criterion 1. In Condition (C2), the coincidence of the modified distributions with the respective original distributions except a point t_0 guarantees that the modified ones almost completely reflect the same information as the original ones. Notice that Condition (C3) requires not only two inequalities to hold, but also the crux term t_0 to be the argument minimum. Under these requirements, we can use a simple score function with the original distributions within domain $V^{\Xi+}$ without caring what t_0 should be.

It is apparent that in this way we may impose some extra *constraints* on distributions $P_{\Xi+}(t)$ and $P_D(t)$ (see Sections 4.4 and 4.5). However, it may still be feasible if the constraints are much weaker than restriction $V^{\Xi+} = V$, and can be satisfied in an IR environment. The method itself is simple and clear: it satisfies the required application and offers a mathematically reasonable interpretation, enabling the divergence measure to be used to generate the discrimination measure.

4.3.2 Modified Discrimination Measure

Assume that the modified distributions $P'_{\Xi+}(t)$ and $P'_D(t)$ satisfying conditions (C1)-(C3) have been found. Then divergence $J(P_{\Xi+}, P_D)$ can be modified to

$$\begin{aligned} J(P'_{\Xi+}, P'_D) &= \sum_{t\in V'} \left(P'_{\Xi+}(t) - P'_D(t)\right) \log \frac{P'_{\Xi+}(t)}{P'_D(t)} = \sum_{t\in V'} \mathbf{ifd}'_J(t) \\ &= \Big[\sum_{t\in V^{\Xi+}-\{t_0\}} \mathbf{ifd}_J(t)\Big] + \mathbf{ifd}'_J(t_0) + \sum_{t\in V'-V^{\Xi+}} \mathbf{ifd}'_J(t), \end{aligned} \tag{4.1}$$

which is meaningful, i.e, the summation exists.

By the foregoing discussion, the modified divergence $J(P'_{\Xi+}, P'_D)$ can be used to measure the difference of the information contained in $P'_{\Xi+}(t)$ and that contained in $P'_D(t)$ about $P'_{\Xi+}(t)$, and vice versa. Thus, each term $t \in V'$ will more or less make contributions to the difference. The measure used to calculate the extent of the contribution of each term can be expressed by

$$\mathbf{ifd}'_J(t) = \begin{cases} \mathbf{ifd}_J(t) & \text{when } t \in V^{\Xi+} - \{t_0\} \\ \mathbf{ifd}'_J(t_0) & \text{when } t = t_0 \in V^{\Xi+} \\ \mathbf{ifd}'_J(t) & \text{when } t \in V' - V^{\Xi+}, \end{cases} \tag{4.2}$$

which is referred to as the *modified discrimination measure* of terms.

4.3.3 Two Inequalities

It can be seen easily that condition (C3) in fact requires two inequalities:

$$\mathbf{ifd}_J(t_0) \geq \mathbf{ifd}'_J(t_0) \quad \text{and} \quad \mathbf{ifd}_J(t_0) \geq \mathbf{ifd}'_J(t_j), \tag{4.3}$$

where $t_0 \in V^{\Xi+}$ and $t_j \in V' - V^{\Xi+}$.

From condition (C2), we can see that the extent of the contributions made by terms $t_i \in V^{\Xi+} - \{t_0\}$ to divergence $J(P'_{\Xi+}, P'_D)$ is equal to the corresponding meaningful values $\mathbf{ifd}_J(t_i)$. As to point t_0, notice that $\mathbf{ifd}_J(t_0)$ is the minimal meaningful value amongst all of meaningful values. Thus, we have

$$\mathbf{ifd}'_J(t_i) = \mathbf{ifd}_J(t_i) \geq \mathbf{ifd}_J(t_0).$$

Consequently, from two inequalities in Eq.(4.3), we can immediately obtain

$$\mathbf{ifd}'_J(t_i) \geq \mathbf{ifd}'_J(t_0) \quad \text{and} \quad \mathbf{ifd}'_J(t_i) \geq \mathbf{ifd}'_J(t_j), \tag{4.4}$$

where $t_i \in V^{\Xi+} - \{t_0\}$ and $t_j \in V' - V^{\Xi+}$.

The inequalities in Eq.(4.4) explicitly indicate that quantity $\mathbf{ifd}'_J(t_i)$ of each term $t_i \in V^{\Xi+}$ will be greater than or equal to quantity $\mathbf{ifd}'_J(t_0)$ of term t_0 and quantities $\mathbf{ifd}'_J(t_j)$ of all terms $t_j \in V' - V^{\Xi+}$, and that the difference between $P'_{\Xi+}(t)$ and $P'_D(t)$ comes mostly from $\mathbf{ifd}'_J(t_i)$ for terms $t_i \in V^{\Xi+} - \{t_0\}$, rather than from terms t_0 and $t_j \in V' - V^{\Xi+}$.

In the next section, we will discuss the score function constructed based on the modified discrimination measure $\mathbf{ifd}'_J(t)$. Notice that we are interested only in the candidate terms $t \in V^{\Xi+}$, and that $\mathbf{ifd}_J(t)$ is meaningful for terms $t \in V^{\Xi+}$ including term t_0, and that $\mathbf{ifd}'_J(t) = \mathbf{ifd}_J(t)$ when $t \in V^{\Xi+} - \{t_0\}$. Thus, we wish to make a simplification so that it is reasonable for us to directly use measure $\mathbf{ifd}_J(t)$ for terms $t \in V^{\Xi+}$ without considering what term t_0 should be. In other words, we wish to be able to ignore the contributions $\mathbf{ifd}'_J(t_0)$ and $\mathbf{ifd}'_J(t_j)$ made by terms t_0 and $t_j \in V' - V^{\Xi+}$, respectively, when we use $\mathbf{ifd}'_J(t)$ as the discrimination measure. We will see that this can easily be achieved if the modified discrimination measure satisfies two inequalities in Eq.(4.3).

Next, a question: whether we can find such modified distributions satisfying conditions (C1)-(C3)? The answer is yes, but it is rather difficult to do. In order to give the reader some sense about the existence of $P'_{\Xi+}(t)$ and $P'_D(t)$, two typical modified methods are proposed in Section 4.6. It should be pointed out that the mathematical interpretations and arguments given in the current chapter are based on the premise that we limit ourselves to consider the discrimination information for only terms $t \in V^{\Xi+}$ in a practical IR environment.

Before the existence of the modified distributions are discussed, let us proceed with our central subject below.

4.4 Association Function $atq_J(t, q)$

As mentioned before, a piece of 'useful information' is viewed as the amount of information in term $t \in V'$ for discriminating two opposite relevance hypotheses. The discrimination measure $\mathbf{ifd}'_J(t)$ can thus be used to measure the amount, and the amount actually provides the extent of the statistical association information of term t with the context of the query. Therefore, a formal definition can be given as follows.

Definition 4.4.1 Let $P_{\Xi+}(t)$ and $P_D(t)$ be discrete probability distributions over $(V, 2^V)$ as expressed in Eq.(3.3). Let $P'_{\Xi+}(t)$ and $P'_D(t)$ be probability distributions over $(V', 2^{V'})$, where $V' \supseteq V^{\Xi+}$, satisfying:

(C1) $P'_{\Xi+}(t) \ll P'_D(t)$ and $P'_D(t) \ll P'_{\Xi+}(t)$ when $t \in V'$,

(C2) $P'_{\Xi+}(t_i) = P_{\Xi+}(t_i)$ and $P'_D(t_i) = P_D(t_i)$ when $t_i \in V^{\Xi+} - \{t_0\}$,

(C3) $\mathbf{ifd}_J(t_0) \geq \mathbf{ifd}'_J(t_0)$ and $\mathbf{ifd}_J(t_0) \geq \mathbf{ifd}'_J(t_j)$ when $t_j \in V' - V^{\Xi+}$,

where term t_0 is the argument minimum of $\mathbf{ifd}_J(t)$ over $t \in V^{\Xi+}$. Then, the association of terms with query q, denoted by $atq_J(t, q)$, can be defined as

$$atq_J(t, q) = \mathcal{Q}(t) \cdot \mathbf{ifd}'_J(t) = \mathcal{Q}(t)\big(P'_{\Xi+}(t) - P'_D(t)\big) \log \frac{P'_{\Xi+}(t)}{P'_D(t)} \qquad (t \in V'),$$

where $\mathcal{Q}(t) \geq 0$ measures the significance of term t concerning query q.

4.5 Score Function $score_J(t)$

A method for selecting the strong associated terms with the query is offered in this section based on the discussions of the divergence given in the previous sections.

In relevance feedback, with Definition 4.4.1, the association score function is defined by

$$score_J(t) = atq_J(t,q) = \mathcal{Q}(t)\cdot \mathbf{ifd}'_J(t)$$
$$= \begin{cases} w_q(t)\cdot \mathbf{ifd}'_J(t) & \text{when } t\in V^q\cap V' \\ \kappa_1 \quad \cdot \mathbf{ifd}'_J(t) & \text{when } t\in V^{\Xi+}-V^q \\ \kappa_2 \quad \cdot \mathbf{ifd}'_J(t) & \text{when } t\in V'-V^{\Xi+}-V^q, \end{cases}$$

where the estimation of $\mathcal{Q}(t)$ was discussed in Section 3.6., and the expression of $\mathbf{ifd}'_J(t)$ is given in Eq.(4.2).

Notice that, like function $score_I(t)$, function $score_J(t)$ only assigns scores to the query terms $t_i\in V^q\cap V^{\Xi+}$, but not to query terms $t_j\in V'-V^{\Xi+}$. However such non-assignments will be very rare.

4.5.1 Reduction of Domain

Recall that, function $score_I(t)$, given in Section 3.6, with domain $t\in V$ can immediately be reduced to the one with domain $t\in V^{\Xi+}$ simply because $\mathbf{ifd}_I(t)=0$ when $t\in V-V^{\Xi+}$ (since $P_{\Xi+}(t)=0$).

However, it is very likely to have $P'_{\Xi+}(t)\cdot P'_D(t)\neq 0$ and $P'_{\Xi+}(t)\neq P'_D(t)$ when $t\in V'-V^{\Xi+}$, and in this case, it has $\mathbf{ifd}'_J(t)\neq 0$. This implies that the contributions to summation $J(P'_{\Xi+}:P'_D)$ may come not only from terms $t\in V^{\Xi+}$, but also from terms $t\in V'-V^{\Xi+}$.

As mentioned before, in practice, all query terms that index documents $d\in D$ can with very high possibility appear in at least one relevant sample document. In this case, the query terms should not belong to domain $V'-V^{\Xi+}$. Thus, we can simply set $\mathcal{Q}(t_j)=\kappa_2$ when $t_j\in V'-V^{\Xi+}$, and $\mathcal{Q}(t_i)=\kappa_1$ when $t_i\in V^{\Xi+}-V^q$. Therefore, when $t_{i_1}\in V^{\Xi+}\cap V^q$, $t_{i_2}\in V^{\Xi+}-V^q$ and $t_j\in V'-V^{\Xi+}$, we have

$$score_J(t_{i_1}) = w_q(t_{i_1})\cdot \mathbf{ifd}'_J(t_{i_1}) \geq \kappa_2\cdot \mathbf{ifd}'_J(t_j) = score_J(t_j),$$
$$score_J(t_{i_2}) = \kappa_1 \quad \cdot \mathbf{ifd}'_J(t_{i_2}) \geq \kappa_2\cdot \mathbf{ifd}'_J(t_j) = score_J(t_j),$$

since $w_q(t_{i_1})\geq \min_{t\in V^q}\{w_q(t)\}\geq \kappa_1\geq\kappa_2\geq 0$ (see Section 3.6) and $\mathbf{ifd}'_J(t_i)\geq \mathbf{ifd}'_J(t_j)\geq 0$ (see condition (C3) and inequalities in Eq.(4.4)). That is, the scores of terms $t_j\in V'-V^{\Xi+}$ will never exceed the scores of terms $t_i\in V^{\Xi+}$ whether t_i is a query term or not. Also, $|S^q|$ is much smaller than $|V^{\Xi+}|$ and we are interested only in the terms with the top scores. Therefore, the score function with domain $t\in V'$ can also be reduced to the one with domain $t\in V^{\Xi+}$:

$$score_J(t) = \mathcal{Q}(t)\big(P'_{\Xi+}(t)-P'_D(t)\big)\log\frac{P'_{\Xi+}(t)}{P'_D(t)}$$
$$= \begin{cases} w_q(t)\cdot\big(P'_{\Xi+}(t)-P'_D(t)\big)\log\frac{P'_{\Xi+}(t)}{P'_D(t)} & \text{when } t\in V^{\Xi+}\cap V^q \\ \kappa_1 \quad \cdot\big(P'_{\Xi+}(t)-P'_D(t)\big)\log\frac{P'_{\Xi+}(t)}{P'_D(t)} & \text{when } t\in V^{\Xi+}-V^q. \end{cases}$$

From the first inequality in Eq.(4.4), we can see that term t_0 makes a minimal contribution to $J(P'_{\Xi+},P'_D)$ among all terms $t\in V^{\Xi+}$. In this case, t_0 is usually not a query term in

practical IR context. Thus, we can set $t_0 \in V^{\Xi+} - V^q$. Therefore, when $t_{i_1} \in V^{\Xi+} \cap V^q$ and $t_{i_2} \in V^{\Xi+} - V^q$, we have

$$\begin{aligned} score_J(t_{i_1}) &= w_q(t_{i_1}) \cdot \mathbf{ifd}'_J(t_{i_1}) \geq \kappa_1 \cdot \mathbf{ifd}'_J(t_0) = score_J(t_0), \\ score_J(t_{i_2}) &= \kappa_1 \quad \cdot \mathbf{ifd}'_J(t_{i_2}) \geq \kappa_1 \cdot \mathbf{ifd}'_J(t_0) = score_J(t_0), \end{aligned}$$

since $\mathbf{ifd}'_J(t_i) \geq \mathbf{ifd}'_J(t_0) \geq 0$ (see condition (C3) and inequalities in Eq.(4.4)). That is, term t_0 obtains the minimal score among all terms $t \in V^{\Xi+}$, and should fall into sub-domain $V^{\Xi+} - S^q$. Therefore, from the application point of view, there is no necessity to consider what term t_0 should be. Notice that $P'_{\Xi+}(t) = P_{\Xi+}(t)$ and $P'_D(t) = P_D(t)$ when $t \in V^{\Xi+}$ except only one term t_0 (see condition (C2)). Thus, we can write a completely equivalent score function by

$$\begin{aligned} score_J(t) &= \mathcal{Q}(t)\big(P_{\Xi+}(t) - P_D(t)\big) \log \frac{P_{\Xi+}(t)}{P_D(t)} \\ &= \begin{cases} w_q(t) \cdot \big(P_{\Xi+}(t) - P_D(t)\big) \log \frac{P_{\Xi+}(t)}{P_D(t)} & \text{when } t \in V^{\Xi+} \cap V^q \\ \kappa_1 \quad \cdot \big(P_R(t) - P_D(t)\big) \log \frac{P_R(t)}{P_D(t)} & \text{when } t \in V^{\Xi+} - V^q, \end{cases} \end{aligned}$$

which is called the association *score* of term t with query q. The estimations of $P_{\Xi+}(t)$ and $P_D(t)$ can be found in Section 3.7.

4.5.2 About Positive Scores

It is important to understand that, in principle, from a higher positive $score_J(t)$ one cannot infer that term t is positively associated with the query. This is because, when $t \in V^{\Xi+}$, $\mathbf{ifd}_{I_{21}}(t)$ can be positive or negative. If $\mathbf{ifd}_{I_{21}}(t) > 0$ (it must be accompanied by $\mathbf{ifd}_{I_{12}}(t) < 0$), then $\mathbf{ifd}_J(t) > 0$ indicates that the algebraic sum is dominated by sub-item $\mathbf{ifd}_{I_{21}}(t)$, and term t contributes quantity $\mathbf{ifd}_J(t)$ for supporting H_2. In this case, the higher the score term t obtains, the more unlikely it is statistically associated with query q. The 'prime culprit' that leads to $\mathbf{ifd}_{I_{21}}(t) > 0$ is $P_{\Xi+}(t) < P_D(t)$.

Fortunately, in practice, we generally have $P_{\Xi+}(t) > P_D(t)$ for all terms $t \in V^{\Xi+}$. Thus, we need not verify $P_{\Xi+}(t) > P_D(t)$ for each of the selected terms. Therefore, each candidate term $t \in V^{\Xi+}$ is assigned a score by function $score_J(t)$, which is always positive, and the terms with top scores should be first considered as the selected terms $t \in S^q$: they actually make the most contributions to the expected divergence $J(P'_{\Xi+}, P'_D)$ among terms $t \in V^{\Xi+}$. Consequently, according to Hypothesis 2 given in Section 3.3, they are more strongly associated with query q than others.

4.5.3 Relationship of Score Functions

The score function can also be written as

$$\begin{aligned} score_J(t) &= \mathcal{Q}(t) \cdot \mathbf{ifd}_J(t) = \mathcal{Q}(t) \cdot \mathbf{ifd}_{I_{12}}(t) + \mathcal{Q}(t) \cdot \mathbf{ifd}_{I_{21}}(t) \\ &= score_{I_{12}}(t) + score_{I_{21}}(t) \qquad (t \in V^{\Xi+}). \end{aligned}$$

From $\mathbf{ifd}_J(t) \geq 0$ we have $score_J(t) \geq 0$ (since $\mathcal{Q}(t) \geq 0$). Also, from $\mathbf{ifd}_{I_{12}}(t) \cdot \mathbf{ifd}_{I_{21}}(t) \leq 0$ we have $score_{I_{12}}(t) \cdot score_{I_{21}}(t) \leq 0$.

Recall we mentioned, in a practical IR context, that we always have $P_{\Xi+}(t) > P_D(t)$ for candidate terms $t \in V^{\Xi+}$. In this case, we have $\mathbf{ifd}_{I_{12}}(t) > 0$ and $\mathbf{ifd}_{I_{21}}(t) < 0$, and then $score_{I_{12}}(t) \geq 0$ and $score_{I_{21}}(t) \leq 0$.

We point out, when term t appears in both relevant and non-relevant documents, that it would contain information for supporting both the relevant hypothesis H_1 and the non-relevant hypothesis H_2. If $P_{\Xi+}(t) > P_D(t)$, then term t contains information $\mathbf{ifd}_{I_{12}}(t) = |\mathbf{ifd}_{I_{12}}(t)|$ for supporting H_1, and also contains information $\mathbf{ifd}_{I_{21}}(t) = -|\mathbf{ifd}_{I_{21}}(t)|$ for supporting H_2. Because the finalized information (the algebraic sum of information that term t conveys) $\mathbf{ifd}_J(t)$ is positive, the information of term t is determined by its positive part $\mathbf{ifd}_{I_{12}}(t)$. The finalized information thus indicates that term t contains information $\mathbf{ifd}_J(t)$ for supporting H_1.

Therefore, function $score_J(t)$ provides the finalized association of term t with the query, namely, it is the algebraic sum of the positive association $score_{I_{12}}(t)$ and the negative association $score_{I_{21}}(t)$. This means that $score_J(t)$ takes into account simultaneously two opposite pieces of relevance information contained in term t, and incorporates them into the association score. Particularly, when $P_{\Xi+}(t) > P_D(t)$, the finalized score indicates that term t is associated with the query to extent $score_J(t)$. In contrast, function $score_I(t) = score_{I_{12}}(t)$, discussed in Section 3.5, offers only the positive association of terms with the query, but ignores the negative association inherent in term t when it also appears in non-relevant documents. So we can see that $score_J(t)$ may measure the extent of association of a term with the query more accurately than $score_I(t)$ does.

4.5.4 In Pseudo-Relevance Feedback Procedure

In pseudo-relevance feedback, i.e., there is no relevance information available, let the sample set Ξ be the top retrieved documents, and all documents in Ξ be viewed as relevant. In this case, we can also construct function $score_J(t)$ using Ξ instead of Ξ^+ as discussed in the case of relevance feedback.

4.6 Two Methods to Modify the Divergence Measure

We are now in a position to investigate the existence of the modified distributions $P'_{\Xi+}(t)$ and $P'_D(t)$ satisfying conditions (C1)-(C3) by giving the concrete forms of $P'_{\Xi+}(t)$ and $P'_D(t)$ satisfying two inequalities in Eq.(4.3).

4.6.1 Method I

Let $t_0 \in V^{\Xi+}$ be an arbitrary term. Notice that we assumed that $|V^d| \geq 2$ for each document $d \in D$. Then, we have $|V^{\Xi+}| \geq 2$ if $V^{\Xi+} \neq \emptyset$. Thus, we obtain $0 < P_{\Xi+}(t_0) < 1$ (see Section 3.7).

In order to construct $P'_{\Xi+}(t)$ and $P'_D(t)$ satisfying conditions (C1) and (C2), let us introduce a *fictitious* '*term*' t^* without containing any information content, and $t^* \notin V$. Let $V' = V^{\Xi+} \cup \{t^*\}$. The strategy adopted here is based on discounting the value of density of $P_{\Xi+}(t_0)$ with a discounting factor $\mu = P_{\Xi+}(t_0)$ (satisfying $0 < \mu < 1$). The discounted value

of density $P_{\Xi+}(t_0) - \mu P_{\Xi+}(t_0) = P_{\Xi+}(t_0) - P^2_{\Xi+}(t_0)$ is restored by redistributing it onto the fictitious term t^*. We may formulate the strategy by the piecewise probability distribution:

$$P'_{\Xi+}(t) = \begin{cases} P_{\Xi+}(t) & \text{when } t \in V' - \{t_0\} - \{t^*\} \\ P^2_{\Xi+}(t_0) & \text{when } t = t_0 \in V' \\ P_{\Xi+}(t_0) - P^2_{\Xi+}(t_0) & \text{when } t = t^* \in V'. \end{cases}$$

A key idea in the strategy is to introduce the fictitious term t^*, and then sum densities $P_D(t)$ for all terms $t \in V - V^{\Xi+}$ onto one point t^* which is assumed not to convey any information. These statements may be further formulated by means of the piecewise probability distribution:

$$P'_D(t) = \begin{cases} P_D(t) & \text{when } t \in V' - \{t_0\} - \{t^*\} \\ P_{\Xi+}(t_0)P_D(t_0) & \text{when } t = t_0 \in V' \\ P_D(t_0) - P_{\Xi+}(t_0)P_D(t_0) + \sum_{t \in V - V^{\Xi+}} P_D(t) & \text{when } t = t^* \in V'. \end{cases}$$

It can be seen readily that $P'_{\Xi+}(t) > 0$ and $P'_D(t) > 0$ hold for every $t \in V'$, and that

$$\begin{aligned} \sum_{t \in V'} P'_{\Xi+}(t) &= \sum_{t \in V' - \{t_0\} - \{t^*\}} P'_{\Xi+}(t) + P'_{\Xi+}(t_0) + P'_{\Xi+}(t^*) \\ &= \sum_{t \in V' - \{t_0\} - \{t^*\}} P_{\Xi+}(t) + P^2_{\Xi+}(t_0) + P_{\Xi+}(t_0) - P^2_{\Xi+}(t_0) \\ &= \sum_{t \in V^{\Xi+} - \{t_0\}} P_{\Xi+}(t) + P_{\Xi+}(t_0) = \sum_{t \in V^{\Xi+}} P_{\Xi+}(t) = 1, \end{aligned}$$

and that

$$\begin{aligned} \sum_{t \in V'} P'_D(t) &= \sum_{t \in V' - \{t_0\} - \{t^*\}} P'_D(t) + P'_D(t_0) + P'_D(t^*) \\ &= \sum_{t \in V' - \{t_0\} - \{t^*\}} P_D(t) + P_{\Xi+}(t_0)P_D(t_0) + P_D(t_0) - P_{\Xi+}(t_0)P_D(t_0) + \sum_{t \in V - V^{\Xi+}} P_D(t) \\ &= \sum_{t \in V^{\Xi+} - \{t_0\}} P_D(t) + P_D(t_0) + \sum_{t \in V - V^{\Xi+}} P_D(t) = \sum_{t \in V} P_D(t) = 1. \end{aligned}$$

That is, $P'_{\Xi+}(t)$ and $P'_D(t)$ satisfy two axioms of probability distribution, they are hence probability distributions over $(V', 2^{V'})$.

Obviously, $P'_{\Xi+}(t)$ and $P'_D(t)$ satisfy conditions (C1) and (C2). Thus, summation $J\big(P'_{\Xi+}, P'_D\big)$ exists. Notice that $V' - V^{\Xi+} = \{t^*\}$. Thus, divergence $J\big(P_{\Xi+}, P_D\big)$ can be modified to $J\big(P'_{\Xi+}, P'_D\big)$ as expressed in Eq.(4.1), in which,

$$\mathbf{ifd}'_J(t_0) = \big(P^2_{\Xi+}(t_0) - P_{\Xi+}(t_0)P_D(t_0)\big) \log \frac{P_{\Xi+}(t_0)}{P_D(t_0)}.$$

$$\mathbf{ifd}'_J(t^*) = \Big[\big(P_{\Xi+}(t_0) - P^2_{\Xi+}(t_0)\big) - \big(P_D(t_0) - P_{\Xi+}(t_0)P_D(t_0) + \sum_{t\in V-V^{\Xi+}} P_D(t)\big)\Big] \times$$

$$\times \log \frac{P_{\Xi+}(t_0) - P^2_{\Xi+}(t_0)}{P_D(t_0) - P_{\Xi+}(t_0)P_D(t_0) + \sum_{t\in V-V^{\Xi+}} P_D(t)}.$$

We can hence immediately write the modified discrimination measure as expressed in Eq.(4.2).

Notice that, in the current modification method, t^* is treated as a fictitious term that does not contain real information content. Thus, t^* is of course impossible to be associated with any given query. So it is clear that there is no need to consider contribution $\mathbf{ifd}'_J(t^*)$ made by term t^* to divergence $J\big(P'_{\Xi+}, P'_D\big)$ at all during a query expansion procedure. On the other hand, it can be seen easily that quantities $\mathbf{ifd}'_J(t)$ are independent of t^* when $t \in V^{\Xi+}$. Thus, our task can be reduced to considering the contributions $\mathbf{ifd}'_J(t)$ made by individual terms $t \in V' - \{t^*\} = V^{\Xi+}$, we need hence only to prove the first inequality given in Eq.(4.3).

The proof that $P'_{\Xi+}(t)$ and $P'_D(t)$ can satisfy condition (C3), that is, that the modified discrimination measure $\mathbf{ifd}'_J(t)$ can satisfy the first inequality in Eq.(4.3), requires Theorem 4.6.1, which is given in Section 10.1. We will see that Theorem 4.6.1 clearly tells us that the difference between $P'_{\Xi+}(t)$ and $P'_D(t)$ over $V' = V^{\Xi+} \cup \{t^*\}$ comes mostly from $\mathbf{ifd}'_J(t)$ for terms $t \in V^{\Xi+} - \{t_0\}$, rather than from term t_0 (since t^* is ignored).

4.6.2 Method II

In Method I, we discussed the issue of modifying the divergence measure through modifying its arguments by means of the discounting factor μ. In fact, in that method, we gave a specific value of μ: $\mu = P_{\Xi+}(t_0)$.

More generally, we can modify the arguments by using the discounting factor μ satisfying $0 < \mu < 1$. Obviously, there are many different strategies to modify distributions $P_{\Xi+}(t)$ and/or $P_D(t)$ so that the modified distributions are absolutely continuous with respect to one another. Notice that $P_D(t) > 0$ for all terms $t \in V$. Thus, for instance, the most natural and simple one probably is to only modify $P_{\Xi+}(t)$ satisfying $P_{\Xi+}(t) > 0$ for $t \in V^{\Xi+}$ to

$$P'_{\Xi+}(t) = \begin{cases} P_{\Xi+}(t) & \text{when } t \in V^{\Xi+} - \{t_0\} \\ \mu P_{\Xi+}(t_0) & \text{when } t = t_0 \in V^{\Xi+} \\ \frac{1-\mu}{|V|-|V^{\Xi+}|} P_{\Xi+}(t_0) & \text{when } t \in V - V^{\Xi+} \end{cases}$$

satisfying $P'_{\Xi+}(t) > 0$ for $t \in V' = V$, where $0 < \mu < 1$ and term $t_0 \in V^{\Xi+}$ is arbitrary. The strategy adopted here is to discount density $P_{\Xi+}(t_0)$ with the discounting factor μ. The discounted density $P_{\Xi+}(t_0) - \mu P_{\Xi+}(t_0)$ is restored by redistributing it evenly onto all terms $t \in V - V^{\Xi+}$. Clearly, $P'_{\Xi+}(t)$ is, in this case, a constant over domain $t \in V - V^{\Xi+}$.

It is readily seen that

$$\sum_{t\in V} P'_{\Xi+}(t) = \sum_{t\in V^{\Xi+}-\{t_0\}} P'_{\Xi+}(t) + P'_{\Xi+}(t_0) + \sum_{t\in V-V^{\Xi+}} P'_{\Xi+}(t)$$

$$= \sum_{t\in V^{\Xi+}-\{t_0\}} P_{\Xi+}(t) + \mu P_{\Xi+}(t_0) + \big(|V| - |V^{\Xi+}|\big)\frac{1-\mu}{|V|-|V^{\Xi+}|} P_{\Xi+}(t_0)$$

$$= \sum_{t \in V^{\Xi+} - \{t_0\}} P_{\Xi+}(t) + \mu P_{\Xi+}(t_0) + (1-\mu) P_{\Xi+}(t_0)$$
$$= \sum_{t \in V^{\Xi+} - \{t_0\}} P_{\Xi+}(t) + P_{\Xi+}(t_0) = \sum_{t \in V^{\Xi+}} P_{\Xi+}(t) = 1.$$

That is, $P'_{\Xi+}(t)$ is a probability distribution on V.

Thus, $P'_{\Xi+}(t)$ and $P'_D(t) = P_D(t)$ satisfy conditions (C1) and (C2). Therefore, the modified divergence $J(P'_{\Xi+}, P'_D)$, as expressed in Eq.(4.1), exists, in which,

$$\mathbf{ifd}'_J(t_0) = \left(\mu P_{\Xi+}(t_0) - P_D(t_0)\right) \log \frac{\mu P_{\Xi+}(t_0)}{P_D(t_0)} \qquad (t_0 \in V^{\Xi+}),$$

$$\mathbf{ifd}'_J(t) = \left(\frac{1-\mu}{|V| - |V^{\Xi+}|} P_{\Xi+}(t_0) - P_D(t)\right) \log \frac{\frac{1-\mu}{|V|-|V^{\Xi+}|} P_{\Xi+}(t_0)}{P_D(t)} \qquad (t \in V - V^{\Xi+}).$$

Then, we can write the modified discrimination measure as given in expression Eq.(4.2).

We will prove the first inequality by Theorem 4.6.2, which is discussed in detail in Section 10.1. We give the proof of the second inequality in [21]. With these two inequalities, we can see that the difference between $P'_{\Xi+}(t)$ and $P_D(t)$ over $V' = V$ comes mostly from $\mathbf{ifd}'_J(t)$ for terms $t \in V^{\Xi+} - \{t_0\}$, rather than for terms t_0 and $t \in V - V^{\Xi+}$.

4.7 Summary

This chapter addresses a formal method based on the basic concept of divergence for automatic query expansion. The meaning of applying divergence to measure the amount of information contained in a term is interpreted, and the discrimination measure is introduced. Because the condition of absolute continuity of the probability distributions with respect to one another is usually not satisfied in a the practical context of IR, this chapter is devoted to a formal analysis and mathematical discussion on the feasibility of applying the divergence to feedback techniques.

¶ A possible way of modifying the discrimination measure for solving the problem is suggested, and the solution is expounded in a general form.

In order to make the modification profitable, the modified distributions $P'_{\Xi+}(t)$ and $P'_D(t)$ should satisfy some conditions. Conditions (C1)-(C3) are discussed, which can make the modified divergence measure: satisfy Criterion 1; almost completely reflect the same information as the original discrimination measure for terms $t \in V^{\Xi+}$; and construct a simple score function.

¶ The existence of the modified discrimination measure satisfying conditions (C1)-(C3) is shown. Two typical methods of modification are formally described.

Taking them in reverse order: the second modification method is very intuitive, and the mathematical thought and theoretical proofs are elegant (see Section 10.1 and [21]). However, for simplification, the original term distributions $P_{\Xi+}(t)$ and $P_D(t)$ must satisfy some extra constraints. Unfortunately, in a realistic IR environment, it is very difficult for us to verify the constraints can be satisfied experimentally.

The first modification method is interesting, and requires a thorough mathematical treatment. As we have seen, for simplification, it needs only constraint $P_{\Xi+}(t_0) \geq P_D(t_0)$, where $t_0 \in V^{\Xi+}$. It is clear that such a constraint is much weaker than constraints required in the second method, and can usually be satisfied in a practical IR environment. The key idea introduced is a fictitious term t^*. Doing it this way not only makes the main concern simple and clear, but also gives a reasonable mathematical interpretation which lets us be able to easily apply the divergence measure to generate the discrimination measure for query expansion.

¶ A modified discrimination measure is naturally introduced based on the modified probability distributions, and the concept of the association of terms with the context of the query is formally defined based on the modified discrimination measure.

¶ The construction of the score function is addressed, and the issue of the reduction of the domain of the score function is formally discussed.

It should be noted especially that, theoretically, a higher positive $score_J(t)$ does not imply that term t must be associated with the query, and condition $P_{\Xi+}(t) > P_D(t)$ should be verified. The reason for this has been clearly explained in the previous sections. Fortunately, in practice, we usually have $P_{\Xi+}(t) > P_D(t)$ for all terms $t \in V^{\Xi+}$. Thus, from the application of view, we may ignore the verification of $P_{\Xi+}(t) > P_D(t)$, and directly use positive $score_J(t)$ for selecting good terms.

In addition, $J(P_R : P_{\bar{R}})$ is symmetric in arguments $P_R(t)$ and $P_{\bar{R}}(t)$. This property might be desirable in some practical applications of IR. However, it requires that $P_R(t) \ll P_{\bar{R}}(t)$ and $P_{\bar{R}}(t) \ll P_R(t)$ for $t \in V$, or $V^R = V^{\bar{R}}$. Such a requirement may be too strong for the IR context. In the next chapter, we will discuss another information measure, which is well-defined for its arguments, i.e., it need not place any requirement on distributions $P_R(t)$ and $P_{\bar{R}}(t)$.

Chapter 5

AQE Based on Information Radius

Information radius was initially introduced to IR theory, as a device for generating the discrimination measure of terms, by Van Rijsbergen in his book [207] in the late seventies. It seems that IR researchers have been comparatively slow to appreciate that an information radius method to the problem of the discrimination information of terms can prove profitable. The study in this chapter is based on Van Rijsbergen's earlier idea, and is a further development and implementation of a methodology initiated there.

In Section 5.1, we intend to give a simple account of the concept of information radius. In Section 5.2, we focus on a detailed discussion of a relevance discrimination measure based on the information radius. In Section 5.3, we consider a symmetric discrimination measure which is a special situation of the discrimination measure. In Section 5.4, we define the concept of association of terms with the context of the query in the sense of the information radius. In Section 5.5, we address the method of constructing a score function and the simplification of the domain of the score function, and illustrate how the score function can be employed in both relevance and pseudo-relevance feedback.

5.1 Information Gain $K(\lambda_1, \lambda_2; P_R, P_{\bar{R}})$

5.1.1 Information Moment

To gain a full appreciation of the discrimination power of the concept of information radius, it is necessary not only to consider its interpretation but also to become acquainted with some other supporting considerations (i.e., some simple properties). An excellent paper about these has been provided by [177]. In addition, it is helpful to become familiar with the concept of information moment and its interpretation. The following is illustrative.

Let $H_1, H_2, ..., H_r$ be competing hypotheses that term t is drawn from the document sets $D_1, D_2, ..., D_r$, respectively. Assume that $P_{D_k}(t) \in \mathcal{P}_n$ defines set D_k with *a priori* probability λ_k ($\lambda_k \geq 0$ for $k = 1, ..., r$ and $\sum_{k=1}^{r} \lambda_k = 1$). Also, let H_* be a hypothesis that term t is drawn from a document set D_* and assume that $P_*(t) \in \mathcal{P}_n$ defines set D_*.

The *information moment* for these r sets $D_1, ..., D_r$ and a set D_* is defined by

$$K_r(\{\lambda_k\}; \{P_{D_k}\} : P_*) = \sum_{k=1}^{r} \lambda_k I(P_{D_k} : P_*) = \sum_{k=1}^{r} \lambda_k I(H_k : H_*|H_k).$$

If we regard λ_k as the probability of $P_{D_k}(t)$ being correct, then the information moment can be interpreted as the expected gain in information on rejecting $P_*(t)$ in favour of $P_{D_k}(t)$ for $k = 1, ..., r$. Particularly, when $r = 1$, we have $\lambda_1 = 1$ and the corresponding information moment is reduced to the directed divergence $K_1(\{1\}; \{P_{D_1}\} : P_*) = I(P_{D_1} : P_*)$.

Now, assume that all $P_{D_1}(t)$, ..., $P_{D_r}(t)$ are known, and let

$$P_\Sigma(t) = \lambda_1 P_{D_1}(t) + ... + \lambda_r P_{D_r}(t).$$

Obviously, we have $P_\Sigma(t) \in \mathcal{P}_n$ by virtue of the convexity of $\mathcal{P}_n$. It was shown [177] that the information moment satisfies the equality

$$K_r(\{\lambda_k\}; \{P_{D_k}\} : P_*) = K_r(\{\lambda_k\}; \{P_{D_k}\} : P_\Sigma) + I(P_\Sigma : P_*).$$

It is clear that the first item on the right side of the equality above is a constant when $\lambda_1, ..., \lambda_r$ and $P_{D_1}(t), ..., P_{D_r}(t)$ are given, and that the second item is the function of distribution $P_*(t) \in \mathcal{P}_n$. Consequently, it can be seen readily that $K_r(\{\lambda_k\}; \{P_{D_k}\} : P_*)$ arrives uniquely at a *minimum* when $P_*(t) = P_\Sigma(t)$, that is,

$$\inf_{P_* \in \mathcal{P}_n} \{K_r(\{\lambda_k\}; \{P_{D_k}\} : P_*)\} = K_r(\{\lambda_k\}; \{P_{D_k}\} : P_\Sigma)$$

since $I(P_\Sigma : P_*) \geq 0$ with equality if and only if $P_\Sigma(t) = P_*(t)$.

5.1.2 Information Radius Measure

If the probability that $P_{D_k}(t)$ should be correct is initially given by λ_k, then the minimum above can be regarded as the expected gain in information on judging which $P_{D_k}(t)$ should be correct.

The *information radius* for these r distributions $P_{D_k}(t)$ with *a priori* probability λ_k, due to Sibson [177], is defined as the minimum, and denoted by

$$K_r(\{\lambda_k\}; \{P_{D_k}\}) = K_r(\{\lambda_k\}; \{P_{D_k}\} : P_\Sigma).$$

Therefore, it can be immediately expressed as

$$\begin{aligned} K_r(\{\lambda_k\}; \{P_{D_k}\}) &= \sum_{k=1}^{r} \lambda_k I(P_{D_k} : P_\Sigma) = \sum_{k=1}^{r} \Big(\lambda_k \sum_{t \in V} P_{D_k}(t) \log \frac{P_{D_k}(t)}{P_\Sigma(t)}\Big) \\ &= \sum_{t \in V} \sum_{k=1}^{r} \lambda_k P_{D_k}(t) \log \frac{P_{D_k}(t)}{\lambda_1 P_{D_1}(t) + ... + \lambda_r P_{D_r}(t)}. \end{aligned}$$

It can easily be seen that $K_r(\{\lambda_k\}; \{P_{D_k}\}) \geq 0$ with equality if and only if $P_{D_{k_1}}(t) = ... = P_{D_{k_s}}(t)$, in which $\lambda_{k_l} > 0$, where $l = 1, ..., s$, $1 \leq s \leq r$ (that is, it vanishes if and only if those $P_{D_{k_l}}(t)$, for which the corresponding coefficient λ_{k_l} are not equal to zero, are identical).

Furthermore, for r *disjoint* probability distributions[1], the information radius is reduced to the entropy of its *a priori* probability distribution $P_\lambda = \{\lambda_1, \lambda_2, ..., \lambda_r\}$:

[1]The r probability distributions $P_{D_k}(t)$, $k = 1, ..., r$, are said to be *disjoint* if $P_{D_k}(t) \geq 0$ when $t \in V_k$ and $P_{D_k}(t) = 0$ when $t \notin V_k$, where $V_1, ..., V_r$ is a partition of V, i.e., $V = V_1 \cup ... \cup V_r$ and $V_k \cap V_{k'} = \emptyset$ $(1 \leq k, k' \leq r; k \neq k')$.

$$
\begin{aligned}
K_r(\{\lambda_k\};\{P_{D_k}\}) &= \Big(\sum_{t\in V_1} + \ldots + \sum_{t\in V_r}\Big)\sum_{k=1}^{r}\lambda_k P_{D_k}(t)\log\frac{P_{D_k}(t)}{\lambda_1 P_{D_1}(t)+\ldots+\lambda_r P_{D_r}(t)} \\
&= \sum_{t\in V_1}\lambda_1 P_{D_1}(t)\log\frac{P_{D_1}(t)}{\lambda_1 P_{D_1}(t)} + \ldots + \sum_{t\in V_r}\lambda_r P_{D_r}(t)\log\frac{P_{D_r}(t)}{\lambda_r P_{D_r}(t)} \\
&= -\lambda_1\log\lambda_1 - \ldots - \lambda_r\log\lambda_r = H(P_\lambda).
\end{aligned}
$$

Notice that, when $\lambda_1\cdot\lambda_2\cdot\ldots\cdot\lambda_r \neq 0$, we have $\lambda_1 P_{D_1}(t)+\lambda_2 P_{D_2}(t)+\ldots+\lambda_r P_{D_r}(t) = 0$ if and only if $P_{D_1}(t) = P_{D_2}(t) = \ldots = P_{D_r}(t) = 0$. Thus, if we assume that $\lambda_k > 0$ for $k = 1,\ldots,r$, then $P_{D_k}(t) \ll P_\Sigma(t)$ holds for $k = 1,\ldots,r$. Therefore, under the assumption, the information radius can be used to compare with arbitrary term distributions over $(V, 2^V)$. Because of this outstanding property, the information radius appears to be of some general interest. There are many practical IR contexts which may consider to apply information radius as a divergence measure, in particular, in situations where an *a priori* probability distribution in the sense of Bayesian statistics is needed.

5.1.3 A Particular Situation

Let us now consider a particular situation where $r = 2$. Let H_1 and H_2 be two opposite hypotheses that term t is drawn from sets $D_1 = R$ and $D_2 = \bar{R}$, respectively. Assume that $P_R(t)$ and $P_{\bar{R}}(t)$ are the term probability distributions over $(V, 2^V)$ under the hypotheses. Also, let H_Σ be a hypothesis that term t is drawn from set $R \cup \bar{R} = D$ characterized by the term probability distribution $P_\Sigma(t) = \lambda_1 P_R(t) + \lambda_2 P_{\bar{R}}(t)$ over $(V, 2^V)$ under the hypothesis. Denote the corresponding information radius as

$$
K(\lambda_1,\lambda_2;P_R,P_{\bar{R}}) = \lambda_1 I(P_R : P_\Sigma) + \lambda_2 I(P_{\bar{R}} : P_\Sigma),
$$

which can be viewed as the expected divergence between distributions $P_R(t)$ and $P_{\bar{R}}(t)$. Based on the interpretation of the information gain given in Section 3.2, if we view λ_1 and λ_2 as the initial probabilities that the respective distributions $P_R(t)$ and $P_{\bar{R}}(t)$ are correct, then the information radius can be interpreted as the expected gain in information on discrimination rejecting $P_\Sigma(t)$ in favour of $P_R(t)$ and $P_{\bar{R}}(t)$ [92].

In applying the information radius to the relevance classification, *a priori* probabilities λ_1 and λ_2 should be given beforehand. The choice of *a priori* probabilities depends on a specific retrieval strategy itself. Generally, for instance, for a given query, if one classifies D into sets R and $\bar{R}$, then, clearly, a natural and reasonable way is to assign $\lambda_1 = \frac{|R|}{|D|}$ and $\lambda_2 = \frac{|\bar{R}|}{|D|}$.

It is readily shown that property $0 \leq K(\lambda_1,\lambda_2;P_R,P_{\bar{R}}) \leq 1$ holds. In fact, by the definition, the lower bound $K(\lambda_1,\lambda_2;P_R,P_{\bar{R}}) \geq 0$ holds as pointed out earlier for the general case, whereas the upper bound can be shown by

$$
\begin{aligned}
&K(\lambda_1,\lambda_2;P_R,P_{\bar{R}}) \\
&= \sum_{t\in V}\Big(\lambda_1 P_R(t)\log\frac{P_R(t)}{\lambda_1 P_R(t)+\lambda_2 P_{\bar{R}}(t)} + \lambda_2 P_{\bar{R}}(t)\log\frac{P_{\bar{R}}(t)}{\lambda_1 P_R(t)+\lambda_2 P_{\bar{R}}(t)}\Big)
\end{aligned}
$$

$$
\begin{aligned}
&= \sum_{t\in V}\Big(\lambda_1 P_R(t)\log\frac{\lambda_1 P_R(t)}{\lambda_1 P_R(t)+\lambda_2 P_{\bar{R}}(t)} + \lambda_2 P_{\bar{R}}(t)\log\frac{\lambda_2 P_{\bar{R}}(t)}{\lambda_1 P_R(t)+\lambda_2 P_{\bar{R}}(t)}\Big) \\
&\quad - \Big(\lambda_1\sum_{t\in V} P_R(t)\log\lambda_1 + \lambda_2\sum_{t\in V} P_{\bar{R}}(t)\log\lambda_2\Big) \\
&\leq 0 + 0 - \big(\lambda_1\log\lambda_1 + \lambda_2\log\lambda_2\big) \leq 1
\end{aligned}
$$

since from calculus we can easily prove that $-\frac{1}{2} \leq x\log x \leq 0$ when $x \in [0,1]$. More of its properties are to be found in [177].

To conclude this section, it is interesting to repeat the important role the composite or, intermediary, distribution $P_\Sigma(t)$ plays in the arguments of the information radius. It is the composite distribution that makes the information moment reach the minimum and then to become the information radius. It is the composite distribution, in which individual component distributions are absolutely continuous with respect to it, that ensures that the information radius always exists, and, due to such an outstanding property, is applicable to wide research areas, particularly, in IR.

5.2 Discrimination Measure $\mathbf{ifd}_K(t)$

Information radius $K\big(\lambda_1,\lambda_2;P_R,P_{\bar{R}}\big)$ is well-defined in comparison with the directed divergence $I\big(P_R:P_{\bar{R}}\big)$ and divergence $J\big(P_R:P_{\bar{R}}\big)$ because both $P_R(t)$ and $P_{\bar{R}}(t)$ are always absolutely continuous with respect to $P_\Sigma(t)$ unconditionally. In what follows, we will always assume that $\lambda_1 \neq 0$ and $\lambda_2 \neq 0$.

5.2.1 Definition of Discrimination Measure

The information radius consists of a sum of items:

$$K\big(\lambda_1,\lambda_2;P_R,P_{\bar{R}}\big) = \sum_{t\in V}\mathbf{ifd}_K(t).$$

Each item

$$\mathbf{ifd}_K(t) = \lambda_1\mathbf{ifd}_{I_{1\Sigma}}(t) + \lambda_2\mathbf{ifd}_{I_{2\Sigma}}(t),$$

which can be positive or negative, indicates '**i**nformation **f**or **d**iscrimination' for each term t. Two sub-items of $\mathbf{ifd}_K(t)$:

$$
\begin{aligned}
\mathbf{ifd}_{I_{1\Sigma}}(t) &= P_R(t)\log\frac{P_R(t)}{\lambda_1 P_R(t)+\lambda_2 P_{\bar{R}}(t)} = P(t|H_1)i(H_1:H_\Sigma|t), \\
\mathbf{ifd}_{I_{2\Sigma}}(t) &= P_{\bar{R}}(t)\log\frac{P_{\bar{R}}(t)}{\lambda_1 P_R(t)+\lambda_2 P_{\bar{R}}(t)} = P(t|H_2)i(H_2:H_\Sigma|t),
\end{aligned}
$$

can also be positive or negative. Consequently, we can make a formal definition as follows.

Definition 5.2.1 Let $P_R(t) = P(t|H_1)$, $P_{\bar{R}}(t) = P(t|H_2)$, and $P_\Sigma(t) = \lambda_1 P_R(t) + \lambda_2 P_{\bar{R}}(t) = P(t|H_\Sigma)$ be discrete probability distributions over $(V, 2^V)$, and derived from sets R, $\bar{R}$ and

$R \cup \bar{R}$, respectively. The information of term t for discrimination on two opposite hypotheses H_1 and H_2 is defined by

$$\begin{aligned}\mathbf{ifd}_K(t) &= \lambda_1 P_R(t) \log \frac{P_R(t)}{\lambda_1 P_R(t) + \lambda_2 P_{\bar{R}}(t)} + \lambda_2 P_{\bar{R}}(t) \log \frac{P_{\bar{R}}(t)}{\lambda_1 P_R(t) + \lambda_2 P_{\bar{R}}(t)} \\ &= \lambda_1 P(t|H_1) i(H_1 : H_\Sigma|t) + \lambda_2 P(t|H_2) i(H_2 : H_\Sigma|t) \qquad (t \in V),\end{aligned}$$

which is referred as to the (relevance) *discrimination measure* of terms.

5.2.2 Interpretation of Discrimination Measure

Similar to Theorem 4.2.1, it is interesting to point out the following theorem.

Theorem 5.2.1 For an arbitrary term $t \in V$ satisfying $P_R(t) \cdot P_{\bar{R}}(t) > 0$, we always have

(1) $\mathbf{ifd}_{I_{1\Sigma}}(t) = 0$ if and only if $P_R(t) = P_{\bar{R}}(t)$, i.e., $\mathbf{ifd}_{I_{2\Sigma}}(t) = 0$;

(2) $\mathbf{ifd}_{I_{1\Sigma}}(t) > 0$ if and only if $P_R(t) > P_{\bar{R}}(t)$, i.e., $\mathbf{ifd}_{I_{2\Sigma}}(t) < 0$.

Proof. If $P_R(t) \neq 0$ and $P_{\bar{R}}(t) \neq 0$, then

(1) $\mathbf{ifd}_{I_{1\Sigma}}(t) = 0$ if and only if $P_R(t) = \lambda_1 P_R(t) + \lambda_2 P_{\bar{R}}(t) = \lambda_1 P_R(t) + (1 - \lambda_1) P_{\bar{R}}(t)$, i.e., $(1 - \lambda_1) P_R(t) = (1 - \lambda_1) P_{\bar{R}}(t)$, i.e., $P_R(t) = P_{\bar{R}}(t)$, i.e., $P_{\bar{R}}(t) = P_R(t)$, i.e., $(1 - \lambda_2) P_{\bar{R}}(t) = (1 - \lambda_2) P_R(t)$, i.e., $P_{\bar{R}}(t) = (1 - \lambda_2) P_R(t) + \lambda_2 P_{\bar{R}}(t) = \lambda_1 P_R(t) + \lambda_2 P_{\bar{R}}(t)$ if and only if $\mathbf{ifd}_{I_{2\Sigma}}(t) = 0$.

(2) $\mathbf{ifd}_{I_{1\Sigma}}(t) > 0$ if and only if $P_R(t) > \lambda_1 P_R(t) + \lambda_2 P_{\bar{R}}(t) = \lambda_1 P_R(t) + (1 - \lambda_1) P_{\bar{R}}(t)$, i.e., $(1 - \lambda_1) P_R(t) > (1 - \lambda_1) P_{\bar{R}}(t)$, i.e., $P_R(t) > P_{\bar{R}}(t)$, i.e., $P_{\bar{R}}(t) < P_R(t)$, i.e., $(1 - \lambda_2) P_{\bar{R}}(t) < (1 - \lambda_2) P_R(t)$, i.e., $P_{\bar{R}}(t) < (1 - \lambda_2) P_R(t) + \lambda_2 P_{\bar{R}}(t) = \lambda_1 P_R(t) + \lambda_2 P_{\bar{R}}(t)$ if and only if $\mathbf{ifd}_{I_{2\Sigma}}(t) < 0$. The proof is complete.

From Theorem 5.2.1, we see that, for $t \in V$, $\mathbf{ifd}_{I_{1\Sigma}}(t)$ and $\mathbf{ifd}_{I_{2\Sigma}}(t)$ are opposite in sign. Thus, similar to the discussion given in Section 3.4, we have the following interpretations.

- ☞ If $P_R(t) = P_{\bar{R}}(t)$, then the discrimination factors $i(H_1 : H_\Sigma|t) = i(H_2 : H_\Sigma|t) = 0$, and term t gives us no discrimination information about the relevance classification, and the corresponding quantity $\mathbf{ifd}_K(t) = 0$.

- ☞ If $P_R(t) > P_{\bar{R}}(t)$, then $(1 - \lambda_1) P_R(t) > (1 - \lambda_1) P_{\bar{R}}(t)$, i.e., $P_R(t) > \lambda_1 P_R(t) + \lambda_2 P_{\bar{R}}(t)$, and the discrimination factor $i(H_1 : H_\Sigma|t) > 0$, term t contributes quantity $\mathbf{ifd}_{I_{1\Sigma}}(t) = |\mathbf{ifd}_{I_{1\Sigma}}(t)|$ for supporting the relevant hypothesis H_1. Whereas from $(1 - \lambda_2) P_{\bar{R}}(t) < (1 - \lambda_2) P_R(t)$, we have $P_{\bar{R}}(t) < \lambda_1 P_R(t) + \lambda_2 P_{\bar{R}}(t)$, and the discrimination factor $i(H_2 : H_\Sigma|t) < 0$, term t contributes quantity $\mathbf{ifd}_{I_{2\Sigma}}(t) = -|\mathbf{ifd}_{I_{2\Sigma}}(t)|$ for supporting the non-relevant hypothesis H_2.

  Thus, if $\mathbf{ifd}_K(t) > 0$, the weighted algebraic sum is dominated by its positive subitem $\mathbf{ifd}_{I_{1\Sigma}}(t)$, and term t contributes quantity $\mathbf{ifd}_K(t)$ for supporting H_1; if $\mathbf{ifd}_K(t) < 0$, the weighted algebraic sum is dominated by its negative sub-item $\mathbf{ifd}_{I_{2\Sigma}}(t)$, and term t contributes quantity $\mathbf{ifd}_K(t)$ for supporting H_2.

- ☞ If $P_R(t) < P_{\bar{R}}(t)$, then $P_R(t) < \lambda_1 P_R(t) + \lambda_2 P_{\bar{R}}(t)$, and $i(H_1 : H_\Sigma|t) < 0$, term t contributes $\mathbf{ifd}_{I_{1\Sigma}}(t) = -|\mathbf{ifd}_{I_{1\Sigma}}(t)|$ for supporting H_1. Whereas $P_{\bar{R}}(t) > \lambda_1 P_R(t) +$

$\lambda_2 P_{\bar{R}}(t)$ and $i(H_2 : H_\Sigma|t) > 0$, term t contributes $\mathbf{ifd}_{I_{2\Sigma}}(t) = |\mathbf{ifd}_{I_{2\Sigma}}(t)|$ for supporting H_2.

Thus, if $\mathbf{ifd}_K(t) > 0$, the weighted algebraic sum is dominated by its positive sub-item $\mathbf{ifd}_{I_{2\Sigma}}(t)$, and term t contributes $\mathbf{ifd}_K(t)$ for supporting H_2; if $\mathbf{ifd}_K(t) < 0$, the weighted algebraic sum is dominated by its negative sub-item $\mathbf{ifd}_{I_{1\Sigma}}(t)$, and term t contributes $\mathbf{ifd}_K(t)$ for supporting H_1.

5.2.3 About Absolute Continuity

Notice that $\mathbf{ifd}_K(t) = 0$ for term t that appears in both sets R and $\bar{R}$ with an equal probability, $P_R(t) = P_{\bar{R}}(t)$, (it also has $\mathbf{ifd}_K(t) = \lambda_1 0 \log \frac{0}{0} + \lambda_2 0 \log \frac{0}{0} = 0$ when $P_R(t) = P_{\bar{R}}(t) = 0$). We can thus see that the contribution, to the expected divergence, of terms unrelated to the relevance classification, will be zero. Thus, the information radius satisfies Criterion 2. This means that $K(\lambda_1, \lambda_2; P_R, P_{\bar{R}})$ emphasizes the importance of those terms with variant probabilities within sets R and $\bar{R}$.

Recall that, in the application of measure $I(P_R : P_{\bar{R}})$, we required that $P_R(t) \ll P_{\bar{R}}(t)$, or $V^R \subseteq V^{\bar{R}}$. Also, in the application of measure $J(P_R, P_{\bar{R}})$, we required that $P_R(t) \ll P_{\bar{R}}(t)$ and $P_{\bar{R}}(t) \ll P_R(t)$, or $V^R = V^{\bar{R}}$. In contrast, we need not make any requirements for $P_R(t)$ and $P_{\bar{R}}(t)$ in the application of measure $K(\lambda_1, \lambda_2; P_R, P_{\bar{R}})$. In fact, properties $P_R(t) \ll P_\Sigma(t)$ and $P_{\bar{R}}(t) \ll P_\Sigma(t)$ inherent in the information radius ensure that, for each term $t \in V$, sub-items $\mathbf{ifd}_{I_{k\Sigma}}(t) < \infty$ for $k = 1, 2$. Therefore, they are meaningfully weighted summed with weight λ_i, and thus, item $\mathbf{ifd}_K(t)$ exists. In the end, the summation over individual items offers the expected divergence between $P_R(t)$ and $P_{\bar{R}}(t)$.

Consequently, by means of an intermediary composite distribution $P_\Sigma(t)$, one can measure the expected divergence between distributions $P_R(t)$ and $P_{\bar{R}}(t)$. This view appears to be appealing because when we cannot use the divergence measures $I(P_R : P_{\bar{R}})$ or $J(P_R : P_{\bar{R}})$ to directly elicit the extent of divergence of $P_{\bar{R}}(t)$ from $P_R(t)$, we would use measure $K(\lambda_1, \lambda_2; P_R, P_{\bar{R}})$ to indirectly capture it instead.

In addition, $K(\lambda_1, \lambda_2; P_R, P_{\bar{R}})$ is not symmetric in arguments $P_R(t)$ and $P_{\bar{R}}(t)$, neither in λ_1 and λ_2. It may be desirable to have a symmetrical divergence measure, which is meaningful in terms of information radius, when there is no particular reason to emphasize either $P_R(t)$ or $P_{\bar{R}}(t)$. A symmetric divergence measure $K(\lambda_1, \lambda_2; P_R, P_{\bar{R}})$ can be easily introduced, which is discussed below.

5.3 Symmetric Discriminant Measure

Now, consider a more particular situation which involves two probability distributions with equal *a priori* probability, $\lambda_1 = \lambda_2 = \frac{1}{2}$, and denote the corresponding information radius by $K(P_R, P_{\bar{R}})$. Thus, we further obtain

$$K(P_R, P_{\bar{R}}) = K(\frac{1}{2}, \frac{1}{2}; P_R, P_{\bar{R}}) = \sum_{t \in V} \mathbf{ifd}_K(t) = \frac{1}{2} \sum_{t \in V} \left(\mathbf{ifd}_{I_{1\Sigma}}(t) + \mathbf{ifd}_{I_{2\Sigma}}(t)\right),$$

in which, for each term $t \in V$, sub-items

$$\mathbf{ifd}_{I_{1\Sigma}}(t) = P_R(t) \log \frac{P_R(t)}{\frac{1}{2}P_R(t) + \frac{1}{2}P_{\bar{R}}(t)} = P(t|H_1) i(H_1 : H_\Sigma|t),$$

$$\mathbf{ifd}_{I_{2\Sigma}}(t) = P_{\bar{R}}(t) \log \frac{P_{\bar{R}}(t)}{\frac{1}{2}P_R(t) + \frac{1}{2}P_{\bar{R}}(t)} = P(t|H_2)i(H_2 : H_\Sigma|t).$$

Obviously, $K(P_R, P_{\bar{R}})$ is symmetric with respect to $P_R(t)$ and $P_{\bar{R}}(t)$.

In the case where the two distributions are disjoint, the information radius $K(P_R, P_{\bar{R}})$ can be reduced to unity. In fact, it can be found that

$$\begin{aligned} K(P_R, P_{\bar{R}}) &= -\lambda_1 \log \lambda_1 - \lambda_2 \log \lambda_2 \\ &= -\frac{1}{2}\log\frac{1}{2} - \frac{1}{2}\log\frac{1}{2} = \frac{1}{2}\log 2 + \frac{1}{2}\log 2 = 1. \end{aligned}$$

If the two distributions *overlap*[2] over some subset $\Gamma \subseteq V^R \cap V^{\bar{R}}$, then the divergence would drop sharply. Particularly, in an extreme case where they overlap over vocabulary V, the divergence vanishes, i.e., $K(P_R, P_R) = 0$.

Thus, it can clearly be seen that the contribution of a term to the expected divergence is completely dependent on the densities of the term. The greater the difference between the densities, the larger the divergence $P_{\bar{R}}(t)$ from $P_R(t)$ at point t, the more the contribution $|\mathbf{ifd}_K(t)|$ term t makes. This characteristic is essential to $K(P_R, P_{\bar{R}})$, as discussed for $K(\lambda_1, \lambda_2; P_R, P_{\bar{R}})$.

Some more interesting discussions on the symmetric discrimination measure is offered in Section 10.2.

5.4 Association Function $atq_K(t, q)$

Let $\Xi^+ \neq \emptyset$ and $\Xi^- \neq \emptyset$ ($\Xi^+ \cup \Xi^- = \Xi$ and $\Xi^+ \cap \Xi^- = \emptyset$) be the respective sets of top relevant and non-relevant sample documents obtained based on the user's assessment in a relevance feedback procedure. Notice that Ξ^- is the non-relevant sample set, and should not be viewed as an approximation of the non-relevant set $\bar{R}$. Let $P_{\Xi+}(t)$ be a term distribution over $(V, 2^V)$ defining set Ξ^+, where $P_{\Xi+}(t) > 0$ when $t \in V^{\Xi^+}$ and $P_{\Xi+}(t) = 0$ when $t \in V - V^{\Xi^+}$. Let $P_{\Xi-}(t)$ be a term distribution over $(V, 2^V)$ defining set Ξ^-, where $P_{\Xi-}(t) > 0$ when $t \in V^{\Xi^-}$ and $P_{\Xi-}(t) = 0$ when $t \in V - V^{\Xi^-}$.

Distributions $P_{\Xi+}(t)$ and $P_{\Xi-}(t)$ are not necessarily absolutely continuous with respect to one another: they are both absolutely continuous with respect to the composite distribution $P_\Sigma(t) = \lambda_1 P_{\Xi+}(t) + \lambda_2 P_{\Xi-}(t)$. Therefore, we can apply the information radius $K(\lambda_1, \lambda_2; P_{\Xi+}, P_{\Xi-})$ to query expansion. This may be an effective way to solve a fairly sticky problem when $P_{\Xi+}(t) \ll P_{\Xi-}(t)$ and/or $P_{\Xi-}(t) \ll P_{\Xi+}(t)$ do not hold in a practical feedback environment.

As we know, the extent of association of term t with the context of the query can be measured by the power of discrimination of term t in favour of relevant hypothesis H_1 against non-relevant hypothesis H_2. The discrimination measure $\mathbf{ifd}_K(t)$ can be used to measure the power. Thus, a formal definition can be made as follows.

Definition 5.4.1 Let $P_{\Xi+}(t)$ and $P_{\Xi-}(t)$ be discrete probability distributions over $(V, 2^V)$,

[2]Two probability distributions are said to *overlap* over some sub-domain $\Gamma \subseteq V$ if their densities coincide over Γ. Particularly, when they overlap over the whole domain, $P_R(t) = P_{\bar{R}}(t)$ for all $t \in V$.

and derived from sets Ξ^+ and Ξ^-, respectively. The association of term t with query q, denoted by $atq_K(t,q)$, is defined as

$$atq_K(t,q) = \mathcal{Q}(t) \cdot \mathbf{ifd}_K(t) = \mathcal{Q}(t)\big(\lambda_1 P_{\Xi^+}(t) \log \frac{P_{\Xi^+}(t)}{\lambda_1 P_{\Xi^+}(t) + \lambda_2 P_{\Xi^-}(t)} + \lambda_2 P_{\Xi^-}(t) \log \frac{P_{\Xi^-}(t)}{\lambda_1 P_{\Xi^+}(t) + \lambda_2 P_{\Xi^-}(t)}\big) \qquad (t \in V),$$

where $\mathcal{Q}(t) \geq 0$ measures the significance of terms $t \in V$ concerning query q.

5.5 Score Function $score_K(t)$

A method to select strongly associated terms with the query is offered in this section based on the discussions on the information radius given in the previous sections.

In relevance feedback, with Definition 5.4.1, the association score function is defined by

$$score_K(t) = atq_K(t,q) = \mathcal{Q}(t) \cdot \mathbf{ifd}_K(t) \qquad (t \in V),$$

where the estimation of $\mathcal{Q}(t)$ was discussed in Section 3.6.

Notice that function $score_K(t)$ also assigns scores to the query terms in $V^q \cap V^{\Xi^+}$, whereas the query terms in $V - V^{\Xi^+}$ might be ignored.

Notice also that the method proposed in this section is not concerned with treating the situation where $\Xi^+ = \emptyset$ and $\Xi^- = \Xi$ (i.e., $\lambda_1 = 0$ and $\lambda_2 = 1$), that is, where no positive relevance information is available and all sample documents are found to be non-relevant. In this case, the user should be required to reformulate his query and submit it to the retrieval system to produce an effective sample set. For the situation where $\Xi^+ = \Xi$ and $\Xi^- = \emptyset$ (i.e., $\lambda_1 = 1$ and $\lambda_2 = 0$), that is, all sample documents are justified to be relevant, the user can terminate his search if he is satisfied that he has found enough documents relevant to his information need. Otherwise, for obtaining more relevant documents, he can enter an iterative (pseudo-relevance) feedback loop by taking an extra 'non-relevant' sample set $\aleph$ and merging it into the sample set Ξ. Thus, we can conduct the selection of good terms as in the situation of pseudo-relevance feedback discussed in Subsection 5.5.5.

5.5.1 Reduction of Domain

Notice that $\Xi = \Xi^+ \cup \Xi^-$ and the whole domain can be partitioned into four sub-domains:

$$V = \big(V^{\Xi^+} \cap V^{\Xi^-}\big) \cup \big(V^{\Xi^+} - V^{\Xi^-}\big) \cup \big(V^{\Xi^-} - V^{\Xi^+}\big) \cup \big(V - V^{\Xi}\big).$$

Thus, the discrimination measure can correspondingly be decomposed as

$$\mathbf{ifd}_K(t) = \begin{cases} \lambda_1 \mathbf{ifd}_{I_{1\Sigma}}(t) + \lambda_2 \mathbf{ifd}_{I_{2\Sigma}}(t) \neq 0 & \text{when } t \in V^{\Xi^+} \cap V^{\Xi^-} \\ \lambda_1 \mathbf{ifd}_{I_{1\Sigma}}(t) + \lambda_2 0 \log \frac{0}{0} > 0 & \text{when } t \in V^{\Xi^+} - V^{\Xi^-} \\ \lambda_1 0 \log \frac{0}{0} + \lambda_2 \mathbf{ifd}_{I_{2\Sigma}}(t) > 0 & \text{when } t \in V^{\Xi^-} - V^{\Xi^+} \\ \lambda_1 0 \log \frac{0}{0} + \lambda_2 0 \log \frac{0}{0} = 0 & \text{when } t \in V - V^{\Xi}. \end{cases}$$

When $t \in V - V^{\Xi}$, $P_{\Xi+}(t) = P_{\Xi-}(t) = 0$ and $\mathbf{ifd}_K(t) = 0$. In this case, term t does not give us any discrimination information for the relevance classification, and $score_K(t) = 0$. Thus, it is not necessary to consider terms in $V - V^{\Xi+}$, and the score function with domain $t \in V$ can immediately be reduced to the one with domain $t \in V^{\Xi}$.

When $t \in V^{\Xi-} - V^{\Xi+} \subseteq V^{\Xi}$, we have $P_{\Xi-}(t) > P_{\Xi+}(t) = 0$ and $\mathbf{ifd}_K(t) = \lambda_2 \mathbf{ifd}_{I_{2\Sigma}}(t) = \lambda_2 P_{\Xi-}(t) \log \frac{1}{\lambda_2} > 0$. Thus, term t contributes $\mathbf{ifd}_{I_{1\Sigma}}(t) = 0$ for supporting H_1, and contributes $\mathbf{ifd}_{I_{2\Sigma}}(t) > 0$ for supporting H_2. Because $\mathbf{ifd}_K(t) > 0$, the weighted algebraic sum is determined by its positive sub-item $\mathbf{ifd}_{I_{2\Sigma}}(t)$, and term t contributes quantity $\mathbf{ifd}_K(t)$ for supporting H_2. In other words, when terms appear in only the non-relevant sample documents, the terms will not offer any statistical information for supporting the relevant hypothesis. Conversely, they provide fully positive information for supporting the non-relevant hypothesis. Such kind of terms might be informative, but would not be associated with the query. Therefore, we should be concerned only with those terms that appear in at least one relevant sample document, and throw all terms $t \in V^{\Xi-} - V^{\Xi+} = V^{\Xi} - V^{\Xi+}$ away. Consequently, the score function with domain $t \in V^{\Xi}$ can further be reduced to the one with domain $t \in V^{\Xi+}$, that is,

$$score_K(t) = \mathcal{Q}(t)\big(\lambda_1 P_{\Xi+}(t) \log \frac{P_{\Xi+}(t)}{\lambda_1 P_{\Xi+}(t) + \lambda_2 P_{\Xi-}(t)} + \lambda_2 P_{\Xi-}(t) \log \frac{P_{\Xi-}(t)}{\lambda_1 P_{\Xi+}(t) + \lambda_2 P_{\Xi-}(t)}\big) \qquad (t \in V^{\Xi+}),$$

which is called the association *score* of term t with query q. The estimation of $P_{\Xi+}(t)$ can be found in Section 3.7. The estimation of $P_{\Xi-}(t)$ can use the same way of estimating $P_D(t)$ found in Section 3.7. In addition, *a priori* probabilities can be given easily by setting $\lambda_1 = \frac{|\Xi^+|}{|\Xi|} > 0$ and $\lambda_2 = \frac{|\Xi^-|}{|\Xi|} > 0$.

Next, when $t \in V^{\Xi+} - V^{\Xi-} \subseteq V^{\Xi+}$, we have $P_{\Xi+}(t) > P_{\Xi-}(t) = 0$ and $\mathbf{ifd}_K(t) = \lambda_1 \mathbf{ifd}_{I_{1\Sigma}}(t) = \lambda_1 P_{\Xi+}(t) \log \frac{1}{\lambda_1} > 0$. Thus, term t contributes $\mathbf{ifd}_{I_{1\Sigma}}(t) > 0$ for supporting H_1, and contributes $\mathbf{ifd}_{I_{2\Sigma}}(t) = 0$ for supporting H_2. Because $\mathbf{ifd}_K(t) > 0$, the weighted algebraic sum is determined by its positive sub-item $\mathbf{ifd}_{I_{1\Sigma}}(t)$, and term t contributes quantity $\mathbf{ifd}_K(t)$ for supporting H_1. In other words, when terms appear in only relevant sample documents, the terms will provide information for supporting the relevant hypothesis. Such kind of terms should be considered as associated with the query.

Finally, when $t \in V^{\Xi+} \cap V^{\Xi-} \subseteq V^{\Xi+}$, we have $P_{\Xi+}(t) > 0$ and $P_{\Xi-}(t) > 0$. In this case, each of $\mathbf{ifd}_{I_{1\Sigma}}(t)$ and $\mathbf{ifd}_{I_{2\Sigma}}(t)$ can be positive or negative, and they are opposite in signs. Recall that $\mathbf{ifd}_{I_{1\Sigma}}(t) > 0$ if and only if $P_{\Xi+}(t) > P_{\Xi-}(t)$, and $\mathbf{ifd}_{I_{2\Sigma}}(t) < 0$ if and only if $P_{\Xi+}(t) > P_{\Xi-}(t)$. From the discussion given in Section 5.2, we can see that, just in the case that $P_{\Xi+}(t) > P_{\Xi-}(t)$ and

$$\mathbf{ifd}_K(t) = \lambda_1 \mathbf{ifd}_{I_{1\Sigma}}(t) + \lambda_2 \mathbf{ifd}_{I_{2\Sigma}}(t) = \lambda_1 \mathbf{ifd}_{I_{1\Sigma}}(t) - \lambda_2 |\mathbf{ifd}_{I_{2\Sigma}}(t)| > 0, \qquad (5.1)$$

term t is able to contribute quantity $\mathbf{ifd}_K(t)$ for supporting H_1. In other words when term t appears in both relevant and non-relevant sample documents, it would contain statistical information for supporting both the relevant and the non-relevant hypotheses. If $P_{\Xi+}(t) > P_{\Xi-}(t)$, then term t contains positive information $\mathbf{ifd}_{I_{1\Sigma}}(t)$ for supporting H_1, and also contains negative information $\mathbf{ifd}_{I_{2\Sigma}}(t)$ for supporting H_2. Further, if the finalized information (the

weighted algebraic sum of information that term t conveys) $\mathbf{ifd}_K(t)$ is positive, then the information of term t is dominated by its positive sub-item $\mathbf{ifd}_{I_{1\Sigma}}(t)$. The finalized information thus indicates that term t contains information $\mathbf{ifd}_K(t)$ for supporting H_1.

5.5.2 About Positive Scores

It should be emphasized here that, only one condition $score_{_K}(t) > 0$ cannot guarantee that term t is more or less positively associated with the query. It is true that a positive score implies that $\mathbf{ifd}_K(t)$ is positive as well (since $\mathcal{Q}(t) \geq 0$), however, this is not enough to infer that term t contains statistical information supporting H_1. This is because, when $t \in V^{\Xi^+} \cap V^{\Xi^-}$, it may have $\mathbf{ifd}_{I_{2\Sigma}}(t) > 0$ (and $\mathbf{ifd}_{I_{1\Sigma}}(t) < 0$), thus $\mathbf{ifd}_K(t) > 0$ indicates that the weighted algebraic sum is dominated by sub-item $\mathbf{ifd}_{I_{2\Sigma}}(t)$, and term t contributes quantity $\mathbf{ifd}_K(t)$ for supporting H_2. In this case, a higher positive score would express that term t is not statistically associated with query q. The 'prime culprit' that leads to $\mathbf{ifd}_{I_{2\Sigma}}(t) > 0$ is $P_{\Xi+}(t) < P_{\Xi-}(t)$.

Recall that, in the application of the directed divergence $I\big(P_{\Xi+} : P_D\big)$ and divergence $J\big(P_{\Xi+}, P_D\big)$, $score_{_I}(t) > 0$ and $score_{_J}(t) > 0$ implies $\mathbf{ifd}_I(t) > 0$ and $\mathbf{ifd}_J(t) > 0$, respectively, and that the terms with positive scores can immediately be inferred to more or less contain information associated with the query. This is because, in practice, we generally have $P_{\Xi+}(t) > P_D(t)$ for all terms $t \in V^{\Xi^+}$ (since the size of set Ξ^+ is very much smaller than the size of collection D). It would not, however, be true that $P_{\Xi+}(t) > P_{\Xi-}(t)$ for $t \in V^{\Xi^+}$ in our current application of the information radius since the size of set Ξ^+ may be quite close to the size of set Ξ^-. Therefore, in order to effectively carry out a feedback process, we must verify that conditions $P_{\Xi+}(t) > P_{\Xi-}(t)$ and $score_{_K}(t) > 0$ can simultaneously be satisfied for each selected term.

In fact, it can be seen easily that the satisfaction with two conditions $P_{\Xi+}(t) > P_{\Xi-}(t)$ and $score_{_K}(t) > 0$ is completely equivalent to the satisfaction with only one condition

$$\lambda_1 \mathbf{ifd}_{I_{1\Sigma}}(t) > \lambda_2 |\mathbf{ifd}_{I_{2\Sigma}}(t)| > 0 \tag{5.2}$$

since $\mathbf{ifd}_{I_{1\Sigma}}(t) > 0$ implies $P_{\Xi+}(t) > P_{\Xi-}(t)$, and inequality (5.2) implies $score_{_K}(t) > 0$.

Therefore, the expansion terms selected, from $t \in V^{\Xi^+}$, should be those that contribute most to the expected divergence $K\big(\lambda_1, \lambda_2; P_{\Xi+}, P_{\Xi-}\big)$. These expansion terms should be either those which obtain positive scores and satisfy $P_{\Xi+}(t) > P_{\Xi-}(t)$, or those which satisfy inequality (5.2). The higher the positive score, the more likely they are strongly associated with the query.

5.5.3 Relationship of Score Functions

Similar to function $score_{_J}(t)$, function $score_{_K}(t)$ can also be decomposed as

$$\begin{aligned} score_{_K}(t) &= \mathcal{Q}(t) \cdot \mathbf{ifd}_K(t) = \mathcal{Q}(t) \cdot \lambda_1 \mathbf{ifd}_{I_{1\Sigma}}(t) + \mathcal{Q}(t) \cdot \lambda_2 \mathbf{ifd}_{I_{2\Sigma}}(t) \\ &= \lambda_1 score_{_{I_{1\Sigma}}}(t) + \lambda_2 score_{_{I_{2\Sigma}}}(t) \qquad (t \in V^{\Xi^+}), \end{aligned}$$

which, obviously, can be positive or negative since $\mathbf{ifd}_K(t)$ can be positive or negative. From $\mathbf{ifd}_{I_{1\Sigma}}(t) \cdot \mathbf{ifd}_{I_{2\Sigma}}(t) \leq 0$ we have $score_{I_{1\Sigma}}(t) \cdot score_{I_{2\Sigma}}(t) \leq 0$.

We can see that function $score_{_K}(t)$ is the weighted algebraic sums of two opposite association scores $score_{I_{1\Sigma}}(t)$ and $score_{I_{2\Sigma}}(t)$: it offers not only the positive associations of

terms with the query, but also the negative associations inherent in terms when they also appear in non-relevant documents. In particular, when $P_{\Xi+}(t) > P_D(t)$ it has $\mathbf{ifd}_{I_{1\Sigma}}(t) > 0$ and $\mathbf{ifd}_{I_{2\Sigma}}(t) < 0$, and then $score_{I_{1\Sigma}}(t) \geq 0$ and $score_{I_{2\Sigma}}(t) \leq 0$. Further, if $score_K(t) > 0$. it implies $\lambda_1 score_{I_{1\Sigma}}(t) > \lambda_2 |score_{I_{2\Sigma}}(t)| \geq 0$, which expresses that term t is more or less statistically associated with the query to the extent given by $score_K(t)$.

5.5.4 A Symmetric Score Function

Particularly, when $|\Xi^+| = |\Xi^-|$ we have $\lambda_1 = \lambda_2 = \frac{1}{2}$. With Definition 5.4.1, we can construct the score function with form

$$score_K(t) = \mathcal{Q}(t) \cdot \mathbf{ifd}_K(t) = \mathcal{Q}(t)\frac{1}{2}\big(\mathbf{ifd}_{I_{1\Sigma}}(t) + \mathbf{ifd}_{I_{2\Sigma}}(t)\big).$$

By eliminating the scale factor $\frac{1}{2}$, we obtain the following equivalent score function:

$$score_K(t) = \mathcal{Q}(t)\big(P_{\Xi+}(t) + P_{\Xi-}(t) + P_{\Xi+}(t)\log\frac{P_{\Xi+}(t)}{P_{\Xi+}(t)+P_{\Xi-}(t)} + P_{\Xi-}(t)\log\frac{P_{\Xi-}(t)}{P_{\Xi+}(t)+P_{\Xi-}(t)}\big) \qquad (t \in V^{\Xi+}).$$

5.5.5 In Pseudo-Relevance Feedback Procedure

In an operational situation where no relevance information is available in advance, we would proceed as follows. Let statistical sample set Ξ be the top α ($= |\Xi|$) documents retrieved in the initial search. All documents $d \in \Xi$ are treated as (*pseudo*) relevant to the query q, and V^{Ξ} constitutes a source of candidate terms.

Let $\aleph$ be a set of documents ranked in the initial search, and $\aleph = \{d_{\beta+1}, d_{\beta+2}, ..., d_{\beta+\gamma}\}$, where $\beta > \alpha$ and $\gamma \geq 1$ are positive numbers, and subscripts $\beta+1, \beta+2, ..., \beta+\gamma$ are ranking numbers. The choice of α, β and γ, depending on a specific retrieval strategy, is not immaterial. For instance, we can take $\alpha = 10$, $\beta = 1000$ and $\gamma = 10$, if we have sufficient belief that documents $d \in \aleph = \{d_{1001}, d_{1002}, ..., d_{1010}\}$ are not very relevant to the query. Thus, we obtain an alternative sample set $\Im = \Xi \cup \aleph$ with $\Xi \cap \aleph = \emptyset$ (generally, $V^{\Xi} \cap V^{\aleph} \neq \emptyset$). Similar to relevance feedback, we thus may define two mutually exclusive and exhaustive events on the sample set by assuming $D_1 = \Xi$ being the set of the pseudo-relevant documents, and $D_2 = \aleph$ being the set of the pseudo-non-relevant documents.

Similarly, let $P_{\Xi}(t)$ be a term distribution over $(V, 2^V)$ defining set Ξ, where $P_{\Xi}(t) > 0$ when $t \in V^{\Xi}$ and $P_{\Xi}(t) = 0$ when $t \in V - V^{\Xi}$. Let $P_{\aleph}(t)$ be a term distribution over $(V, 2^V)$ defining set $\aleph$, where $P_{\aleph}(t) > 0$ when $t \in V^{\aleph}$ and $P_{\aleph}(t) = 0$ when $t \in V - V^{\aleph}$. The estimation of $P_{\Xi}(t)$ and $P_{\aleph}(t)$ can be derived in the same way as the estimation of $P_{\Xi+}(t)$ and $P_D(t)$, respectively, which can be found in Section 3.7.

If $\gamma \neq \alpha$, then we can use $K\big(\lambda_1, \lambda_2; P_{\Xi}, P_{\aleph}\big)$ to construct the score function as discussed in the relevance feedback situation above with $\lambda_1 = \frac{\alpha}{\alpha+\gamma}$ and $\lambda_2 = \frac{\gamma}{\alpha+\gamma}$.

If $\gamma = \alpha$, then we have $\lambda_1 = \lambda_2 = \frac{1}{2}$, and we can apply $K\big(P_{\Xi}, P_{\aleph}\big)$ to construct the score function as discussed in the relevance feedback situation above.

5.6 Summary

In this chapter, information radius is introduced as a device for constructing the discrimination measure of terms. A simple account of the concept of information radius is given. The application of the information radius to measurement of the amount of information contained in a term is interpreted.

¶ Unlike $I(P_R, P_{\bar{R}})$ and $J(P_R, P_{\bar{R}})$, in the applications of $K(\lambda_1, \lambda_2; P_R, P_{\bar{R}})$, *a priori* probability distribution $P_\lambda = \{\lambda_1, \lambda_2\}$ must be provided beforehand. The choice of P_λ depends on a specific model itself.

Unlike $I(P_R, P_{\bar{R}})$ and $J(P_R, P_{\bar{R}})$, two term distributions of $K(\lambda_1, \lambda_2; P_R, P_{\bar{R}})$ can be completely disjoint, i.e., $V^R \cap V^{\bar{R}} = \emptyset$ (recall that $I(P_R, P_{\bar{R}})$ requires $V^R \subseteq V^{\bar{R}}$ and $J(P_R, P_{\bar{R}})$ requires $V^R = V^{\bar{R}}$). In this case, $K(\lambda_1, \lambda_2; P_R, P_{\bar{R}})$ is reduced to the entropy of its *a priori* probability distribution.

¶ Like $I(P_R, P_{\bar{R}})$ and $J(P_R, P_{\bar{R}})$, if two term distributions of $K(\lambda_1, \lambda_2; P_R, P_{\bar{R}})$ overlap over some sub-domain $\Gamma \subseteq V^R \cap V^{\bar{R}}$, i.e., $P_R(t) = P_{\bar{R}}(t)$ for $t \in \Gamma$, then the expected divergence would drop sharply. Particularly, when $P_R(t) = P_{\bar{R}}(t)$ for $t \in V$, then $K(\lambda_1, \lambda_2; P_R, P_R) = 0$.

Like $I(P_R, P_{\bar{R}})$ and $J(P_R, P_{\bar{R}})$, $K(\lambda_1, \lambda_2; P_R, P_{\bar{R}})$ emphasizes the importance of those terms with variant probabilities within sets V^R and $V^{\bar{R}}$.

¶ The discrimination measure, and a symmetric discrimination measure (as a special case), based on the information radius is discussed. The concept of association of terms with the context of the query is then defined.

It is important to notice that divergence $K(P_R, P_{\bar{R}})$ cannot be reduced to $\Phi(t)$ by considering $K(P_R, P_{\bar{R}}) = 1 + \frac{1}{2}\Phi(t)$, and by eliminating coefficients 1 and $\frac{1}{2}$, when we want to use a symmetric discrimination measure (see Section 10.2 for a detailed discussion about $\Phi(t)$). This is because $\Phi(t)$ is non-positive and does not possess Criterion 2, so it cannot serve as a divergence measure.

¶ The method of constructing a score function and the simplification of the domain of the score function is addressed, and how the score function can be employed in both relevance and pseudo-relevance feedback is illustrated.

It should be especially pointed out that only one condition $score_K(t) > 0$ cannot infer that term t contains statistical information supporting H_1. As mentioned, when $t \in V^{\Xi+} \cap V^{\Xi-}$, it may have $\mathbf{ifd}_{I_{2\Sigma}}(t) > 0$ (in this case $\mathbf{ifd}_{I_{1\Sigma}}(t) < 0$), and positive quantity $\mathbf{ifd}_K(t)$ is dominated by its sub-item $\mathbf{ifd}_{I_{2\Sigma}}(t)$, and term t contributes $\mathbf{ifd}_K(t)$ for supporting H_2. In order to effectively select good terms, two conditions $P_{\Xi+}(t) > P_{\Xi-}(t)$ and $score_K(t) > 0$ should be simultaneously verified for each candidate term.

$K(\lambda_1, \lambda_2; P_R, P_{\bar{R}})$ is not symmetric in arguments $P_R(t)$ and $P_{\bar{R}}(t)$, neither in λ_1 and λ_2. Nevertheless, a symmetric divergence measure can be easily introduced by setting $\lambda_1 = \lambda_2 = \frac{1}{2}$. Also, there is no need to require the absolute continuity of $P_R(t)$ and $P_{\bar{R}}(t)$ in the applications of the information radius since $P_R(t) \ll P_\Sigma(t)$ and $P_{\bar{R}}(t) \ll P_\Sigma(t)$ always hold unconditionally (when $\lambda_1 \neq 0$ and $\lambda_2 \neq 0$). Such an outstanding property is not possessed by the divergence measures $I(P_R, P_{\bar{R}})$ and $J(P_R, P_{\bar{R}})$. In the next chapter, we will give another way to look at the information radius based on the concept of entropy increase.

Chapter 6

AQE Based on Jensen Difference

Our interest is to analyse the expected divergence of two term probability distributions derived from the relevant and non-relevant document sets, respectively. The analysis is expected to able to reveal some semantic relations between terms. The study addressed in this chapter is to view this issue in another way: the comparison of the diversities of terms in the sets of relevant and non-relevant documents, and the set of the combination of these two sets.

In Section 6.1, we discuss the diversity measure, and introduce three typical entropy functions used as the diversity measures. In Section 6.2, we discuss Jensen difference by means of the concavity of the diversity measure, and the Jensen difference for three entropy functions are derived. In Section 6.3, we discuss the measure of entropy increase, which is a special case of the Jensen difference, and the appropriateness of the application of three entropy functions is carefully investigated.

6.1 Diversity Measure $H\left(\lambda_1 P_{D_1} + \lambda_2 P_{D_2}\right)$

Diversity measure is a general concept, which can be applicable to observations belonging to any sample space when it satisfies some concavity properties.

6.1.1 Diversity Measure

A diversity measure can be conceived as a function from the space of term probability distributions into the real line. It reflects differences between terms within a set of documents.

Before a definition of the diversity measure can be given, it is necessary to understand the notion of *concavity*[1] of a function.

Then, use symbol H to indicate *Heterogeneity*, the concept of diversity measure can be defined as follows.

[1] A function $f(x)$ is said to be *concave* over some interval (a, b) if for arbitrary two points $x_1, x_2 \in (a, b)$ and two numbers λ_1, λ_2 satisfying $0 \leq \lambda_1, \lambda_2 \leq 1$ and $\lambda_1 + \lambda_2 = 1$,

$$f(\lambda_1 x_1 + \lambda_2 x_2) \geq \lambda_1 f(x_1) + \lambda_2 f(x_2).$$

Geometrically, a concave function always lies above any chord. Examples of concave functions include $\log x$ for $(0, +\infty)$ and $-x^2$ for $(-\infty, +\infty)$, etc.

Definition 6.1.1 A function $H(\cdot)$ mapping $\mathcal{P}_n$ into the real line $[0, +\infty)$ is said to be a measure of *diversity* if

(C_1) $H(P_{D_k}) \geq 0$ with equality if and only if $P_{D_k}(t)$ is *degenerate*[2],

(C_2) $H(\cdot)$ is a concave function, that is,

$$H(\lambda_1 P_{D_1} + \lambda_2 P_{D_2}) \geq \lambda_1 H(P_{D_1}) + \lambda_2 H(P_{D_2})$$

with equality if and only if $P_{D_1}(t) = P_{D_2}(t)$ $(0 \leq \lambda_1, \lambda_2 \leq 1$ and $\lambda_1 + \lambda_2 = 1)$,

which is referred to as the diversity within a document set D_k. The quantity of $H(P_{D_k})$ reflects the difference between terms within set D_k.

The condition C_1 is a natural one since a measure of diversity should preferably be non-negative and take the value zero only when all terms of a document set, in an extreme case, are 'identical'.

Consider two term distributions $P_{D_1}(t)$ and $P_{D_2}(t)$ and a composite distribution $P_{\Sigma}(t) = \lambda_1 P_{D_1}(t) + \lambda_2 P_{D_2}(t)$. Condition C_2 is motivated by considering that the amount of diversity in a mixture of document sets should not be smaller than the average of the diversities within the individual sets, that is, the diversity possibly increases by mixing the sets of documents. We may formulate this requirement by condition C_2. In other words, the condition C_2 is equivalent to saying that H is a concave function. The concavity of H reflects the intuitive requirement that two terms drawn from different sets are on the average more different than those coming from the same document set.

6.1.2 Entropy Functions as Diversity Measures

A variety of diversity measures have been introduced through the concept of *entropy* in information theory. In fact, an entropy function $H(\cdot)$ can be directly conceived of as a function defined on set $\mathcal{P}_n$, satisfying some postulates. Some of the postulates are that it is non-negative, attains the maximum for the *uniform*[3] distribution $P_I(t)$, and has the minimum when the term distribution is degenerate. In some sense, an entropy function is an index of similarity of an arbitrary distribution $P_{D_k}(t)$ with the uniform distribution $P_I(t)$, and hence a measure of diversity [131].

Mathai & Rathie [126] and Aczel & Daroczy [1] consider three general forms for entropy functions:

$$H(P_{D_k}) = -\frac{\sum_{t \in V} P_{D_k}^{\beta_t}(t) \log P_{D_k}(t)}{\sum_{t \in V} P_{D_k}^{\beta_t}(t)},$$

$$H(P_{D_k}) = \frac{1}{1-\alpha} \log \left(\frac{\sum_{t \in V} P_{D_k}^{\alpha+\beta_t-1}(t)}{\sum_{t \in V} P_{D_k}^{\beta_t}(t)} \right) \qquad (\alpha > 0,\ \alpha \neq 1),$$

$$H(P_{D_k}) = \frac{1}{2^{1-\alpha}-1} \left(\frac{\sum_{t \in V} P_{D_k}^{\alpha+\beta_t-1}(t)}{\sum_{t \in V} P_{D_k}^{\beta_t}(t)} - 1 \right) \qquad (\alpha > 0,\ \alpha \neq 1).$$

[2] A probability distribution $P = \{p_1, p_2, ..., p_n\}$ is said to be *degenerate* if $p_k = 1$ $(1 \leq k \leq n)$ and $p_i = 0$ $(1 \leq i \leq n,\ i \neq k)$.

[3] A probability distribution $P = \{p_1, p_2, ..., p_n\}$ is said to be *uniform*, denoted P_I, if $p_i = \frac{1}{|V|}$ $(1 \leq i \leq n)$.

When $\beta_t = 1$ for all terms $t \in V$ we obtain the familiar expressions introduced by Shannon [176], Rényi [145], Havrda & Charvát [80]:

$$H_{Sh}(P_{D_k}) = -\sum_{t \in V} P_{D_k}(t) \log P_{D_k}(t),$$

$$H_{Re}(P_{D_k}) = \frac{1}{1-\alpha} \log\Big(\sum_{t \in V} P_{D_k}^{\alpha}(t)\Big) \qquad (\alpha > 0,\ \alpha \neq 1),$$

$$H_{HC}(P_{D_k}) = \frac{1}{1-2^{1-\alpha}}\Big(1 - \sum_{t \in V} P_{D_k}^{\alpha}(t)\Big) \qquad (\alpha > 0,\ \alpha \neq 1).$$

All these measures are non-negative and take value zero when and only when $P_{D_k}(t)$ is degenerate. They all attain the maximum when $P_{D_k}(t) = \frac{1}{|V|}$ for every term $t \in V$. Thus they satisfy condition C_1.

It can be verified that H_{Sh} satisfies the concavity condition C_2, H_{Re} satisfies C_2 for only $0 < \alpha < 1$, while H_{HC} satisfies the concavity condition C_2 for any $\alpha > 0$ ($\alpha \neq 1$) [144].

In addition, for $\alpha = 1$, $H_{Re}(P_{D_k})$ and $H_{HC}(P_{D_k})$ are defined in the limiting sense:

$$\lim_{\alpha \to 1} H_{Re}(P_{D_k}) = \lim_{\alpha \to 1} H_{HC}(P_{D_k}) = H_{Sh}(P_{D_k}).$$

For more properties of these functions see [1, 126].

Nayak studied the relationships between these three entropy functions for the different values α. We will not further discuss this topic in this thesis. The interested reader is referred to [131].

6.2 Jensen Difference

If we have a mixture of several document sets, it would be of interest to know how much of the diversity in the mixture set is due to diversity within the sets and how much due to between the sets. Rao [143, 144] refers to this problem as decomposition of diversity.

The concavity of the diversity measure provides the decomposition of the total diversity in a composite distribution as defined in the following.

Definition 6.2.1 Let $P_{D_k}(t)$ define the document set D_k with *a priori* probability λ_k ($k = 1, 2, ..., r$). Define the *decomposition* [144] by

$$H\Big(\sum_{k=1}^{r} \lambda_k P_{D_k}\Big) = \sum_{k=1}^{r} \lambda_k H(P_{D_k}) + J_H(\{\lambda_k\}; \{P_{D_k}\}),$$

in which, the first item on the right side of equation above is the average diversity within the term distributions, and the second item

$$J_H(\{\lambda_k\}; \{P_{D_k}\}) = H\Big(\sum_{k=1}^{r} \lambda_k P_{D_k}\Big) - \sum_{k=1}^{r} \lambda_k H(P_{D_k})$$

is referred to as the *Jensen difference* with respect to entropy H, which is non-negative and vanishes when $P_{D_1}(t) = \cdots = P_{D_r}(t)$. Jensen difference provides a measure of overall differences among the term distributions $P_{D_k}(t)$ ($k = 1, 2, \cdots, r$).

Thus, for the three entropy functions given in the last section, from the derivation given in Section 10.3, we can immediately obtain the following results.

For Shannon's entropy, the Jensen difference can be written as

$$J_{H_{Sh}}(\{\lambda_k\};\{P_{D_k}\}) = H_{Sh}\Big(\sum_{k=1}^{r}\lambda_k P_{D_k}\Big) - \sum_{k=1}^{r}\lambda_k H_{Sh}(P_{D_k})$$
$$= \sum_{t\in V}\sum_{k=1}^{r}\lambda_k P_{D_k}(t)\log\frac{P_{D_k}(t)}{\sum_{k=1}^{r}\lambda_k P_{D_k}(t)},$$

which is the *information radius* defined by Sibson [177].

For Rényi's entropy, when $0<\alpha<1$, the Jensen difference can be written as

$$J_{H_{Re}}(\{\lambda_k\};\{P_{D_k}\}) = H_{Re}\Big(\sum_{k=1}^{r}\lambda_k P_{D_k}\Big) - \sum_{k=1}^{r}\lambda_k H_{Re}(P_{D_k})$$
$$= \frac{1}{1-\alpha}\log\frac{\sum_{t\in V}\big(\sum_{k=1}^{r}\lambda_k P_{D_k}(t)\big)^{\alpha}}{\prod_{k=1}^{r}\big(\sum_{t\in V}P_{D_k}^{\alpha}(t)\big)^{\lambda_k}}.$$

For the entropy of Havrda & Charvát, when $\alpha>0$ ($\alpha\neq 1$), the Jensen difference can be written as

$$J_{H_{HC}}(\{\lambda_k\};\{P_{D_k}\}) = H_{HC}\Big(\sum_{k=1}^{r}\lambda_k P_{D_k}\Big) - \sum_{k=1}^{r}\lambda_k H_{HC}(P_{D_k})$$
$$= \frac{1}{1-2^{1-\alpha}}\sum_{t\in V}\Big(\big(\sum_{k=1}^{r}\lambda_k P_{D_k}^{\alpha}(t)\big) - \big(\sum_{k=1}^{r}\lambda_k P_{D_k}(t)\big)^{\alpha}\Big).$$

6.3 Appropriateness of Applications

Let us now consider a particular situation where $r=2$. Assume that $P_R(t)$ and $P_{\bar{R}}(t)$ are discrete probability distributions over $(V,2^V)$, and derived from sets R and $\bar{R}$, respectively. Denote the corresponding Jensen difference by

$$J_H(\lambda_1,\lambda_2;P_R,P_{\bar{R}}) = H(\lambda_1 P_R+\lambda_2 P_{\bar{R}}) - \lambda_1 H(P_R) - \lambda_1 H(P_{\bar{R}}),$$

which is referred as to the measure of *entropy increase* of term t.

From the condition C_2, the entropy increase $J_H(\lambda_1,\lambda_2;P_R,P_{\bar{R}})$ is positive if distributions $P_R(t)$ and $P_{\bar{R}}(t)$ are different, equals to zero when $P_R(t)=P_{\bar{R}}(t)$, and hence may be considered a direct measure of overall differences between $P_R(t)$ and $P_{\bar{R}}(t)$. In other words, $J_H(\lambda_1,\lambda_2;P_R,P_{\bar{R}})$ provides the excess variability, representing the amount of difference between the sets R and $\bar{R}$. Then we obtain a measure of differences between $P_R(t)$ and $P_{\bar{R}}(t)$ induced by the diversity $H(\cdot)$, which is not necessary symmetric in $P_R(t)$ and $P_{\bar{R}}(t)$, respectively.

In practice, we may need a symmetric entropy increase measure. For this reason, take $\lambda_1=\lambda_2=\frac{1}{2}$, and denote the corresponding entropy increase measure by

$$J_H(P_R,P_{\bar{R}}) = H(\tfrac{1}{2}P_R+\tfrac{1}{2}P_{\bar{R}}) - \tfrac{1}{2}H(P_R) - \tfrac{1}{2}H(P_{\bar{R}}).$$

which is symmetric in $P_R(t)$ and $P_{\bar{R}}(t)$, and thus can be viewed as 'distance' between $P_R(t)$ and $P_{\bar{R}}(t)$ induced by the diversity H.

In applications to IR, $J_H(\lambda_1, \lambda_2; P_R, P_{\bar{R}})$ should be chosen to effectively reflect some intrinsic difference between terms related to a specific classification procedure.

6.3.1 Entropy Function H_{Sh}

Shannon's entropy has many properties that agree with the intuitive notion of what a measure of information should be [38]. As mentioned in Section 3.1, measure $i(t) = -\log P(t)$ can be interpreted as the uncertainty concerning the occurrence of term t before an experiment is performed. Then, Shannon's entropy

$$H_{Sh}(P) = \sum_{t \in V} P(t) i(t) = -\sum_{t \in V} P(t) \log P(t)$$

can be thought of as a measure of the expected uncertainty concerning the occurrence of term t (as a random variable), that is, a measure of the amount of information required in the expectation to describe the random variable t.

Notice that we adopt here the convention $0 \log 0 = 0$ since it is rather natural that adding items with zero probability does not affect the degree of uncertainty, i.e., does not change the Shannon entropy. It is shown that the entropy, $H_{Sh}(P)$, of a discrete random variable may be finite, even when the random variable takes on a denumerable number of values.

Particularly, the concavity of $H_{Sh}(P)$ is typically useful in IR applications: it provides a natural measure of divergence between distributions $P_R(t)$ and $P_{\bar{R}}(t)$. For Shannon's entropy, we have the measure of increase in entropy,

$$\begin{aligned} J_{H_{Sh}}(\lambda_1, \lambda_2; P_R, P_{\bar{R}}) = \sum_{t \in V} \Big(& \lambda_1 P_R(t) \log \frac{P_R(t)}{\lambda_1 P_R(t) + \lambda_2 P_{\bar{R}}(t)} \\ & + \lambda_2 P_{\bar{R}}(t) \log \frac{P_{\bar{R}}(t)}{\lambda_1 P_R(t) + \lambda_2 P_{\bar{R}}(t)} \Big), \end{aligned}$$

which is the divergence measure $K(\lambda_1, \lambda_2; P_R, P_{\bar{R}})$, and the expression of each of its items

$$\Gamma_{H_{Sh}}(t) = \lambda_1 P_R(t) \log \frac{P_R(t)}{\lambda_1 P_R(t) + \lambda_2 P_{\bar{R}}(t)} + \lambda_2 P_{\bar{R}}(t) \log \frac{P_{\bar{R}}(t)}{\lambda_1 P_R(t) + \lambda_2 P_{\bar{R}}(t)}$$

is the discrimination measure $\mathbf{ifd}_K(t)$. Thus, another way of looking at the information radius and its individual items is by the entropy increase.

We have given detailed discussions on $K(\lambda_1, \lambda_2; P_R, P_{\bar{R}})$ and $\mathbf{ifd}_K(t)$, and on how we can apply them to establish the association concept and construct the association score function, in the last chapter.

6.3.2 Entropy Function H_{Re}

For Rényi's entropy, when $0 < \alpha < 1$, we have the entropy increase measure

$$\begin{aligned} J_{H_{Re}}(\lambda_1, \lambda_2; P_R, P_{\bar{R}}) &= \frac{1}{1-\alpha} \log \frac{\sum_{t \in V} \left(\lambda_1 P_R(t) + \lambda_2 P_{\bar{R}}(t)\right)^\alpha}{\left(\sum_{t \in V} P_R^\alpha(t)\right)^{\lambda_1} \left(\sum_{t \in V} P_{\bar{R}}^\alpha(t)\right)^{\lambda_2}} \\ &= \log \Big(\sum_{t \in V} \frac{\left(\lambda_1 P_R(t) + \lambda_2 P_{\bar{R}}(t)\right)^\alpha}{\left(\sum_{t \in V} P_R^\alpha(t)\right)^{\lambda_1} \left(\sum_{t \in V} P_{\bar{R}}^\alpha(t)\right)^{\lambda_2}} \Big)^{\frac{1}{1-\alpha}}. \end{aligned}$$

Notice that function log is monotonically increasing of its argument:

$$\Big(\sum_{t\in V}\frac{\big(\lambda_1 P_R(t)+\lambda_2 P_{\bar{R}}(t)\big)^{\alpha}}{\big(\sum_{t\in V}P_R^{\alpha}(t)\big)^{\lambda_1}\big(\sum_{t\in V}P_{\bar{R}}^{\alpha}(t)\big)^{\lambda_2}}\Big)^{\frac{1}{1-\alpha}}.$$

Notice also that, it has $a=\frac{1}{1-\alpha}>1$ when $0<\alpha<1$. Thus, function χ^a is monotonically increasing since its argument

$$\chi=\sum_{t\in V}\frac{\big(\lambda_1 P_R(t)+\lambda_2 P_{\bar{R}}(t)\big)^{\alpha}}{\big(\sum_{t\in V}P_R^{\alpha}(t)\big)^{\lambda_1}\big(\sum_{t\in V}P_{\bar{R}}^{\alpha}(t)\big)^{\lambda_2}}\geq 0.$$

Therefore, the order of the contributions made by individual terms to $J_{H_{Re}}(\lambda_1,\lambda_2;P_R,P_{\bar{R}})$ is the same as that of argument χ.

Thus, considering the contributions made by terms to measure $J_{H_{Re}}$ can be reduced to considering the contributions made by terms of argument χ, in which, each item of χ can be written by

$$\Gamma_{H_{Re}}(t)=\frac{\big(\lambda_1 P_R(t)+\lambda_2 P_{\bar{R}}(t)\big)^{\alpha}}{\big(\sum_{t\in V}P_R^{\alpha}(t)\big)^{\lambda_1}\big(\sum_{t\in V}P_{\bar{R}}^{\alpha}(t)\big)^{\lambda_2}}.$$

For some term $t\in V^R$ satisfying $P_R(t)=P_{\bar{R}}(t)\neq 0$,

$$\frac{\big(\lambda_1 P_R(t)+\lambda_2 P_R(t)\big)^{\alpha}}{\big(\sum_{t\in V}P_R^{\alpha}(t)\big)^{\lambda_1}\big(\sum_{t\in V}P_R^{\alpha}(t)\big)^{\lambda_2}}=\frac{P_R^{\alpha}(t)(\lambda_1+\lambda_2)^{\alpha}}{\big(\sum_{t\in V}P_R^{\alpha}(t)\big)^{\lambda_1+\lambda_2}}=\frac{P_R^{\alpha}(t)}{\sum_{t\in V}P_R^{\alpha}(t)}\neq 0,$$

that is, $J_{H_{Re}}(\lambda_1,\lambda_2;P_R,P_{\bar{R}})$ does not possess Criterion 2. Consequently, the expression, $\Gamma_{H_{Re}}(t)$, of the individual items of $J_{H_{Re}}(\lambda_1,\lambda_2;P_R,P_{\bar{R}})$ should not be an appropriate discrimination measure of terms.

6.3.3 Entropy Function H_{HC}

For the entropy of Havrda & Charvát, when $\alpha>0$ ($\alpha\neq 1$), we have the entropy increase measure

$$J_{H_{HC}}(\lambda_1,\lambda_2;P_R,P_{\bar{R}})=\frac{1}{1-2^{1-\alpha}}\sum_{t\in V}\Big(\big(\lambda_1 P_R^{\alpha}(t)+\lambda_2 P_{\bar{R}}^{\alpha}(t)\big)-\big(\lambda_1 P_R(t)+\lambda_2 P_{\bar{R}}(t)\big)^{\alpha}\Big).$$

Each of its items is

$$\frac{1}{1-2^{1-\alpha}}\Big(\big(\lambda_1 P_R^{\alpha}(t)+\lambda_2 P_{\bar{R}}^{\alpha}(t)\big)-\big(\lambda_1 P_R(t)+\lambda_2 P_{\bar{R}}(t)\big)^{\alpha}\Big).$$

It is easily seen that $J_{H_{HC}}(\lambda_1,\lambda_2;P_R,P_{\bar{R}})$ possesses Criterion 2. In fact, for some term $t\in V^R$ satisfying $P_R(t)=P_{\bar{R}}(t)\neq 0$, we have

$$\big(\lambda_1 P_R^{\alpha}(t)+\lambda_2 P_R^{\alpha}(t)\big)-\big(\lambda_1 P_R(t)+\lambda_2 P_R(t)\big)^{\alpha}=P_R^{\alpha}(t)-P_R^{\alpha}(t)=0.$$

However, its items would not be an appropriate discrimination measure of terms. In order to explain this point, let us consider a simple case when $\alpha=2$:

$$J_{H_{HC}}(\lambda_1,\lambda_2;P_R,P_{\bar{R}})=\frac{1}{1-2^{1-2}}\sum_{t\in V}\Big(\big(\lambda_1 P_R^{2}(t)+\lambda_2 P_{\bar{R}}^{2}(t)\big)-\big(\lambda_1 P_R(t)+\lambda_2 P_{\bar{R}}(t)\big)^{2}\Big)$$

$$= 2\sum_{t\in V}\Big(\lambda_1 P_R^2(t) + \lambda_2 P_{\bar{R}}^2(t) - \lambda_1^2 P_R^2(t) - \lambda_2^2 P_{\bar{R}}^2(t) - 2\lambda_1\lambda_2 P_R(t)P_{\bar{R}}(t)\Big)$$

$$= 2\sum_{t\in V}\Big(\lambda_1 P_R^2(t)(1-\lambda_1) + \lambda_2 P_{\bar{R}}^2(t)(1-\lambda_2) - 2\lambda_1\lambda_2 P_R(t)P_{\bar{R}}(t)\Big)$$

$$= 2\sum_{t\in V}\Big(\lambda_1 P_R^2(t)\lambda_2 + \lambda_2 P_{\bar{R}}^2(t)\lambda_1 - 2\lambda_1\lambda_2 P_R(t)P_{\bar{R}}(t)\Big) = 2\lambda_1\lambda_2\sum_{t\in V}\big(P_R(t) - P_{\bar{R}}(t)\big)^2.$$

It is reduced to the *Euclidean distance* apart from a scale factor $2\lambda_1\lambda_2$. Each of its items is

$$\Gamma_{H_{HC}}(t) = 2\lambda_1\lambda_2\big(P_R(t) - P_{\bar{R}}(t)\big)^2 = 2\lambda_1\lambda_2\big|P_R(t) - P_{\bar{R}}(t)\big|^2.$$

In order to discuss the appropriateness of $\Gamma_{H_{HC}}(t)$ as a discrimination measure, let us recall our analysis given in the previous chapters (see the summary of the signs of the discrimination measures $\mathbf{ifd}_I(t), \mathbf{ifd}_J(t), \mathbf{ifd}_K(t)$ given in Table 6.3.1), we knew that,

- If $P_R(t) > P_{\bar{R}}(t)$, term t contributes quantity $\mathbf{ifd}_I(t) > 0$ for supporting H_1.
  If $P_R(t) < P_{\bar{R}}(t)$, term t contributes quantity $\mathbf{ifd}_I(t) < 0$ for supporting H_1.
- If $P_R(t) > P_{\bar{R}}(t)$, term t contributes quantity $\mathbf{ifd}_J(t) > 0$ for supporting H_1.
  If $P_R(t) < P_{\bar{R}}(t)$, term t contributes quantity $\mathbf{ifd}_J(t) > 0$ for supporting H_2.
- If $P_R(t) > P_{\bar{R}}(t)$ and $\mathbf{ifd}_K(t) > 0$, term t contributes $\mathbf{ifd}_K(t)$ for supporting H_1.
  If $P_R(t) < P_{\bar{R}}(t)$ and $\mathbf{ifd}_K(t) > 0$, term t contributes $\mathbf{ifd}_K(t)$ for supporting H_2.

Table 6.3.1 The signs of the discrimination measures

	$P_R(t) > P_{\bar{R}}(t)$	$P_R(t) < P_{\bar{R}}(t)$	$P_R(t) = P_{\bar{R}}(t)$
$\mathbf{ifd}_I(t)$	$\mathbf{ifd}_I(t) > 0$ Ⓢ H_1	$\mathbf{ifd}_I(t) < 0$ Ⓢ H_1	$\mathbf{ifd}_I(t) = 0$
$\mathbf{ifd}_J(t)$	$\mathbf{ifd}_{I_{12}}(t) > 0$ Ⓢ H_1 $\mathbf{ifd}_{I_{21}}(t) < 0$ Ⓢ H_2	$\mathbf{ifd}_{I_{12}}(t) < 0$ Ⓢ H_1 $\mathbf{ifd}_{I_{21}}(t) > 0$ Ⓢ H_2	$\mathbf{ifd}_{I_{12}}(t) = 0$ $\mathbf{ifd}_{I_{21}}(t) = 0$
	$\mathbf{ifd}_J(t) > 0$ Ⓢ H_1	$\mathbf{ifd}_J(t) > 0$ Ⓢ H_2	$\mathbf{ifd}_J(t) = 0$
$\mathbf{ifd}_K(t)$	$\mathbf{ifd}_{I_{1\Sigma}}(t) > 0$ Ⓢ H_1 $\mathbf{ifd}_{I_{2\Sigma}}(t) < 0$ Ⓢ H_2	$\mathbf{ifd}_{I_{1\Sigma}}(t) < 0$ Ⓢ H_1 $\mathbf{ifd}_{I_{2\Sigma}}(t) > 0$ Ⓢ H_2	$\mathbf{ifd}_{I_{1\Sigma}}(t) = 0$ $\mathbf{ifd}_{I_{2\Sigma}}(t) = 0$
	$\mathbf{ifd}_K(t) > 0$ Ⓢ H_1 $\mathbf{ifd}_K(t) < 0$ Ⓢ H_2	$\mathbf{ifd}_K(t) > 0$ Ⓢ H_2 $\mathbf{ifd}_K(t) < 0$ Ⓢ H_1	$\mathbf{ifd}_K(t) = 0$

Symbol '$\mathbf{ifd}(t)$ Ⓢ H_k' expresses that the discrimination measure $\mathbf{ifd}(t)$ supports hypothesis H_k; and H_1 and H_2 are two hypotheses that term t are drawn from R and $\bar{R}$, respectively.

We can easily see that whether term t supports the relevant hypothesis depends mainly on the relationship between $P_R(t)$ and $P_{\bar{R}}(t)$, rather than on the mathematical sign of the discrimination measures. These results clearly tell us that, as long as $P_R(t) > P_{\bar{R}}(t)$, term t is deemed to contain statistical information supporting H_1. Conversely, so long as $P_R(t) < P_{\bar{R}}(t)$, term t supports H_2. Thus, $P_R(t) - P_{\bar{R}}(t)$ is in fact the simplest discrimination measure.

One might think that, in a pseudo-relevance feedback procedure, $P_\Xi(t) \geq P_D(t)$ holds for all terms $t \in V^\Xi$, and thus, function $2\lambda_1\lambda_2\big|P_\Xi(t) - P_D(t)\big|^2$ can be consistent with function $P_\Xi(t) - P_D(t)$ in the sense of reflecting the relationship between $P_R(t)$ and $P_{\bar{R}}(t)$. However, in this case, it has $\lambda_1 = \frac{|\Xi|}{|D|} \to 0$ as $|D| = N \to +\infty$ and $\lambda_2 = \frac{|D|}{|D|} = 1$, which implies that the

entropy increase measures discussed in this chapter would not be suitable for the situation of pseudo-relevance feedback.

On the other hand, in a relevance feedback procedure, we use Ξ^+ instead of R, and take $\Xi^- = \Xi - \Xi^+$ ($\not\approx \bar{R} \approx D$). In this case, one can very easily set $\lambda_1 = \frac{|\Xi^+|}{|\Xi|}$ and $\lambda_2 = \frac{|\Xi^-|}{|\Xi|}$. However, in practice, we cannot assume that condition $P_{\Xi^+}(t) \geq P_{\Xi^-}(t)$ holds for all terms $t \in V^{\Xi^+}$. Thus, function $2\lambda_1\lambda_2\left|P_{\Xi^+}(t) - P_{\Xi^-}(t)\right|^2$ cannot give the relationship between $P_R(t)$ and $P_{\bar{R}}(t)$. Consequently, expression $\Gamma_{H_{HC}}(t)$ is not an appropriate discrimination measure of terms.

6.4 Summary

This chapter attempts to study the application of the concept of entropy, or entropy increase, to IR, by introducing the more general concepts of diversity and Jensen difference. Three typical entropy functions are discussed, and the appropriateness of applying them as a divergence measure is investigated.

¶ The concavity of $H_{Sh}(P)$ provides a natural measure of divergence between distributions $P_R(t)$ and $P_{\bar{R}}(t)$. It turns out that the entropy increase measure $J_{H_{Sh}}(\lambda_1, \lambda_2; P_R, P_{\bar{R}}) = K(\lambda_1, \lambda_2; P_R, P_{\bar{R}})$, and each of its items $\Gamma_{H_{Sh}}(t) = \mathbf{ifd}_K(t)$. Thus, with the entropy increase, we have another way to interpret the concept of the information radius and its individual items.

¶ For entropy $H_{Re}(P)$, when $0 < \alpha < 1$, for term $t \in V^R$ satisfying $P_R(t) = P_{\bar{R}}(t) \neq 0$, it has $\Gamma_{H_{Re}}(t) \neq 0$, that is, $J_{H_{Re}}(\lambda_1, \lambda_2; P_R, P_{\bar{R}})$ does not possess Criterion 2. Thus, the entropy increase measure $J_{H_{Re}}(\lambda_1, \lambda_2; P_R, P_{\bar{R}})$ should not be an appropriate divergence measure of term distributions.

¶ It can be seen that, when $\alpha > 0$ ($\alpha \neq 1$), $J_{H_{HC}}(\lambda_1, \lambda_2; P_R, P_{\bar{R}})$ possesses Criterion 2. However, when $\alpha = 2$, it is reduced to the *Euclidean distance* apart from a scale factor $2\lambda_1\lambda_2$, and each of its items is $\Gamma_{H_{HC}}(t) = 2\lambda_1\lambda_2\left|P_R(t) - P_{\bar{R}}(t)\right|^2$, which cannot give the relationship between $P_R(t)$ and $P_{\bar{R}}(t)$. Thus, the entropy increase measure $J_{H_{HC}}(\lambda_1, \lambda_2; P_R, P_{\bar{R}})$ might not be an appropriate divergence measure of term distributions either.

Chapter 7

AQE Based on Expected Mutual Information

This chapter is devoted to the discussion of an interesting subject: discrimination on the mutual information, or dependence, of terms. The formalism of the discrimination measures is based on the concept of expected mutual information [67, 106].

The discrimination on mutual information of terms was formally introduced to IR theory, as a device for identifying good terms, by Van Rijsbergen [206, 207]. This chapter attempts a further investigation into the issue based on the study initiated there. We will see that the formal method proposed in this chapter not only covers the method EMIM as a special case, but also suggests a general form of the definition and estimation of mutual information within a more general probabilistic framework.

In Section 7.1, after distinguishing term state distribution from term distribution, we intend to give a formal interpretation of the notion of amount of mutual information contained in a given term pair. In Section 7.2, we focus on the mathematical study of estimating term state distributions. Three particular methods are considered, and then a general framework for the estimation is established. In Section 7.3, we make an in-depth investigation into dependence discrimination measures and reveal some important relationships between them, which underpin the method proposed in this chapter. In Section 7.4, we give an insight into the concept of dependence of terms. Section 7.5, we devote to introducing the concepts of mutual association in the sense of the mutual information of terms. Three basic concepts: term-based association, set-based association and query-based association, are discussed. In Section 7.6, two score functions are proposed, and their relationship is analysed. In Section 7.7, we address extensions of our methods to other information entities. Also, the reader is referred to Section 10.6 where three examples are given which elaborate on all computations encountered in this chapter.

7.1 Information Gain $I(\delta_i, \delta_j)$

The objective of this section is to apply the idea of information theory to IR theory by interpretations of the notion of amount of mutual information contained in a given term pair. The formal method proposed in this chapter will be based on these interpretations.

Before entering into a formal discussion on the mutual information of terms, let us first clarify the difference between the notions of a term state distribution and a term distribution.

7.1.1 Term State Distribution and Term Distribution

To speak of the mutual information of terms, we must regard the term state distribution as defined on a different probability space from the term distribution.

A term is usually thought of having its state values *present* or *absent* in some document. Thus, for an arbitrary term t, we need to introduce a variable δ taking values from set $\Omega = \{1, 0\}$, where $\delta = 1$ expresses that term t is present and $\delta = 0$ expresses that term t is absent. That is, if we denote $t^1 = t$ and $t^0 = \bar{t}$, then, it has $t^\delta = t$, $\bar{t}$ when $\delta = 1, 0$, respectively. We call $\Omega = \{1, 0\}$ a *state value space*, and each element in Ω a *state value*, of the term.

Similarly, for an arbitrary term pair (t_i, t_j), we introduce a variable pair (δ_i, δ_j) taking values from set $\Omega \times \Omega = \{(1,1), (1,0), (0,1), (0,0)\}$. We call $\Omega \times \Omega$ a state value space, and each element in $\Omega \times \Omega$ a state value, of the term pair.

In information retrieval, the notion of term state distribution should be carefully distinguished from the notion of term distribution. For a given document d and term $t \in V^d$, its state distribution, denoted by $P_d(\delta) = P(t^\delta | d)$, should be over the state value space $\Omega = \{1, 0\}$. Whereas the term distribution, denoted by $p_d(t) = p(t|d)$, should be derived from document d, and over term set V^d. Generally, the term state distribution can easily be derived in terms of the term distribution.

More precisely, assume that, for each document $d \in D$, set V^d satisfies $2 \leq |V^d| \leq n$ (i.e., each document has at least two distinct terms). Also, assume that frequencies $f_d(t)$ for all terms $t \in V^d$ have been obtained. Based on the statistical data within document d, each document can be characterized by

$$p_d(t) = \frac{f_d(t)}{||d||} = \frac{f_d(t)}{\sum_{t' \in V^d} f_d(t')} \qquad (t \in V^d), \tag{7.1}$$

which is the term probability distribution over $(V^d, 2^{V^d})$. Obviously, under assumption $|V^d| \geq 2$, we have $0 < p_d(t) < 1$ for every $t \in V^d$. (Notice that $p_d(t)$ can also be defined over $(V, 2^V)$, a detailed discussion about this is given in Section 10.5).

Based on the term distribution, for a given term $t \in V^d$, the probabilities of the different state values concerning document d can be written down

$$P_d(\delta = 1) = p_d(t) \quad \text{and} \quad P_d(\delta = 0) = 1 - p_d(t), \tag{7.2}$$

which is over $\Omega = \{1, 0\}$. Clearly, $0 < P_d(\delta) < 1$ for $\delta = 1, 0$.

It can easily be seen that different terms in a given document are likely to obtain different term state distributions depending on the term distribution which is uniquely determined by the statistical data within the document.

Example 7.1.1 Suppose that we are given a document $d = \{t_1, t_2, t_2, t_3, t_4\}$. From which we have $V^d = \{t_1, t_2, t_3, t_4\}$, and the term distribution:

$$p_d(t_1) = \frac{1}{6}, \quad p_d(t_2) = \frac{3}{6}, \quad p_d(t_3) = \frac{1}{6}, \quad p_d(t_4) = \frac{1}{6}.$$

Thus, the term state distributions for individual terms $t \in V^d$ are:

$$P_d(\delta_1 = 1) = \frac{1}{6} \quad \text{and} \quad P_d(\delta_1 = 0) = 1 - \frac{1}{6} = \frac{5}{6};$$
$$P_d(\delta_2 = 1) = \frac{3}{6} \quad \text{and} \quad P_d(\delta_2 = 0) = 1 - \frac{3}{6} = \frac{3}{6};$$

and so on. ♠

For a term pair (t_i, t_j), the estimation of the term (joint) state distribution $P_d(\delta_i, \delta_j)$ is a more complicated task, which we shall specially discuss in the next section.

It should especially be pointed out that the term taking a certain state value δ should be looked upon as complex because many other term state values may be dependent on the state value. This is a central issue on which we shall concentrate in this chapter.

7.1.2 Mutual Information Contained in a Term Pair

The concept of expected mutual information is a very familiar one. More formal discussions about it can be found in [106]. An account for discrete variables is given as follows.

Let H_1 and H_2 be two opposite hypotheses related to a term pair (t_i, t_j). Similar to the discussion on the information gain $i(H_1 : H_2|t)$ given in Section 3.2, the logarithm of the likelihood ratio,

$$i(H_1 : H_2|(t_i, t_j)) = \log \frac{P((t_i, t_j)|H_1)}{P((t_i, t_j)|H_2)} = \log \frac{P(H_1|(t_i, t_j))}{P(H_2|(t_i, t_j))} - \log \frac{P(H_1)}{P(H_2)}, \tag{7.3}$$

measures the amount of information contained in term pair (t_i, t_j) for discrimination in favour of H_1 against H_2.

Particularly, if we assume that hypothesis H_1: terms t_i and t_j are dependent with a joint probability distribution $P(\delta_i, \delta_j)$ over $\Omega \times \Omega$, and that hypothesis H_2: terms t_i and t_j are independent with the product of marginal distributions $P(\delta_i)$ and $P(\delta_j)$ both over Ω. Then, when H_1 is true, measure

$$i(H_1 : H_2|(\delta_i, \delta_j)) = \log \frac{P((\delta_i, \delta_j)|H_1)}{P((\delta_i, \delta_j)|H_2)} = \log \frac{P(\delta_i, \delta_j)}{P(\delta_i)P(\delta_j)}$$

is the information gained from term pair (t_i, t_j) for discrimination in support of the dependent hypothesis H_1 against the independent hypothesis H_2 when (t_i, t_j) has the state value (δ_i, δ_j). In practice, $i(H_1 : H_2|(\delta_i, \delta_j))$ is referred to as the *mutual information* of terms t_i and t_j under the corresponding state value (δ_i, δ_j). Notice that, in contrast to Eq.(7.3), we here adopt notation $i(H_1 : H_2|(\delta_i, \delta_j))$ instead of $i(H_1 : H_2|(t_i, t_j))$ as a more general expression which can be applicable to any state value of a given term pair. In the remainder of this chapter 'amount of mutual information' and 'extent of dependence' are treated synonymously, and they can be measured by using the same measure $i(H_1 : H_2|(\delta_i, \delta_j))$.

It is interesting to notice that the mutual information between terms t_i and t_j under the state value (δ_i, δ_j) can also be written as

$$\begin{aligned} i(H_1 : H_2|(\delta_i, \delta_j)) &= \log \frac{P(\delta_i, \delta_j)}{P(\delta_i)P(\delta_j)} \\ &= \log \frac{P(\delta_i|\delta_j)}{P(\delta_i)} = \log P(\delta_i|\delta_j) - \log P(\delta_i) = i(\delta_i) - i(\delta_i|\delta_j) \end{aligned}$$

$$= \log \frac{P(\delta_j|\delta_i)}{P(\delta_j)} = \log P(\delta_j|\delta_i) - \log P(\delta_j) = i(\delta_j) - i(\delta_j|\delta_i)$$
$$= i(H_1 : H_2|(\delta_j, \delta_i)),$$

provided that $P(\delta_i) > 0$ and $P(\delta_j) > 0$. Thus, $i(H_1 : H_2|(\delta_i, \delta_j))$ is symmetric in δ_i and δ_j, and it can be positive or negative.

Particularly, when $(\delta_i, \delta_j) = (1, 1)$, we have

$$i(H_1 : H_2|(t_i, t_j)) = i(t_i) - i(t_i|t_j) = i(t_j) - i(t_j|t_i) = i(H_1 : H_2|(t_j, t_i)),$$

provided that $P(t_i) > 0$ and $P(t_j) > 0$.

Therefore, another way of looking at $i(H_1 : H_2|(t_i, t_j))$ would be as follows. Intuitively, when $i(H_1 : H_2|(t_i, t_j)) > 0$, it is a measure of the decrease in the uncertainty about the occurrence of term t_i (or term t_j) caused by the occurrence of term t_j (or term t_i). This can also be thought of as a measure of the positive information concerning the occurrence of term t_i (or term t_j) provided by the occurrence of term t_j (or term t_i).

Conversely, when $i(H_1 : H_2|(t_i, t_j)) < 0$, it is a measure of the increase in the uncertainty about the occurrence of t_i (or t_j) caused by the occurrence of t_j (or t_i). This can also be thought of as a measure of the negative information concerning the occurrence of t_i (or t_j) provided by the occurrence of t_j (or t_i).

7.1.3 Expected Mutual Information Measure

Let us further assume that the joint state distribution is absolutely continuous with respect to the product of its marginal distributions. Then, the *expected mutual information* contained in term pair (t_i, t_j) is defined [67, 106, 176] by

$$I(\delta_i, \delta_j) = \sum_{\delta_i, \delta_j = 1,0} P(\delta_i, \delta_j) \log \frac{P(\delta_i, \delta_j)}{P(\delta_i)P(\delta_j)} = \sum_{\delta_i, \delta_j = 1,0} P(\delta_i, \delta_j) i(H_1 : H_2|(\delta_i, \delta_j)).$$

It can be verified that $I(\delta_i, \delta_j) \geq 0$, with equality if and only if $P(\delta_i, \delta_j) = P(\delta_i)P(\delta_j)$ for $\delta_i, \delta_i = 1, 0$ [106]. This property tells us that, in the expectation, the mutual information received from term pair (t_i, t_j) is positive. There is no mutual information if terms t_i and t_j are statistically independent.

The essential reason for the assumption that $P(\delta_i, \delta_j) \ll P(\delta_i) \cdot P(\delta_j)$ for $\delta_i, \delta_j = 1, 0$ is to ensure that $I(\delta_i, \delta_j)$ is well-defined. That is, $P(\delta_i, \delta_j) i(H_1 : H_2|(\delta_i, \delta_j)) \neq \infty$ for $\delta_i, \delta_j = 1, 0$ under the notational conventions $0 \cdot \log\left(\frac{0}{0}\right) = 0$ and $0 \cdot \log\left(\frac{0}{a}\right) = 0$ (where $0 < a < +\infty$).

In practice, the information contained in a document is generally regarded as an *information entity*, and documents (as stored units) are usually treated as independent. Thus, if two terms appear in different documents, it is unlikely that they are statistically dependent (again, not in the sense of a semantic relation). Therefore, in this chapter, we restrict the 'distance' between terms to a document, that is, the mutual information of terms is estimated by using statistical data within a document. Also, one can estimate the mutual information of terms by using the statistical data within an information *subentity* (such as, a local context, abstract, summary, passages, etc.), or within an information *superentity* (such as, a set of the relevant sample documents). In these cases, the information subentity or superentity would in effect be considered as a new independent information entity.

As we will see at later stages, with such a restriction, i.e., considering the mutual information of terms within a specific information entity rather than within an extremely large and all-embracing information source (the whole collection), the score functions can be easily designed, and the total computation involved in the method proposed in the current chapter is not expensive, nor complex.

Thus, for a given information entity, denoted by E, let the joint state distribution of term pair (t_i, t_j) be $P_E(\delta_i, \delta_j)$, and its corresponding marginal distributions be $P_E(\delta_i)$ and $P_E(\delta_j)$. All these distributions are derived by using the statistical data within entity E. Assume that $P_E(\delta_i, \delta_j) \ll P_E(\delta_i) \cdot P_E(\delta_j)$ for $\delta_i, \delta_j = 1, 0$. Then the expected mutual information of terms t_i and t_j concerning entity E can be expressed as

$$I_E(\delta_i, \delta_j) = \sum_{\delta_i, \delta_j = 1,0} P_E(\delta_i, \delta_j) \log \frac{P_E(\delta_i, \delta_j)}{P_E(\delta_i) P_E(\delta_j)}. \tag{7.4}$$

It should be emphasized, in studying mutual information of terms, that our interest is in the fact that quantity $I_E(\delta_i, \delta_j)$ is 'concerning entity E'. In fact, for a given term pair (t_i, t_j), quantities $I_E(\delta_i, \delta_j)$ elicited from the statistical data within the different entities are likely to be different.

To speak of the mutual information of terms concerning a certain entity E, based on Shannon's basic ideas, the joint and marginal state distributions on which $I_E(\delta_i, \delta_j)$ is based must be set up. Thus, in the next section, we will focus mainly on the issue of the estimation of these distributions.

7.2 Estimation of Term State Distributions $P_d(\delta_i)$ and $P_d(\delta_i, \delta_j)$

There are various possible ways to estimate the state distributions. Usually, the estimation of the marginal state distributions $P_E(\delta_i)$ and $P_E(\delta_j)$ can be given easily by, for instance, the term probability distribution over V^E. The marginal state distributions can also be estimated based on a non-negative function over V^E, which may not be a term probability distribution; we shall shortly see such an example.

Thus, the main aim of this section is to study the estimation of the joint state distribution $P_d(\delta_i, \delta_j)$, which needs more complicated mathematical treatment. We start by considering three particular methods (as examples), in order to try to get ideas of what happens in the general situation, and then a general framework for the estimation is established.

7.2.1 Method A: Using Term Co-occurrence Data

To begin with, let us consider a given document. It will be found that it is very easily extended to other kinds of information entities.

For two arbitrary terms $t_i, t_j \in V^d$, using the statistical data of the co-occurrence of terms within document d, the estimation of $P_d(\delta_i, \delta_j)$ can be expressed as

$$P_d(\delta_i = 1, \delta_j = 1) = \frac{f_d(t_i) f_d(t_j)}{\sum_{i' < j'; t_{i'}, t_{j'} \in V^d} f_d(t_{i'}) f_d(t_{j'})} = \gamma_d(t_i, t_j),$$

$$\begin{aligned} P_d(\delta_i = 1, \delta_j = 0) &= P_d(\delta_i = 1) - P_d(\delta_i = 1, \delta_j = 1) = p_d(t_i) - \gamma_d(t_i, t_j), \\ P_d(\delta_i = 0, \delta_j = 1) &= P_d(\delta_j = 1) - P_d(\delta_i = 1, \delta_j = 1) = p_d(t_j) - \gamma_d(t_i, t_j), \\ P_d(\delta_i = 0, \delta_j = 0) &= 1 - P_d(\delta_i = 1) - P_d(\delta_j = 1) + P_d(\delta_i = 1, \delta_j = 1) \\ &= 1 - p_d(t_i) - p_d(t_j) + \gamma_d(t_i, t_j), \end{aligned} \tag{7.5}$$

where $\gamma_d(t_i, t_j)$ is a positive function and $p_d(t)$ is given in Eq.(7.1). Obviously, $P_d(\delta_i, \delta_j)$ in Eq.(7.5) is uniquely determined by $\gamma_d(t_i, t_j)$ and $p_d(t)$.

We first prove that Eq.(7.5) constitutes a probability distribution over $\Omega \times \Omega$ by establishing the following theorem and corollary.

Theorem 7.2.1 Given $t_j, t_j \in V^d$, for the expression given in Eq.(7.5), we have:

(1) $p_d(t_i) \geq \gamma_A(t_i, t_j)$ if and only if

$$\sum_{i'<j'; t_{i'}, t_{j'} \in V^d - \{t_j\}} f_d(t_{i'}) f_d(t_{j'}) \geq f_d(t_j) f_d(t_j);$$

(2) $p_d(t_j) \geq \gamma_A(t_i, t_j)$ if and only if

$$\sum_{i'<j'; t_{i'}, t_{j'} \in V^d - \{t_i\}} f_d(t_{i'}) f_d(t_{j'}) \geq f_d(t_i) f_d(t_i).$$

The detailed proof of this theorem is given in Section 10.4.

Corollary 7.2.1 For given $t_j, t_j \in V^d$, the expression given in Eq.(7.5) is a probability distribution if

$$\sum_{i'<j'; t_{i'}, t_{j'} \in V^d - \{t_j\}} f_d(t_{i'}) f_d(t_{j'}) \geq f_d(t_j) f_d(t_j),$$

$$\sum_{i'<j'; t_{i'}, t_{j'} \in V^d - \{t_i\}} f_d(t_{i'}) f_d(t_{j'}) \geq f_d(t_i) f_d(t_i).$$

Proof. Denote the denominator of probability $P_d(\delta_i = 1, \delta_j = 1)$ by

$$\varpi = \sum_{i'<j'; t_{i'}, t_{j'} \in V^d} f_d(t_{i'}) f_d(t_{j'}).$$

Obviously, $\varpi > 0$ (since $|V^d| \geq 2$), thus $P_d(\delta_i = 1, \delta_j = 1) > 0$ since $\gamma_d(t_i, t_j) = \frac{f_d(t_i) f_d(t_j)}{\varpi} > 0$. Also, $P_d(\delta_i = 1, \delta_j = 0) \geq 0$ since, under the corresponding condition, $p_d(t_i) \geq \gamma_d(t_i, t_j)$ by Theorem 7.2.1. Again, $P_d(\delta_i = 0, \delta_j = 1) \geq 0$ since, under the corresponding condition, $p_d(t_j) \geq \gamma_d(t_i, t_j)$ by Theorem 7.2.1. Finally, $P_d(\delta_i = 0, \delta_j = 0) \geq 0$ is derived as follows. From $f_d(t_i) + f_d(t_j) \leq ||d||$, we have

$$\frac{f_d(t_i) + f_d(t_j)}{||d||} \leq 1 \leq 1 + \frac{f_d(t_i) f_d(t_j)}{\varpi},$$

it follows immediately,

$$1 - p_d(t_i) - p_d(t_j) + \gamma_d(t_i, t_j) = 1 - \frac{f_d(t_i)}{||d||} - \frac{f_d(t_j)}{||d||} + \frac{f_d(t_i) f_d(t_j)}{\varpi} \geq 0.$$

Also, it is easily seen $\sum_{\delta_i,\delta_j=1,0} P_d(\delta_i,\delta_j) = 1$. The proof is complete.

Corollary 7.2.1 tells us that $P_d(\delta_i,\delta_j)$ in Eq.(7.5) is a probability distribution if it satisfies two inequalities given in the corollary. Whereas Theorem 7.2.1 states that these two inequalities can be verified as long as we have $p_d(t_i) \geq \gamma_d(t_i,t_j)$ and $p_d(t_j) \geq \gamma_d(t_i,t_j)$, which are much easier to verify in practical applications.

Let us see an example below, which will help to clarify the above ideas and understand the computations involved in Eq.(7.5).

Example 7.2.1 For a document $d = \{t_1,t_2,t_2,t_2,t_3,t_4\}$ in Example 7.1.1, we find

$$\begin{aligned}\varpi &= \sum_{i'<j';\, t_{i'},t_{j'}\in V^d} f_d(t_{i'})f_d(t_{j'}) \\ &= f_d(t_1)f_d(t_2) + f_d(t_1)f_d(t_3) + f_d(t_1)f_d(t_4) + \\ &\quad f_d(t_2)f_d(t_3) + f_d(t_2)f_d(t_4) + f_d(t_3)f_d(t_4) \\ &= 1\times 3 + 1\times 1 + 1\times 1 + 3\times 1 + 3\times 1 + 1\times 1 = 12.\end{aligned}$$

Thus, for instance, for term pair (t_1,t_2), we have

$$\begin{aligned}P_d(\delta_1=1,\delta_2=1) &= \gamma_d(t_1,t_2) = \frac{f_d(t_1)f_d(t_2)}{\varpi} = \frac{1\times 3}{12} = \frac{3}{12} > 0, \\ P_d(\delta_1=1,\delta_2=0) &= p_d(t_1) - \gamma_d(t_1,t_2) = \frac{1}{6} - \frac{3}{12} = -\frac{1}{12} < 0, \\ P_d(\delta_1=0,\delta_2=1) &= p_d(t_2) - \gamma_d(t_1,t_2) = \frac{3}{6} - \frac{3}{12} = \frac{3}{12} > 0, \\ P_d(\delta_1=0,\delta_2=0) &= 1 - p_d(t_1) - p_d(t_2) + \gamma_d(t_1,t_2) = 1 - \frac{1}{6} - \frac{3}{6} + \frac{3}{12} = \frac{7}{12} > 0,\end{aligned}$$

from which we can conclude that $P_d(\delta_1,\delta_2)$ is not a probability distribution since $p_d(t_1) - \gamma_d(t_1,t_2) < 0$. Also, we can verify this in another way:

$$\begin{aligned}&\sum_{i'<j';\, t_{i'},t_{j'}\in V^d-\{t_2\}} f_d(t_{i'})f_d(t_{j'}) = \sum_{i'<j';\, t_{i'},t_{j'}\in\{t_1,t_3,t_4\}} f_d(t_{i'})f_d(t_{j'}) \\ &= f_d(t_1)f_d(t_3) + f_d(t_1)f_d(t_4) + f_d(t_3)f_d(t_4) = 1+1+1 < 9 = f_d(t_2)f_d(t_2),\end{aligned}$$

$$\begin{aligned}&\sum_{i'<j';\, t_{i'},t_{j'}\in V^d-\{t_1\}} f_d(t_{i'})f_d(t_{j'}) = \sum_{i'<j';\, t_{i'},t_{j'}\in\{t_2,t_3,t_4\}} f_d(t_{i'})f_d(t_{j'}) \\ &= f_d(t_2)f_d(t_3) + f_d(t_2)f_d(t_4) + f_d(t_3)f_d(t_4) = 3+3+1 > 1 = f_d(t_1)f_d(t_1).\end{aligned}$$

That is, the first inequality given in Corollary 7.2.1 is not satisfied. ♠

From the above example, we can see that, as documents become longer, factor ϖ would become larger rapidly (then $\gamma_d(t_i,t_j)$ remains positive but becomes smaller rapidly). Thus, it should not be a problem to satisfy $p_d(t_i) \geq \gamma_d(t_i,t_j)$ and $p_d(t_j) \geq \gamma_d(t_i,t_j)$, for arbitrary terms $t_i,t_j \in V^d$, in practice.

Clearly, it is important to compute function $\gamma_d(t_i,t_j)$ in Eq.(7.5). From the proof of Theorem 7.2.1, we can see that the denominator of probability $P_d(\delta_i=1,\delta_j=1)$ can also be

expressed as

$$\begin{aligned}\varpi &= f_d(t_{i_1}) \sum_{j'=i_2,...,i_s} f_d(t_{j'}) + f_d(t_{i_2}) \sum_{j'=i_3,...,i_s} f_d(t_{j'}) + ... + f_d(t_{i_{s-1}}) \sum_{j'=i_s} f_d(t_{j'}) \\ &= \sum_{i'=i_1,i_2,...,i_{s-1}} \Big[f_d(t_{i'}) \sum_{j'>i'} f_d(t_{j'}) \Big],\end{aligned} \tag{7.6}$$

which might give a simpler formula for computing ϖ. An example of the computation can be found in Section 10.6 (see Example A).

The thought line of how individual probabilities in Eq.(7.5) are obtained is rather clear and intuitive: $P_d(\delta_i = 1, \delta_j = 0)$ and $P_d(\delta_i = 0, \delta_j = 1)$ are derived by means of constraints

$$\begin{aligned}P_d(\delta_i = 1, \delta_j = 0) + P_d(\delta_i = 1, \delta_j = 1) &= P_d(\delta_i = 1), \\ P_d(\delta_i = 0, \delta_j = 1) + P_d(\delta_i = 1, \delta_j = 1) &= P_d(\delta_j = 1),\end{aligned}$$

and probability $P_d(\delta_i = 0, \delta_j = 0)$ is derived by using another constraint

$$\sum_{\delta_i,\delta_j=1,0} P_d(\delta_i, \delta_j) = 1.$$

It is worth explaining the derivation of probability $P_d(\delta_i = 1, \delta_j = 1)$, which is normally more interesting for us, in more detail. Its numerator $f_d(t_i)f_d(t_j)$ characterizes the co-occurrence frequencies of t_i and t_j in document d. Its denominator ϖ, the sum of all possible numerators $f_d(t_{i'})f_d(t_{j'})$ for $i' < j'; i', j' \in \{i_1, i_2, ..., i_s\}$, is a normalization factor for the probability. An alternative way of looking at $P_d(\delta_i = 1, \delta_j = 1)$ is through an $n \times n$ matrix, called the *co-occurrence frequency matrix* of terms, which is discussed in Section 10.5.

It should be pointed out, in the above estimation, that we suppose that individual documents are represented by $M_d = [f_d(t)]_{1\times n}$. However, the estimation discussed here is independent of the specific document representation scheme. This implies that the estimation method can be applied to a more general representation scheme $M_d = [w_d(t)]_{1\times n}$. In this case, one need only estimate the state distributions using $w_d(t)$ (which should satisfy $w_d(t) > 0$ for all $t \in V^d$) instead of $f_d(t)$ as discussed above. Notice that Eq.(7.5) is determined by functions $\gamma_d(t_i, t_j)$ and $p_d(t)$. Thus, for $t_i, t_j \in V^d$, we can write down

$$\begin{aligned}P_d(\delta_i = 1) = p_d(t_i) &= \frac{w_d(t_i)}{\sum_{i' \in V^d} w_d(t')}, \\ P_d(\delta_i = 1, \delta_j = 1) = \gamma_d(t_i, t_j) &= \frac{w_d(t_i) w_d(t_j)}{\sum_{i'<j'; t_{i'}, t_{j'} \in V^d} w_d(t_{i'}) w_d(t_{j'})},\end{aligned}$$

and so forth.

Now, let us discuss further the extension of the above estimation to other information entities. First, we consider the relevant sample set Ξ^+. In this case, all documents in Ξ^+ are merged together to form a new (larger) document. Notice that $f_{\Xi^+}(t) = \sum_{d\in\Xi^+} f_d(t)$ (i.e., it is the sum of the frequencies of term t in individual documents in Ξ^+) and $||\Xi^+|| = \sum_{d\in\Xi^+} ||d||$ (i.e., it is the sum of the lengths of individual documents in Ξ^+). Thus, one can estimate the state distributions using the statistical data within set Ξ^+ as discussed above. That is, for

$t_i, t_j \in V^{\Xi+}$, we have

$$P_{\Xi+}(\delta_i = 1) = p_{\Xi+}(t_i) = \frac{f_{\Xi+}(t_i)}{||\Xi^+||},$$
$$P_{\Xi+}(\delta_i = 1, \delta_j = 1) = \gamma_{\Xi+}(t_i, t_j) = \frac{f_{\Xi+}(t_i) f_{\Xi+}(t_j)}{\sum_{i'<j'; t_{i'}, t_{j'} \in V^{\Xi+}} f_{\Xi+}(t_{i'}) f_{\Xi+}(t_{j'})}. \tag{7.7}$$

Next, consider a sub-document d_0 of document d as an information entity. In this case, the sub-document is viewed as a new (smaller) document, and frequency $f_{d_0}(t) \leq f_d(t)$ for every term $t \in V^{d_0} \subseteq V^d$ (it should satisfy $|V^{d_0}| \geq 2$). Then, one can also estimate the state distributions using the statistical data within sub-document d_0. That is, for $t_i, t_j \in V^{d_0}$, we may have

$$P_{d_0}(\delta_i = 1) = p_{d_0}(t_i) = \frac{f_{d_0}(t_i)}{||d_0||},$$
$$P_{d_0}(\delta_i = 1, \delta_j = 1) = \gamma_{d_0}(t_i, t_j) = \frac{f_{d_0}(t_i) f_{d_0}(t_j)}{\sum_{i'<j'; t_{i'}, t_{j'} \in V^{d_0}} f_{d_0}(t_{i'}) f_{d_0}(t_{j'})}. \tag{7.8}$$

7.2.2 Method B: Using Conditional Probabilities

As with our discussion of Method A, we start by considering a given document, and then extend the consideration to other information entities.

A typical example for the derivation of conditional probability in probability theory is the problem of randomly drawing a ball from an urn, without replacement. Before we can give an alternative estimation of distribution $P_d(\delta_i, \delta_j)$, let us first see a simple example which may clarify the idea involved.

Let us examine an urn containing six balls numbered $1, 2, 2, 2, 3, 4$. Except for the number assignment, the balls are identical in every detail; they are indistinguishable. Let two balls be drawn at random from the urn, one after the other, and their numbers noted; the first ball drawn is *not* returned to the urn before the second ball is drawn. For $i = 1, 2$, let A_i be the event that the ball drawn on the ith draw bears number i; we write that the outcome (event) of two draws is (A_1, A_2). We are thus seeking probabilities $P(A_1, A_2)$, $P(A_1, \bar{A}_2)$, $P(\bar{A}_1, A_2)$ and $P(\bar{A}_1, \bar{A}_2)$.

It is intuitively appealing that the conditional probability of drawing a ball with number 2 on the second draw, given that ball numbered 1 was drawn on the first draw, is $P(A_2|A_1) = \frac{3}{6-1}$, since before the second draw there were $6-1$ balls in the urn, of which 3 balls bore number 2. Also, the conditional probability of drawing a ball that does not bear number 2 on the second draw, given that ball of number 1 was drawn on the first draw, is $P(\bar{A}_2|A_1) = 1 - \frac{3}{6-1}$. Similarly, we can have $P(\bar{A}_1|A_2) = 1 - \frac{1}{6-1}$. Then, with the conditional probability formula given in probability theory, we can immediately write down

$$P(A_1, A_2) = P(A_1)P(A_2|A_1) = \frac{1}{6}\frac{3}{6-1} = \frac{3}{30},$$
$$P(A_1, \bar{A}_2) = P(A_1)P(\bar{A}_2|A_1) = \frac{1}{6}\left(1 - \frac{3}{6-1}\right) = \frac{2}{30},$$
$$P(\bar{A}_1, A_2) = P(A_2)P(\bar{A}_1|A_2) = \frac{3}{6}\left(1 - \frac{1}{6-1}\right) = \frac{12}{30},$$
$$P(\bar{A}_1, \bar{A}_2) = 1 - P(A_1, A_2) - P(A_1, \bar{A}_2) - P(\bar{A}_1, A_2) = \frac{13}{30}.$$

It should be noted that the last probability $P(\bar{A}_1, \bar{A}_2)$ is given according to a constraint $P(A_1, A_2) + P(A_1, \bar{A}_2) + P(\bar{A}_1, A_2) + P(\bar{A}_1, \bar{A}_2) = 1$. In fact, substituting

$$P(A_1) = P(A_1, A_2 \cup \bar{A}_2) = P(A_1, A_2) + P(A_1, \bar{A}_2),$$
$$P(\bar{A}_1) = P(\bar{A}_1, A_2 \cup \bar{A}_2) = P(\bar{A}_1, A_2) + P(\bar{A}_1, \bar{A}_2),$$

into $1-P(A_1) = P(\bar{A}_1)$, we can immediately obtain the constraint $1-[P(A_1, A_2)+P(A_1, \bar{A}_2)] = P(\bar{A}_1, A_2) + P(\bar{A}_1, \bar{A}_2)$.

The example above may be a prototype of our problem, which, as stated, involves the probabilities of four state values of term pairs, such as, (t_1, t_2), from a given document d.

Let us return to Example 7.2.1. In like manner, for document $d = \{t_1, t_2, t_2, t_2, t_3, t_4\}$ (thus $||d|| = 6$, $V^d = \{t_1, t_2, t_3, t_4\}$, $|V^d| = 4 \geq 2$, $f_d(t_1) = 1$ and $f_d(t_2) = 3$), let terms $t_1, t_2, t_2, t_2, t_3, t_4$ correspond to the balls bearing numbers $1, 2, 2, 2, 3, 4$, respectively. Also let propositions (t_1, t_2), $(t_1, \bar{t}_2)$, $(\bar{t}_1, t_2)$ and $(\bar{t}_1, \bar{t}_2)$ correspond to events (A_1, A_2), $(A_1, \bar{A}_2)$, $(\bar{A}_1, A_2)$ and $(\bar{A}_1, \bar{A}_2)$, respectively. Then we obtain immediately the same results as the above experiment of drawing balls as follows.

$$P_d(\delta_1 = 1, \delta_2 = 1) = P_d(\delta_1 = 1)P_d(\delta_2 = 1|\delta_1 = 1) = \frac{f_d(t_1)}{||d||}\frac{f_d(t_2)}{||d|| - 1} = \frac{3}{30},$$
$$P_d(\delta_1 = 1, \delta_2 = 0) = P_d(\delta_1 = 1)P_d(\delta_2 = 0|\delta_1 = 1) = \frac{f_d(t_1)}{||d||}\left(1 - \frac{f_d(t_2)}{||d|| - 1}\right) = \frac{2}{30},$$
$$P_d(\delta_1 = 0, \delta_2 = 1) = P_d(\delta_2 = 1)P_d(\delta_1 = 0|\delta_2 = 1) = \frac{f_d(t_2)}{||d||}\left(1 - \frac{f_d(t_1)}{||d|| - 1}\right) = \frac{12}{30},$$
$$P_d(\delta_1 = 0, \delta_2 = 0) = 1 - P_d(\delta_1 = 1, \delta_2 = 1) - P_d(\delta_1 = 1, \delta_2 = 0) - P_d(\delta_1 = 0, \delta_2 = 1) = \frac{13}{30}.$$

From the viewpoint of statistics, terms $t_1, t_2, ..., t_n$ are simply treated as some distinct abstract symbols without explicitly taking into consideration the real semantic meaning of individual terms. A document d can then be viewed as a 'multi-set' of the symbols since many of the symbols (terms) may not occur only once. Statistics is concerned with statistical frequencies $f_d(t_1), f_d(t_2), ..., f_d(t_n)$ of these symbols or, symbol weights $w_d(t_1), w_d(t_2), ..., w_d(t_n)$ derived from the statistical frequencies. Therefore, as abstract symbols, there is no essential difference between $t_1, t_2, ..., t_n$ and $1, 2, ..., n$ (numbers $1, 2, ..., n$ can also be thought of as abstract symbols). Thus, saying 'a ball bearing number i is drawn from an urn' is equivalent to saying 'number i is taken from an urn', and is equivalent to saying 'symbol t_i is found in a multi-set (i.e., term t_i occurs in a document)'. Also, saying 'number j is found on the second draw, given that number i was found on the first draw (without replacement)' is equivalent to saying 'symbol t_j is found in the remainder of the multi-set, given that symbol t_i was found'. Given the occurrence of term t_i will lead to a change of the probability distribution of the occurrence of other terms in document d. For instance, for document d in Example 7.2.1, the probability of the occurrence of term t_2 is $P_d(\delta_2 = 1) = p_d(t_2) = \frac{3}{6}$, but the conditional probability of the occurrence of term t_2, given that term t_1 occurred, is $P_d(\delta_2 = 1|\delta_1 = 1) = p_d(t_2|t_1) = \frac{3}{6-1} \neq \frac{3}{6}$.

Now let us establish the same fact in general. For arbitrary terms $t_i, t_j \in V^d$, using the conditional probabilities, the joint state distribution $P_d(\delta_i, \delta_j)$ can be formulated by

$$
\begin{aligned}
P_d(\delta_i = 1, \delta_j = 1) &= P_d(\delta_i = 1) P_d(\delta_j = 1 | \delta_i = 1) \\
&= \frac{f_d(t_i)}{||d||} \frac{f_d(t_j)}{||d|| - 1} = \gamma_d(t_i, t_j), \\
P_d(\delta_i = 1, \delta_j = 0) &= P_d(\delta_i = 1) P_d(\delta_j = 0 | \delta_i = 1) \\
&= \frac{f_d(t_i)}{||d||} \left(1 - \frac{f_d(t_j)}{||d|| - 1}\right) = p_d(t_i) - \gamma_d(t_i, t_j), \\
P_d(\delta_i = 0, \delta_j = 1) &= P_d(\delta_j = 1) P_d(\delta_i = 0 | \delta_j = 1) \\
&= \frac{f_d(t_j)}{||d||} \left(1 - \frac{f_d(t_i)}{||d|| - 1}\right) = p_d(t_j) - \gamma_d(t_i, t_j), \\
P_d(\delta_i = 0, \delta_j = 0) &= 1 - P_d(\delta_i = 1, \delta_j = 1) - P_d(\delta_i = 1, \delta_j = 0) - P_d(\delta_i = 0, \delta_j = 1) \\
&= 1 - p_d(t_i) - p_d(t_j) + \gamma_d(t_i, t_j),
\end{aligned}
\tag{7.9}
$$

where $\gamma_d(t_i, t_j)$ is a positive function and $p_d(t)$ is given in Eq.(7.1). Obviously, $P_d(\delta_i, \delta_j)$ in Eq.(7.9) is uniquely determined by $\gamma_d(t_i, t_j)$ and $p_d(t)$.

It is interesting to observe that the two joint distributions given in Eq.(7.5) and Eq.(7.9) are rather different, whereas they share the same marginal distributions given in Eq.(7.2). In practice, the joint probability distribution generally cannot be uniquely determined by its marginal probability distributions, that is, the character of a bivariable random vector can not be defined by the character of its individual components.

Similar to the discussion of Method A, if we consider the relevant sample set Ξ^+ as an information entity, for $t_i, t_j \in V^{\Xi^+}$, we can write

$$
\begin{aligned}
P_{\Xi^+}(\delta_i = 1) &= p_{\Xi^+}(t_i) = \frac{f_{\Xi^+}(t_i)}{||\Xi^+||}, \\
P_{\Xi^+}(\delta_i = 1, \delta_j = 1) &= \gamma_{\Xi^+}(t_i, t_j) = \frac{f_{\Xi^+}(t_i)}{||\Xi^+||} \frac{f_{\Xi^+}(t_j)}{||\Xi^+|| - 1}.
\end{aligned}
\tag{7.10}
$$

Also, if we consider a sub-document d_0 of document d as an information entity, for $t_i, t_j \in V^{d_0}$, we can write

$$
\begin{aligned}
P_{d_0}(\delta_i = 1) &= p_{d_0}(t_i) = \frac{f_{d_0}(t_i)}{||d_0||}, \\
P_{d_0}(\delta_i = 1, \delta_j = 1) &= \gamma_{d_0}(t_i, t_j) = \frac{f_{d_0}(t_i)}{||d_0||} \frac{f_{d_0}(t_j)}{||d_0|| - 1}.
\end{aligned}
\tag{7.11}
$$

7.2.3 Method C: Using Document Frequency Data

In some probabilistic methods involving the consideration of statistically dependent terms, one would state that the binary assumption suffices to specify the dependence of terms. The method discussed below is under this assumption.

Consider the relevant sample set Ξ^+ as an information entity, on which it is assumed that the statistical data $F_{\Xi^+}(t_i)$ (the number of documents in which term t_i occurs) and $F_{\Xi^+}(t_i, t_j)$ (the number of documents in which terms t_i and t_j co-occur) can be obtained. Define a function

$$
\phi_{\Xi^+}(t) = \frac{F_{\Xi^+}(t)}{|\Xi^+|} \qquad (t \in V^{\Xi^+}). \tag{7.12}
$$

It is clear that $0 < \phi_{\Xi^+}(t) \leq 1$ for every $t \in V^{\Xi^+}$ (since term t occurs in at least one document and at most all documents in Ξ^+).

Based on function $\phi_{\Xi^+}(t)$, for each term $t \in V^{\Xi^+}$, define

$$P_{\Xi^+}(\delta = 1) = \phi_{\Xi^+}(t) \quad \text{and} \quad P_{\Xi^+}(\delta = 0) = 1 - \phi_{\Xi^+}(t), \tag{7.13}$$

which is a probability distribution over $\Omega = \{1, 0\}$. Clearly, $0 < P_{\Xi^+}(\delta = 1) \leq 1$ and $0 \leq P_{\Xi^+}(\delta = 0) < 1$.

For arbitrary terms $t_i, t_j \in V^{\Xi^+}$, using the statistical data of the document frequencies concerning set Ξ^+, it is very easy to directly derive probabilities $P_{\Xi^+}(\delta_i, \delta_j)$ for $\delta_i, \delta_j = 1, 0$. Notice that the (total) number of documents in the relevant sample set is $|\Xi^+|$. Then, the probability that terms t_i and t_j co-occur should be $\frac{F_{\Xi^+}(t_i,t_j)}{|\Xi^+|}$ since the number of documents in which t_i and t_j co-occur is $F_{\Xi^+}(t_i, t_j)$. Also, the probability that term t_i occurs but term t_j does not occur should be $\frac{F_{\Xi^+}(t_i)-F_{\Xi^+}(t_i,t_j)}{|\Xi^+|}$ since the number of documents in which t_i occurs but t_j does not occur is $F_{\Xi^+}(t_i) - F_{\Xi^+}(t_i, t_j)$. Similarly, we can give the probability that term t_i does not occur but term t_j occurs $\frac{F_{\Xi^+}(t_j)-F_{\Xi^+}(t_i,t_j)}{|\Xi^+|}$. Finally, the probability that neither of the terms t_i and t_j occur should be $\frac{W(t_i,t_j)}{|\Xi^+|}$, where $W(t_i, t_j) = |\Xi^+| - F_{\Xi^+}(t_i) - F_{\Xi^+}(t_j) + F_{\Xi^+}(t_i, t_j)$ is the number of documents in which none of terms t_i and t_j occur. Therefore, the estimation of distribution $P_d(\delta_i, \delta_j)$ can be expressed as

$$\begin{aligned}
P_{\Xi^+}(\delta_i = 1, \delta_j = 1) &= \frac{F_{\Xi^+}(t_i, t_j)}{|\Xi^+|} = \gamma_{\Xi^+}(t_i, t_j), \\
P_{\Xi^+}(\delta_i = 1, \delta_j = 0) &= \frac{F_{\Xi^+}(t_i) - F_{\Xi^+}(t_i, t_j)}{|\Xi^+|} = \phi_{\Xi^+}(t_i) - \gamma_{\Xi^+}(t_i, t_j), \\
P_{\Xi^+}(\delta_i = 0, \delta_j = 1) &= \frac{F_{\Xi^+}(t_j) - F_{\Xi^+}(t_i, t_j)}{|\Xi^+|} = \phi_{\Xi^+}(t_j) - \gamma_{\Xi^+}(t_i, t_j), \\
P_{\Xi^+}(\delta_i = 0, \delta_j = 0) &= \frac{|\Xi^+| - F_{\Xi^+}(t_i) - F_{\Xi^+}(t_j) + F_{\Xi^+}(t_i, t_j)}{|\Xi^+|} \\
&= 1 - \phi_{\Xi^+}(t_i) - \phi_{\Xi^+}(t_j) + \gamma_{\Xi^+}(t_i, t_j),
\end{aligned} \tag{7.14}$$

where $\gamma_{\Xi^+}(t_i, t_j)$ is a non-negative function. Obviously, $P_{\Xi^+}(\delta_i, \delta_j)$ in Eq.(7.14) is uniquely determined by $\gamma_{\Xi^+}(t_i, t_j)$ and $\phi_{\Xi^+}(t)$.

An alternative way to derive distribution $P_{\Xi^+}(\delta_i, \delta_j)$ is to use a conditional probability formula: the conditional probability of observing term t_j occurs, given that term t_i occurred, is $P_d(\delta_j = 1 | \delta_i = 1) = \frac{F_{\Xi^+}(t_i,t_j)}{F_{\Xi^+}(t_i)}$, since before the observation there were $F_{\Xi^+}(t_i)$ documents in Ξ^+, in which t_i occurred. Also, the conditional probability of observing term t_j does not occur, given that term t_i occurred, is $P_d(\delta_j = 0 | \delta_i = 1) = 1 - \frac{F_{\Xi^+}(t_i,t_j)}{F_{\Xi^+}(t_i)}$. Similarly, we have $P_d(\delta_i = 0 | \delta_j = 1) = 1 - \frac{F_{\Xi^+}(t_i,t_j)}{F_{\Xi^+}(t_j)}$. Then, we can immediately write the expression:

$$\begin{aligned}
P_{\Xi^+}(\delta_i = 1, \delta_j = 1) &= P_d(\delta_i = 1) P_d(\delta_j = 1 | \delta_i = 1) \\
&= \frac{F_{\Xi^+}(t_i)}{|\Xi^+|} \frac{F_{\Xi^+}(t_i, t_j)}{F_{\Xi^+}(t_i)} = \frac{F_{\Xi^+}(t_i, t_j)}{|\Xi^+|}, \\
P_{\Xi^+}(\delta_i = 1, \delta_j = 0) &= P_d(\delta_i = 1) P_d(\delta_j = 0 | \delta_i = 1) \\
&= \frac{F_{\Xi^+}(t_i)}{|\Xi^+|} \left[1 - \frac{F_{\Xi^+}(t_i, t_j)}{F_{\Xi^+}(t_i)}\right] = \frac{F_{\Xi^+}(t_i)}{|\Xi^+|} - \frac{F_{\Xi^+}(t_i, t_j)}{|\Xi^+|},
\end{aligned}$$

$$
\begin{aligned}
P_{\Xi^+}(\delta_i = 0, \delta_j = 1) &= P_d(\delta_j = 1) P_d(\delta_i = 0 | \delta_j = 1) \\
&= \frac{F_{\Xi^+}(t_j)}{|\Xi^+|}\left[1 - \frac{F_{\Xi^+}(t_i, t_j)}{F_{\Xi^+}(t_j)}\right] = \frac{F_{\Xi^+}(t_j)}{|\Xi^+|} - \frac{F_{\Xi^+}(t_i, t_j)}{|\Xi^+|}, \\
P_{\Xi^+}(\delta_i = 0, \delta_j = 0) &= 1 - P_d(\delta_i = 1, \delta_j = 1) - P_d(\delta_i = 1, \delta_j = 0) - P_d(\delta_i = 0, \delta_j = 1) \\
&= 1 - \frac{F_{\Xi^+}(t_i)}{|\Xi^+|} - \frac{F_{\Xi^+}(t_i)}{|\Xi^+|} + \frac{F_{\Xi^+}(t_i, t_j)}{|\Xi^+|}.
\end{aligned}
$$

These results are in agreement with the ones given in Eq.(7.14).

Next, let us analyse the absolute continuity of distribution $P_{\Xi^+}(\delta_i, \delta_j)$ with respect to product $P_{\Xi^+}(\delta_i) \cdot P_{\Xi^+}(\delta_j)$ by the following theorem.

Theorem 7.2.3 For arbitrary terms $t_i, t_j \in V^{\Xi^+}$, $P_{\Xi^+}(\delta_i, \delta_j) \ll P_{\Xi^+}(\delta_i) \cdot P_{\Xi^+}(\delta_j)$ for $\delta_i, \delta_j = 1, 0$.

The detailed proof of this theorem is given in Section 10.4.

By the above theorem, we now can substitute estimates Eq.(7.13) and Eq.(7.14) into measure $I_{\Xi^+}(\delta_i, \delta_j)$, called $EMIM$ in [207], and further obtain:

$$
\begin{aligned}
I_{\Xi^+}(\delta_i, \delta_j) = {} & \gamma_{\Xi^+}(t_i, t_j) \log \frac{\gamma_{\Xi^+}(t_i, t_j)}{\phi_{\Xi^+}(t_i)\phi_{\Xi^+}(t_j)} \\
& + \left(\phi_{\Xi^+}(t_i) - \gamma_{\Xi^+}(t_i, t_j)\right) \log \frac{\phi_{\Xi^+}(t_i) - \gamma_{\Xi^+}(t_i, t_j)}{\phi_{\Xi^+}(t_i)\left(1 - \phi_{\Xi^+}(t_j)\right)} \\
& + \left(\phi_{\Xi^+}(t_j) - \gamma_{\Xi^+}(t_i, t_j)\right) \log \frac{\phi_{\Xi^+}(t_j) - \gamma_{\Xi^+}(t_i, t_j)}{\left(1 - \phi_{\Xi^+}(t_i)\right)\phi_{\Xi^+}(t_j)} \\
& + \left(1 - \phi_{\Xi^+}(t_i) - \phi_{\Xi^+}(t_j) + \gamma_{\Xi^+}(t_i, t_j)\right) \times \\
& \quad \times \log \frac{1 - \phi_{\Xi^+}(t_i) - \phi_{\Xi^+}(t_j) + \gamma_C(t_i, t_j)}{\left(1 - \phi_{\Xi^+}(t_i)\right)\left(1 - \phi_{\Xi^+}(t_j)\right)} \\
= {} & \Big[\frac{F_{\Xi^+}(t_i, t_j)}{|\Xi^+|} \log \frac{F_{\Xi^+}(t_i, t_j)}{F_{\Xi^+}(t_i)F_{\Xi^+}(t_j)} \\
& + \frac{F_{\Xi^+}(t_i) - F_{\Xi^+}(t_i, t_j)}{|\Xi^+|} \log \frac{F_{\Xi^+}(t_i) - F_{\Xi^+}(t_i, t_j)}{F_{\Xi^+}(t_i)\left(|\Xi^+| - F_{\Xi^+}(t_j)\right)} \\
& + \frac{F_{\Xi^+}(t_j) - F_{\Xi^+}(t_i, t_j)}{|\Xi^+|} \log \frac{F_{\Xi^+}(t_j) - F_{\Xi^+}(t_i, t_j)}{\left(|\Xi^+| - F_{\Xi^+}(t_i)\right)F_{\Xi^+}(t_j)} \\
& + \frac{|\Xi^+| - F_{\Xi^+}(t_i) - F_{\Xi^+}(t_j) + F_{\Xi^+}(t_i, t_j)}{|\Xi^+|} \times \\
& \quad \times \log \frac{|\Xi^+| - F_{\Xi^+}(t_i) - F_{\Xi^+}(t_j) + F_{\Xi^+}(t_i, t_j)}{\left(|\Xi^+| - F_{\Xi^+}(t_i)\right)\left(|\Xi^+| - F_{\Xi^+}(t_j)\right)} \Big] + \\
& \Big[\frac{F_{\Xi^+}(t_i, t_j)}{|\Xi^+|} + \frac{F_{\Xi^+}(t_i) - F_{\Xi^+}(t_i, t_j)}{|\Xi^+|} + \frac{F_{\Xi^+}(t_j) - F_{\Xi^+}(t_i, t_j)}{|\Xi^+|} + \\
& + \frac{|\Xi^+| - F_{\Xi^+}(t_i) - F_{\Xi^+}(t_j) + F_{\Xi^+}(t_i, t_j)}{|\Xi^+|} \Big] \times \log |\Xi^+| \\
= {} & \Big[F_{\Xi^+}(t_i, t_j) \log \frac{F_{\Xi^+}(t_i, t_j)}{F_{\Xi^+}(t_i)F_{\Xi^+}(t_j)}
\end{aligned}
$$

$$
\begin{aligned}
&+ \big(F_{\Xi^+}(t_i) - F_{\Xi^+}(t_i,t_j)\big) \log \frac{F_{\Xi^+}(t_i) - F_{\Xi^+}(t_i,t_j)}{F_{\Xi^+}(t_i)\big(|\Xi^+| - F_{\Xi^+}(t_j)\big)} \\
&+ \big(F_{\Xi^+}(t_j) - F_{\Xi^+}(t_i,t_j)\big) \log \frac{F_{\Xi^+}(t_j) - F_{\Xi^+}(t_i,t_j)}{\big(|\Xi^+| - F_{\Xi^+}(t_i)\big)F_{\Xi^+}(t_j)} \\
&+ \big(|\Xi^+| - F_{\Xi^+}(t_i) - F_{\Xi^+}(t_j) + F_{\Xi^+}(t_i,t_j)\big)\times \\
&\quad \times \log \frac{|\Xi^+| - F_{\Xi^+}(t_i) - F_{\Xi^+}(t_j) + F_{\Xi^+}(t_i,t_j)}{\big(|\Xi^+| - F_{\Xi^+}(t_i)\big)\big(|\Xi^+| - F_{\Xi^+}(t_j)\big)} \Big] \times \frac{1}{|\Xi^+|} + \log|\Xi^+| \\
= \; & emim_{\Xi^+}(\delta_i,\delta_j) \times \frac{1}{|\Xi^+|} + \log|\Xi^+|,
\end{aligned}
$$

in which, a scale factor $\frac{1}{|\Xi^+|}$ and a constant $\log|\Xi^+|$ are independent of all term pairs $(t_i,t_j) \in V^{\Xi^+} \times V^{\Xi^+}$ (where $i \neq j$). Thus, we can eliminate the scale factor and constant. Adopting Van Rijsbergen's notation [207], we obtain the following equivalent (i.e., strictly monotone), but simpler, measure:

$$
emim_{\Xi^+}(\delta_i,\delta_j) = n_{11} \log \frac{n_{11}}{n_{1\cdot}n_{\cdot 1}} + n_{10} \log \frac{n_{10}}{n_{1\cdot}n_{\cdot 0}} + n_{01} \log \frac{n_{01}}{n_{0\cdot}n_{\cdot 1}} + n_{00} \log \frac{n_{00}}{n_{0\cdot}n_{\cdot 0}},
$$

in which,

$$
\begin{aligned}
n_{11} &= F_{\Xi^+}(t_i,t_j), & n_{1\cdot} &= F_{\Xi^+}(t_i), \\
n_{10} &= F_{\Xi^+}(t_i) - F_{\Xi^+}(t_i,t_j), & n_{0\cdot} &= |\Xi^+| - F_{\Xi^+}(t_i), \\
n_{01} &= F_{\Xi^+}(t_j) - F_{\Xi^+}(t_i,t_j), & n_{\cdot 1} &= F_{\Xi^+}(t_j), \\
n_{00} &= |\Xi^+| - F_{\Xi^+}(t_i) - F_{\Xi^+}(t_j) + F_{\Xi^+}(t_i,t_j), & n_{\cdot 0} &= |\Xi^+| - F_{\Xi^+}(t_j).
\end{aligned}
$$

An essential difference between Eq.(7.13) and Eq.(7.14) and n_{11} through $n_{\cdot 0}$ is that the former are normalized by factor $|\Xi^+|$ but the latter are not.

It is very important to notice that, from the relation

$$
emim_{\Xi^+}(\delta_i,\delta_j) = \big[I_{\Xi^+}(\delta_i,\delta_j) - \log|\Xi^+|\big] \times |\Xi^+|,
$$

we can clearly see that $I_{\Xi^+}(\delta_i,\delta_j) \geq 0$ cannot infer $emim_{\Xi^+}(\delta_i,\delta_j) \geq 0$. The following theorem is interesting.

Theorem 7.2.4 For arbitrary terms $t_i, t_j \in V^{\Xi^+}$, $emim_{\Xi^+}(\delta_i,\delta_j) \leq 0$.

The proof of the above theorem is simple (see Section 10.4).

The fact that the individual items of $emim_{\Xi^+}(\delta_i,\delta_j)$ are non-positive can also be seen directly by the relations:

$$
n_{1\cdot} = n_{11} + n_{10}, \quad n_{0\cdot} = n_{01} + n_{00}, \quad n_{\cdot 1} = n_{11} + n_{01}, \quad n_{\cdot 0} = n_{10} + n_{00}.
$$

In Theorem 7.2.3, we proved that, when $\phi_{\Xi^+}(t_i) = 1$, i.e., $F_{\Xi^+}(t_i) = |\Xi^+|$, $P_{\Xi^+}(\delta_i,\delta_j)$ was still absolutely continuous with respect to $P_{\Xi^+}(\delta_i) \cdot P_{\Xi^+}(\delta_j)$. The following theorem (the proof is given in Section 10.4) tells us another interesting fact.

Theorem 7.2.5 For terms $t_i, t_j \in V^{\Xi^+}$, if $F_{\Xi^+}(t_i) = |\Xi^+|$ then $I_{\Xi^+}(\delta_i,\delta_j) = 0$.

Theorem 7.2.5 tells us that, when $F_{\Xi^+}(t_i) = |\Xi^+|$, the occurrence of term t_i (in all sample documents) will not provide any information about the occurrence of term t_j (in some sample documents). Thus, these two terms are independent of one another concerning set Ξ^+.

Consequently, in IR applications, we should require that the sample set satisfies $|\Xi^+| > 1$. In fact, in order to avoid too many terms having $F_{\Xi^+}(t_i) = |\Xi^+|$, we usually take a relatively larger sample set, for instance, $|\Xi^+| \geq 5$.

Some examples on aspects of the above discussions can be found in Section 10.6 (see Example C).

Estimations $I_{\Xi^+}(\delta_i, \delta_j)$ and $emim_{\Xi^+}(\delta_i, \delta_j)$ are well-known to all IR researchers. It was initially introduced by Van Rijsbergen in his earlier book and papers [206, 207]. It is interesting to notice that the ways of deriving the estimation of $I_{\Xi^+}(\delta_i, \delta_j)$ and $emim_{\Xi^+}(\delta_i, \delta_j)$ given here are very different from that of the one given there.

7.2.4 A General Framework for Estimation

Having described three methods of estimating $P_E(\delta_i, \delta_j)$ by giving specific expressions in Eq.(7.5), Eq.(7.9) and Eq.(7.14), respectively, you may realize that their consistency would suggest unified expressions.

The Unified Expressions

Given an information entity E, introduce a *positive* function $\psi_E : V^E \to (0, 1)$, and a *non-negative* function $\gamma_E : V^E \times V^E \to [0, 1]$ satisfying $\gamma_E(t_i, t_j) \leq \psi_E(t_i)$, and $\gamma_E(t_i, t_j) \leq \psi_E(t_j)$, and $\gamma_E(t_i, t_j) \geq \psi_E(t_i) + \psi_E(t_j) - 1$.

Based on function $\psi_E(t)$, for a given term $t \in V^E$, define

$$P_E(\delta = 1) = \psi_E(t) \quad \text{and} \quad P_E(\delta = 0) = 1 - \psi_E(t). \tag{7.15}$$

Also, based on function $\gamma(t_i, t_j)$, for a given term pair $(t_i, t_j) \in V^E \times V^E$ $(i \neq j)$, define

$$P_E(\delta_i = 1, \delta_j = 1) = \gamma_E(t_i, t_j).$$

Then, we can give a unified expression for the estimation of $P_E(\delta_i, \delta_j)$ as follows.

$$\begin{aligned} P_E(\delta_i = 1, \delta_j = 1) &= \gamma_E(t_i, t_j), \\ P_E(\delta_i = 1, \delta_j = 0) &= \psi_E(t_i) - \gamma_E(t_i, t_j), \\ P_E(\delta_i = 0, \delta_j = 1) &= \psi_E(t_j) - \gamma_E(t_i, t_j), \\ P_E(\delta_i = 0, \delta_j = 0) &= 1 - \psi_E(t_i) - \psi_E(t_j) + \gamma_E(t_i, t_j), \end{aligned} \tag{7.16}$$

which is obviously uniquely determined by $\gamma_E(t_i, t_j)$ and $\psi_E(t)$. The following discussion is necessary.

A Few Points of Discussion

There are a few important points to make about the unified expressions above.

* From $0 < \psi_E(t) < 1$, it is clear that $P_E(\delta) > 0$ for $\delta = 1, 0$ and $\sum_{\delta=1,0} P_E(\delta) = 1$. Thus, $P_E(\delta)$ is a probability distribution over Ω.

* From the way we introduce function $\gamma_E(t_i, t_j)$, it can be seen

$$\begin{aligned} &\gamma_E(t_i, t_j) \geq 0, \\ &\psi_E(t_i) - \gamma_E(t_i, t_j) \geq 0, \\ &\psi_E(t_j) - \gamma_E(t_i, t_j) \geq 0, \\ &1 - \psi_E(t_i) - \psi_E(t_j) + \gamma_E(t_i, t_j) \geq 0, \end{aligned} \tag{7.17}$$

i.e., $P_E(\delta_i, \delta_j) \geq 0$ for $\delta_i, \delta_j = 1, 0$. Also, $\sum_{\delta_i, \delta_j = 1,0} P_E(\delta_i, \delta_j) = 1$. Thus, $P_E(\delta_i, \delta_j)$ is a probability distribution over $\Omega \times \Omega$.

* It can be easily verified that

$$\sum_{\delta_j = 1,0} P_E(\delta_i = 1, \delta_j) = \gamma_E(t_i, t_j) + \psi_E(t_i) - \gamma_E(t_i, t_j) = \psi_E(t_i) = P_E(\delta_i = 1),$$

$$\begin{aligned} \sum_{\delta_j = 1,0} P_E(\delta_i = 0, \delta_j) &= \psi_E(t_j) - \gamma_E(t_i, t_j) + 1 - \psi_E(t_i) - \psi_E(t_j) + \gamma_E(t_i, t_j) \\ &= 1 - \psi_E(t_i) = P_E(\delta_i = 0); \end{aligned}$$

and that

$$\sum_{\delta_i = 1,0} P_E(\delta_i, \delta_j = 1) = \gamma_E(t_i, t_j) + \psi_E(t_j) - \gamma_E(t_i, t_j) = \psi_E(t_j) = P_E(\delta_j = 1),$$

$$\begin{aligned} \sum_{\delta_i = 1,0} P_E(\delta_i, \delta_j = 0) &= \psi_E(t_i) - \gamma_E(t_i, t_j) + 1 - \psi_E(t_i) - \psi_E(t_j) + \gamma_E(t_i, t_j) \\ &= 1 - \psi_E(t_j) = P_E(\delta_j = 0), \end{aligned}$$

from which, we immediately obtain

$$\sum_{\delta_j = 1,0} P_E(\delta_i, \delta_j) = P_E(\delta_i) \quad \textit{and} \quad \sum_{\delta_i = 1,0} P_E(\delta_i, \delta_j) = P_E(\delta_j).$$

Therefore, $P_E(\delta_i)$ and $P_E(\delta_j)$ are the marginal distributions of distribution $P_E(\delta_i, \delta_j)$.

* Finally, for arbitrary terms $t_i, t_j \in V^E$, $P_E(\delta_i, \delta_j) \ll P_E(\delta_i) \cdot P_E(\delta_j)$ for $\delta_i, \delta_j = 1, 0$. In fact, $P_E(\delta_i) \cdot P_E(\delta_j) \neq 0$ since $0 < P_E(\delta) < 1$ for $\delta = 0, 1$.

Notice that function $\psi_E(t)$ may or may not be a probability distribution over V^E. Some examples can be found in the above methods: $p_d(t)$ in Method A and Method B is a probability distribution over V^d, whereas $\phi_{\Xi^+}(t)$ in Method C is not a probability distribution over V^{Ξ^+} since its denominator is a scale factor $|\Xi^+|$ rather than a normalization factor $\sum_{t \in V^{\Xi^+}} F_{\Xi^+}(t)$.

Generally, function $\psi_E(t)$ is required to satisfy $0 < \psi_E(t) < 1$ so as to guarantee $P_E(\delta_i) \cdot P_E(\delta_j) \neq 0$, and $P_E(\delta_i, \delta_j) \ll P_E(\delta_i) \cdot P_E(\delta_j)$ for $\delta_i, \delta_j = 1, 0$. The estimations given in Methods A and B are for this case. However, one can also define $0 \leq \psi_E(t) \leq 1$. In this case, it is necessary to verify $P_E(\delta_i, \delta_j) \ll P_E(\delta_i) \cdot P_E(\delta_j)$, particularly for those points such that $\psi_E(t_i) = 0$ and/or $\psi_E(t_j) = 0$, and $\psi_E(t_i) = 1$ and/or $\psi_E(t_j) = 1$. An example for this can be found in Method C, in which, $0 < \phi_E(t) \leq 1$, and Theorem 7.2.3 serves for proving the absolute continuity for arbitrary terms $t_i, t_j \in V^{\Xi^+}$.

Obviously, $P_E(\delta)$ is a probability distribution if $0 < \psi_E(t) < 1$. Also, $P_E(\delta_i, \delta_j)$ is a probability distribution if it satisfies all conditions given in Eq.(7.17). ($\sum_{\delta_i,\delta_j=1,0} P_E(\delta_i, \delta_j) = 1$ must hold for any estimation possessing the form of the unified expression). Thus, the key for any estimation of $P_E(\delta_i, \delta_j)$ is to verify these conditions. Corollary 7.2.1 is an example of the verification for the estimation given in Method A. The estimates given by Methods C and B are derived from the conditional probability formulae, and thus are clearly probability distributions.

For any estimate possessing the form of the unified expression, it is shown that $P_E(\delta_i)$ and $P_E(\delta_j)$ must be the marginal distributions of distribution $P_E(\delta_i, \delta_j)$. The estimates given by Methods A, B and C are examples for this case.

Remark

The issue of the mutual information of terms is an active research subject in IR. A variety of methods have been developed in order to assign a 'similarity' value to every pair of terms, and then some decision(s) are made on those values. The discrimination measures can influence retrieval performance significantly. However, it seems that only the '*form*' of the measures has frequently been a focus of research in IR literature, whereas the problem of verification of the probability distributions was often ignored as an unimportant matter. This implicitly means that a function with the form

$$i(x, y) = \log \frac{P(x, y)}{P_1(x)P_2(y)}$$

would be a 'mutual information measure', and that the discussion on $P(x, y)$, $P_1(x)$ and $P_2(y)$ in the function is trivial.

It is not true indeed. In fact, if expressions $P(x, y)$, $P_1(x)$ and $P_2(y)$ are not probability distributions, then function $i(x, y)$ would not be a mutual information measure in an information-theoretic sense. Neither would it be a mutual information measure if $P_1(x)$ and $P_2(y)$ are not marginal distributions of the joint distribution $P(x, y)$, even though they are all probability distributions. It may even not converge if $P(x, y) \ll P_1(x) \cdot P_2(y)$ does not hold. The mathematical interpretation for all these points can be found in Section 7.1.

Section 7.5, we will devote to a detailed account of the concept of association of a term with the context of the query based on the state distributions. Before doing so let us give an in-depth investigation of the discrimination measure in the sense of the mutual information of terms.

7.3 Discrimination Measure $\text{ifd}_M(t)$

So far, we have concentrated on developing a unified method for tackling various estimations of the state distributions. Before seeing how to apply our knowledge of these estimates to more practical problems, we need to study further the dependence discrimination measures and find out their relations, which underpin the method proposed in this chapter.

7.3.1 Definition of Discrimination Measures

Having discussed some relevance discrimination measures in the previous chapters, you may have thought of, in order to measure the expected mutual information for a given term pair,

that we need first to derive the contributions made by its individual state values to the expected mutual information. Thus, let us return to Eq.(7.4). We use symbol M to indicate 'Mutual information'. For a given information entity E and terms $t_i, t_j \in V^E$, denote the first item of $I_E(\delta_i, \delta_j)$ by

$$\mathbf{ifd}_M^E(t_i, t_j) = P_E(\delta_i = 1, \delta_j = 1) i_E(H_1 : H_2 | (\delta_i = 1, \delta_j = 1)),$$

which indicates information for discrimination of the dependence of terms t_i and t_j when proposition (t_i, t_j) is true in entity E; also, denote the second item by

$$\mathbf{ifd}_M^E(t_i, \bar{t}_j) = P_E(\delta_i = 1, \delta_j = 0) i_E(H_1 : H_2 | (\delta_i = 1, \delta_j = 0)),$$

which indicates information for discrimination of the dependence terms t_i and t_j when proposition $(t_i, \bar{t}_j)$ is true in entity E; and so forth.

Then, the expected mutual information of terms t_i and t_j concerning the information entity E can be expressed as a sum of the items,

$$\begin{aligned} I_E(\delta_i, \delta_j) &= \mathbf{ifd}_M^E(t_i, t_j) + \mathbf{ifd}_M^E(t_i, \bar{t}_j) + \mathbf{ifd}_M^E(\bar{t}_i, t_j) + \mathbf{ifd}_M^E(\bar{t}_i, \bar{t}_j) \\ &= \sum_{\delta_i, \delta_j = 0,1} \mathbf{ifd}_M^E(t_i^{\delta_i}, t_j^{\delta_j}). \end{aligned}$$

As in the foregoing, the amount of the mutual information $i_E(H_1 : H_2 | (\delta_i, \delta_j))$ measures the power of terms t_i and t_j to discriminate two opposite dependence hypotheses H_1 and H_2 under the corresponding state value (δ_i, δ_j). The magnitude of probability $P_E(\delta_i, \delta_j)$ measures the significance of the corresponding state value (δ_i, δ_j) in determining the power of discrimination. Thus, quantity $\mathbf{ifd}_M^E(t_i^{\delta_i}, t_j^{\delta_j})$ indicates the mutual 'information of terms t_i and t_j for discrimination' in favour of the dependent hypothesis H_1 against the independent hypothesis H_2, under the state values (δ_i, δ_j) concerning entity E. We may formulate these statements by a more formal definition as follows.

Definition 7.3.1 Given an information entity E and terms $t_i, t_j \in V^E$, let the joint state distribution of (t_i, t_j) be $P_E(\delta_i, \delta_j)$, and its corresponding marginal distributions be $P_E(\delta_i)$ and $P_E(\delta_j)$. Assume $P_E(\delta_i, \delta_j) \ll P_E(\delta_i) \cdot P_E(\delta_j)$ for $\delta_i, \delta_j = 1, 0$. The information for discriminating their dependence under state value (δ_i, δ_j) is defined by

$$\mathbf{ifd}_M^E(t_i^{\delta_i}, t_j^{\delta_j}) = P_E(\delta_i, \delta_j) \log \frac{P_E(\delta_i, \delta_j)}{P_E(\delta_i) P_d(\delta_j)} = P_E(\delta_i, \delta_j) i_E(H_1 : H_2 | (\delta_i, \delta_j)),$$

which is referred to as the (dependence) discrimination measure of term pair (t_i, t_j), and $i(H_1 : H_2 | (\delta_i, \delta_j))$ the (dependence) discrimination factor of term pair (t_i, t_j).

In most of the discussion in this chapter, we assume terms $t_i, t_j \in V^E$. However, for convenient discussion at the application stage, we also make the following definition.

Definition 7.3.2 Given an information entity E and terms $t_i, t_j \in V$, for $t_i \notin V^E$ or/and $t_j \notin V^E$, define

$$\mathbf{ifd}_M^E(t_i^{\delta_i}, t_j^{\delta_j}) = 0 \qquad \text{for } \delta_i, \delta_j = 1, 0.$$

That is, the contributions made by individual state values (δ_i, δ_j) to summation $I_E(\delta_i, \delta_j)$ will receive zero, and hence we have $I_E(\delta_i, \delta_j) = 0$. This is equivalent to stating that two

terms contain mutual information (whether more or less) only when they co-occur in some entity E (see Section 10.5 for the appropriateness of such a definition).

It is important to notice, unlike the expected information discussed in previous chapters, that the contributions to $I_E(\delta_i, \delta_j)$ made by the statistical quantities should refer to the individual state values for $\delta_i, \delta_j = 1, 0$, rather than to term pair (t_i, t_j) itself.

7.3.2 Interpretation of Discrimination Measures

Similar to the discussion given in Section 3.4, for a given state value (δ_i, δ_j), we have the following interpretations.

- ☞ If $P_E(\delta_i, \delta_j) = P_E(\delta_i) \cdot P_E(\delta_j)$, then the discrimination factor $i_E(H_1 : H_2|(\delta_i, \delta_j)) = 0$, and the state value (δ_i, δ_j) gives us no discrimination information about the dependence judgement, and the corresponding quantity $\mathbf{ifd}_M^E(t_i^{\delta_i}, t_j^{\delta_j}) = 0$.
- ☞ If $P_E(\delta_i, \delta_j) > P_E(\delta_i) \cdot P_E(\delta_j)$, then the discrimination factor $i_E(H_1 : H_2|(\delta_i, \delta_j)) > 0$, the discrimination measure indicates that the state value (δ_i, δ_j) contributes quantity $\mathbf{ifd}_M^E(t_i^{\delta_i}, t_j^{\delta_j}) > 0$ for supporting the dependent hypothesis H_1.
- ☞ If $P_E(\delta_i, \delta_j) < P_E(\delta_i) \cdot P_E(\delta_j)$, then the discrimination factor $i_E(H_1 : H_2|(\delta_i, \delta_j)) < 0$, the discrimination measure indicates that the state value (δ_i, δ_j) contributes quantity $\mathbf{ifd}_M^E(t_i^{\delta_i}, t_j^{\delta_j}) < 0$ for supporting the dependent hypothesis H_1.

Recall that, if $P_R(t) > P_{\bar{R}}(t)$ and $\mathbf{ifd}_I(t) > 0$ (or $\mathbf{ifd}_J(t) > 0$, or $\mathbf{ifd}_K(t) > 0$), then term t contributes a positive information quantity for supporting the relevant hypothesis H_1. In like manner, for a given state value (δ_i, δ_j), in order to discriminate whether it contributes a positive information quantity for supporting the dependent hypothesis H_1, a key point is to derive the relation between $P_E(\delta_i, \delta_j)$ and $P_E(\delta_i) \cdot P_E(\delta_j)$. Obviously, different state values might have different relations. Thus, we have to obtain all relations for $\delta_i, \delta_j = 1, 0$ for determining the corresponding signs of measures $\mathbf{ifd}_M^E(t_i^{\delta_i}, t_j^{\delta_j})$.

7.3.3 Properties of Discrimination Measures

Now, substituting estimates $P_E(\delta)$ and $P_E(\delta_i, \delta_j)$ given in Eq.(7.15) and Eq.(7.16), respectively, into $\mathbf{ifd}_M^E(t_i^{\delta_i}, t_j^{\delta_j})$, we obtain the following (four) general expressions of the discrimination measures:

$$
\begin{aligned}
\mathbf{ifd}_M^E(t_i, t_j) &= \gamma_E(t_i, t_j) \log \frac{\gamma_E(t_i, t_j)}{\psi_E(t_i)\psi_E(t_j)}, \\
\mathbf{ifd}_M^E(t_i, \bar{t}_j) &= \big(\psi_E(t_i) - \gamma_E(t_i, t_j)\big) \log \frac{\psi_E(t_i) - \gamma_E(t_i, t_j)}{\psi_E(t_i)\big(1 - \psi_E(t_j)\big)}, \\
\mathbf{ifd}_M^E(\bar{t}_i, t_j) &= \big(\psi_E(t_j) - \gamma_E(t_i, t_j)\big) \log \frac{\psi_E(t_j) - \gamma_E(t_i, t_j)}{\big(1 - \psi_E(t_i)\big)\psi_E(t_j)}, \\
\mathbf{ifd}_M^E(\bar{t}_i, \bar{t}_j) &= \big(1 - \psi_E(t_i) - \psi_E(t_j) + \gamma_E(t_i, t_j)\big) \log \frac{1 - \psi_E(t_i) - \psi_E(t_j) + \gamma_E(t_i, t_j)}{\big(1 - \psi_E(t_i)\big)\big(1 - \psi_E(t_j)\big)}.
\end{aligned}
\tag{7.18}
$$

Obviously, for a given entity E, the expressions are uniquely determined by functions $\gamma_E(t_i, t_j)$ and $\psi_E(t_i)$.

The following theorem enables us to gain an insight into the signs of $\mathbf{ifd}_M^E(t_i^{\delta_i}, t_j^{\delta_j})$ for $\delta_i, \delta_j = 1, 0$ by deriving only one single relation. The relation is the one between $P_E(\delta_i, \delta_j)$ and $P_E(\delta_i) \cdot P_E(\delta_j)$ under $(\delta_i, \delta_j) = (1, 1)$, i.e., between $\gamma_E(t_i, t_j)$ and $\psi_E(t_i) \cdot \psi_E(t_j)$. The signs of inequalities in the theorem should be carefully noted.

Theorem 7.3.1 Given $t_i, t_j \in V^E$, suppose that $I_E(\delta_i, \delta_j)$ is estimated by using the unified expressions given in Eq.(7.15) and Eq.(7.16), we have

(1) if $\gamma_E(t_i, t_j) = \psi_E(t_i)\psi_E(t_j)$, then $\mathbf{ifd}_M^E(t_i, t_j) = 0$, $\mathbf{ifd}_M^E(\bar{t}_i, \bar{t}_j) = 0$, $\mathbf{ifd}_M^E(t_i, \bar{t}_j) = 0$ and $\mathbf{ifd}_M^E(\bar{t}_i, t_j) = 0$;

(2) if $\gamma_E(t_i, t_j) > \psi_E(t_i)\psi_E(t_j)$, then $\mathbf{ifd}_M^E(t_i, t_j) > 0$, $\mathbf{ifd}_M^E(\bar{t}_i, \bar{t}_j) > 0$, $\mathbf{ifd}_M^E(t_i, \bar{t}_j) \leq 0$ and $\mathbf{ifd}_M^E(\bar{t}_i, t_j) \leq 0$;

(3) if $\gamma_E(t_i, t_j) < \psi_E(t_i)\psi_E(t_j)$, then $\mathbf{ifd}_M^E(t_i, t_j) \leq 0$, $\mathbf{ifd}_M^E(\bar{t}_i, \bar{t}_j) \leq 0$, $\mathbf{ifd}_M^E(t_i, \bar{t}_j) \geq 0$ and $\mathbf{ifd}_M^E(\bar{t}_i, t_j) \geq 0$.

Theorem 7.3.1 (for the detailed proof see Section 10.4) shows clearly that if we use the estimates given by Eq.(7.15) and Eq.(7.16) then

- a single relation between $\gamma_E(t_i, t_j)$ and $\psi_E(t_i) \cdot \psi_E(t_j)$ can entirely determine all signs of $\mathbf{ifd}_M^E(t_i^{\delta_i}, t_j^{\delta_j})$ for $\delta_i, \delta_j = 1, 0$;
- the signs of $\mathbf{ifd}_M^E(t_i, t_j)$ and $\mathbf{ifd}_M^E(\bar{t}_i, \bar{t}_j)$ are always the same, so are the signs of $\mathbf{ifd}_M^E(t_i, \bar{t}_j)$ and $\mathbf{ifd}_M^E(\bar{t}_i, t_j)$;
- the signs of $\mathbf{ifd}_M^E(t_i, t_j)$ and $\mathbf{ifd}_M^E(\bar{t}_i, \bar{t}_j)$ are always opposite to the signs of $\mathbf{ifd}_M^E(t_i, \bar{t}_j)$ and $\mathbf{ifd}_M^E(\bar{t}_i, t_j)$.

Hereafter, we call $\mathbf{ifd}_M^E(t_i, t_j)$ and $\mathbf{ifd}_M^E(\bar{t}_i, \bar{t}_j)$ the *consistent* mutual information of terms, which means that terms t_i and t_j have identical state values, $\delta_i = \delta_j$, i.e., either they co-occur or neither of them occurs; we call $\mathbf{ifd}_M^E(t_i, \bar{t}_j)$ and $\mathbf{ifd}_M^E(\bar{t}_i, t_j)$ the *inconsistent* mutual information of terms, which means that terms t_i and t_j have opposite state values, $\delta_i \neq \delta_j$, i.e., one of them occurs but another one does not occur.

Particularly, for the estimates of the joint state distribution given by Methods A, B and C, we have the following corollaries, respectively.

Corollary 7.3.1 For the estimates given by Eq.(7.1) and Eq.(7.5), $\mathbf{ifd}_M^d(t_i, t_j) > 0$, $\mathbf{ifd}_M^d(\bar{t}_i, \bar{t}_j) > 0$, $\mathbf{ifd}_M^d(t_i, \bar{t}_j) \leq 0$ and $\mathbf{ifd}_M^d(\bar{t}_i, t_j) \leq 0$ always hold if

$$\sum_{i'<j'; t_{i'}, t_{j'} \in V^d - \{t_j\}} f_d(t_{i'}) f_d(t_{j'}) \geq f_d(t_j) f_d(t_j),$$

$$\sum_{i'<j'; t_{i'}, t_{j'} \in V^d - \{t_i\}} f_d(t_{i'}) f_d(t_{j'}) \geq f_d(t_i) f_d(t_i).$$

Proof. By Corollary 7.2.1, $P_d(\delta_i, \delta_j)$ in Eq.(7.5) is a probability distribution. Also, for arbitrary terms $t_i, t_j \in V^d$,

$$\varpi = \sum_{i'<j'; t_{i'}, t_{j'} \in V^d} f_d(t_{i'}) f_d(t_{j'}) < \sum_{t_{i'}, t_{j'} \in V^d} f_d(t_{i'}) f_d(t_{j'}) = \|d\|^2,$$

from which we have

$$\gamma_d(t_i,t_j)=\frac{f_d(t_i)f_d(t_j)}{\varpi}>\frac{f_d(t_i)}{||d||}\frac{f_d(t_j)}{||d||}=p_d(t_i)p_d(t_j).$$

Take $P_E(\delta_i,\delta_j)=P_d(\delta_i,\delta_j)$, $\gamma_E(t_i,t_j)=\gamma_d(t_i,t_j)$ and $\psi_E(t)=p_d(t)$. Thus, from (1) of Theorem 7.3.1, we can see that four inequalities given in this corollary hold. The proof is complete.

Corollary 7.3.2 For the estimates given by Eq.(7.1) and Eq.(7.9), $\mathbf{ifd}^d_M(t_i,t_j)>0$, $\mathbf{ifd}^d_M(\bar{t}_i,\bar{t}_j)>0$, $\mathbf{ifd}^d_M(t_i,\bar{t}_j)\leq 0$ and $\mathbf{ifd}^d_M(\bar{t}_i,t_j)\leq 0$ always hold.

Proof. $P_d(\delta_i,\delta_j)$ in Eq.(7.9) is a probability distribution. Also, for arbitrary terms $t_i,t_j\in V^d$,

$$||d||-1<||d||,$$

from which we have

$$\gamma_d(t_i,t_j)=\frac{f_d(t_i)}{||d||}\frac{f_d(t_j)}{||d||-1}>\frac{f_d(t_i)}{||d||}\frac{f_d(t_j)}{||d||}=p_d(t_i)p_d(t_j).$$

Take $P_E(\delta_i,\delta_j)=P_d(\delta_i,\delta_j)$, $\gamma_E(t_i,t_j)=\gamma_d(t_i,t_j)$ and $\psi_E(t)=p_d(t)$. Thus, from (1) of Theorem 7.3.1, we can see that four inequalities given in this corollary hold. The proof is complete.

Corollaries 7.3.1 and 7.3.2 tell us that, with Methods A and B, the signs of the consistent mutual information are always positive, and the signs of the inconsistent mutual information are always non-positive. This is because, in this case, relation $\gamma_d(t_i,t_j)>p_d(t_i)p_d(t_j)$ holds for arbitrary terms $t_i,t_j\in V^d$. Consequently, the use of estimates given by Eq.(7.1) and Eq.(7.5), or Eq.(7.1) and Eq.(7.9), asserts that terms co-occurring in some document must be more or less statistically dependent since quantity $\mathbf{ifd}^d_M(t_i,t_j)>0$ supports the dependent hypothesis H_1.

Corollary 7.3.3 For the estimates given by Eq.(7.13) and Eq.(7.14),

(1) if $\frac{F_{\Xi^+}(t_i,t_j)}{|\Xi^+|}>\frac{F_{\Xi^+}(t_i)}{|\Xi^+|}\frac{F_{\Xi^+}(t_j)}{|\Xi^+|}$, then $\mathbf{ifd}^{\Xi^+}_M(t_i,t_j)>0$, $\mathbf{ifd}^{\Xi^+}_M(\bar{t}_i,\bar{t}_j)>0$, $\mathbf{ifd}^{\Xi^+}_M(t_i,\bar{t}_j)\leq 0$ and $\mathbf{ifd}^{\Xi^+}_M(\bar{t}_i,t_j)\leq 0$;

(2) if $\frac{F_{\Xi^+}(t_i,t_j)}{|\Xi^+|}=\frac{F_{\Xi^+}(t_i)}{|\Xi^+|}\frac{F_{\Xi^+}(t_j)}{|\Xi^+|}$, then $\mathbf{ifd}^{\Xi^+}_M(t_i,t_j)=0$, $\mathbf{ifd}^{\Xi^+}_M(\bar{t}_i,\bar{t}_j)=0$, $\mathbf{ifd}^{\Xi^+}_M(t_i,\bar{t}_j)=0$ and $\mathbf{ifd}^{\Xi^+}_M(\bar{t}_i,t_j)=0$;

(3) if $\frac{F_{\Xi^+}(t_i,t_j)}{|\Xi^+|}<\frac{F_{\Xi^+}(t_i)}{|\Xi^+|}\frac{F_{\Xi^+}(t_j)}{|\Xi^+|}$, then $\mathbf{ifd}^{\Xi^+}_M(t_i,t_j)\leq 0$, $\mathbf{ifd}^{\Xi^+}_M(\bar{t}_i,\bar{t}_j)\leq 0$, $\mathbf{ifd}^{\Xi^+}_M(t_i,\bar{t}_j)\geq 0$ and $\mathbf{ifd}^{\Xi^+}_M(\bar{t}_i,t_j)\geq 0$.

Proof. $P_{\Xi^+}(\delta_i,\delta_j)$ in Eq.(7.14) is a probability distribution. Take $P_E(\delta_i,\delta_j)=P_{\Xi^+}(\delta_i,\delta_j)$, $\gamma_E(t_i,t_j)=\gamma_{\Xi^+}(t_i,t_j)=\frac{F_{\Xi^+}(t_i,t_j)}{|\Xi^+|}$ and $\psi_E(t)=\phi_{\Xi^+}(t)=\frac{F_{\Xi^+}(t)}{|\Xi^+|}$. The proof is complete.

Corollary 7.3.3 tells us that, with Method C, the signs of the consistent mutual information are always opposite to the signs of the inconsistent mutual information. However, unlike with Methods A and B, the signs of the consistent mutual information can be positive or negative, similarly for the inconsistent mutual information. This is because Method C does not ensure that relation $\gamma_{\Xi^+}(t_i,t_j)>\phi_{\Xi^+}(t_i)\phi_{\Xi^+}(t_j)$ holds for arbitrary terms $t_i,t_j\in V^{\Xi^+}$. Therefore,

Corollary 7.3.3 cannot directly assert that, if estimates Eq.(7.13) and Eq.(7.14) are used, terms co-occurring in some relevant sample set can be statistically dependent since the sign of $\mathbf{ifd}_M^{\Xi+}(t_i, t_j)$ would be very different from term pair to term pair. We can illustrate this point by an example below.

Example 7.3.1 Suppose $\Xi^+ = \{d_1, d_2, d_3\}$, and $V^{d_1} = \{t_1, t_2, t_3, t_4, t_5\}$, $V^{d_2} = \{t_1, t_4, t_5, t_7\}$, $V^{d_3} = \{t_4, t_7, t_8\}$. Then, we have $F_{\Xi+}(t_1) = 2$, $F_{\Xi+}(t_2) = 1$, $F_{\Xi+}(t_5) = 2$, $F_{\Xi+}(t_7) = 2$, $F_{\Xi+}(t_1, t_2) = 1$, $F_{\Xi+}(t_5, t_7) = 1$. Thus,

$$\frac{3}{9} = \frac{1}{3} = \frac{F_{\Xi+}(t_1, t_2)}{|\Xi^+|} > \frac{F_{\Xi+}(t_1)}{|\Xi^+|}\frac{F_{\Xi+}(t_2)}{|\Xi^+|} = \frac{2}{3}\frac{1}{3} = \frac{2}{9},$$

from which we know that $\mathbf{ifd}_M^{\Xi+}(t_1, t_2) > 0$, $\mathbf{ifd}_M^{\Xi+}(\bar{t}_1, \bar{t}_2) > 0$, $\mathbf{ifd}_M^{\Xi+}(t_1, \bar{t}_2) < 0$, $\mathbf{ifd}_M^{\Xi+}(\bar{t}_1, t_2) < 0$, and that terms t_1 and t_2 are statistically dependent when they co-occur. Also,

$$\frac{3}{9} = \frac{1}{3} = \frac{F_{\Xi+}(t_5, t_7)}{|\Xi^+|} < \frac{F_{\Xi+}(t_5)}{|\Xi^+|}\frac{F_{\Xi+}(t_7)}{|\Xi^+|} = \frac{2}{3}\frac{2}{3} = \frac{4}{9},$$

from which we know that $\mathbf{ifd}_M^{\Xi+}(t_5, t_7) < 0$, $\mathbf{ifd}_M^{\Xi+}(\bar{t}_5, \bar{t}_7) < 0$, $\mathbf{ifd}_M^{\Xi+}(t_5, \bar{t}_7) > 0$, $\mathbf{ifd}_M^{\Xi+}(\bar{t}_5, t_7) > 0$, and that terms t_5 and t_7 are not statistically dependent when they co-occur. ♠

Now, compare the first items of $I_{\Xi+}(\delta_i, \delta_j)$ and $emim_{\Xi+}(\delta_i, \delta_j)$. By Corollary 7.3.3, we know that from a single relation between $\gamma_{\Xi+}(t_i, t_j) = \frac{F_{\Xi+}(t_i, t_j)}{|\Xi^+|}$ and $\phi_{\Xi+}(t_i)\phi_{\Xi+}(t_j) = \frac{F_{\Xi+}(t_i)}{|\Xi^+|}\frac{F_{\Xi+}(t_j)}{|\Xi^+|}$ we can infer all the signs of the discrimination measures $\mathbf{ifd}_M^{\Xi+}(t_i^{\delta_i}, t_j^{\delta_j})$ for $\delta_i, \delta_j = 1, 0$, and then determine whether term pair (t_i, t_j) is statistically dependent under its individual state values. However, the inference and determination cannot be made from the relation between $n_{11} = F_{\Xi+}(t_i, t_j)$ and $n_{1\cdot}n_{\cdot 1} = F_{\Xi+}(t_i)F_{\Xi+}(t_j)$. In fact, by Theorem 7.2.4, we know that the individual items of $emim_{\Xi+}(\delta_i, \delta_j)$ are always non-positive (see Example C given in Section 10.6).

7.4 On Dependence of Terms

Based on the ideas developed in the last section, we can gain insight into the concept of the dependence of terms.

7.4.1 Dependence in Broad and Narrow Senses

Suppose $t_i, t_j \in V^E$. Now imagine that you find that t_j is a very important term, and that $I_E(\delta_i, \delta_j)$ obtains a rather high value. Now ask yourself the question: 'Is term t_i the one that I am certainly interested in?'.

If you try to give an answer 'Yes', think about the kind of case where E is a document d, and $\mathbf{ifd}_M^d(t_i, t_j) \leq 0$, $\mathbf{ifd}_M^d(t_i, \bar{t}_j) > 0$, $\mathbf{ifd}_M^d(\bar{t}_i, t_j) > 0$, $\mathbf{ifd}_M^d(\bar{t}_i, \bar{t}_j) \leq 0$, and another question: 'What do these discrimination measures tell us substantively?'.

The answer to the second question is clear: the *positive* value $I_d(\delta_i, \delta_j)$ would be dominated by the positive quantities $\mathbf{ifd}_M^d(t_i, \bar{t}_j)$ and/or $\mathbf{ifd}_M^d(\bar{t}_i, t_j)$. The higher value the measure $I_d(\delta_i, \delta_j)$ has, the larger quantities the measures $\mathbf{ifd}_M^d(t_i, \bar{t}_j)$ and/or $\mathbf{ifd}_M^d(\bar{t}_i, t_j)$ provide, and the more they indicate that t_i and t_j should not co-occur in d.

Consider a further situation where t_j is a unique term of the query and d is a relevant document. Then, in the above case, a higher value $I_d(\delta_i, \delta_j)$ indicates that the occurrence of query term t_j in document d should accompany the absence of term t_i from d. Thus, in order to better match the relevant document with the query, the selection of term t_i as an expansion term for query expansion would not be what we desire. The answer of the first question is now apparent.

In IR, a higher value of $I_E(\delta_i, \delta_j)$ would not imply that term t_i is the one that we are surely interested in. In other words, a term very 'dependent' on term t_j might not mean that it is a good one with respect to the query (suppose $V^q = \{t_j\}$).

It is certainly true that each of the discrimination measures $\mathbf{ifd}_M^E(t_i, t_j)$, $\mathbf{ifd}_M^E(t_i, \bar{t}_j)$, $\mathbf{ifd}_M^E(\bar{t}_i, t_j)$ and $\mathbf{ifd}_M^E(\bar{t}_i, \bar{t}_j)$ can be used to measure the extent of dependence of terms t_i and t_j. Also, it is certainly true that the larger the quantities the measures offer, the higher the extent term t_i is statistically dependent on term t_j (and vice versa). However, the *implications* of the dependence for the individual measures are very different. Remember that we always emphasize 'the dependence *under the state value* (δ_i, δ_j)'. This emphasis is necessary because it clearly indicates that it is the state value (δ_i, δ_j) that supports the dependence. Such kind of dependence we can call *dependence in a broad sense.*

In IR, we generally agree to concentrate our attention on the statistical data of the co-occurrence of terms. Thus, the dependence of terms of which we usually speak is the one when terms *co-occur*, i.e., $(\delta_i, \delta_j) = (1, 1)$, which we can call *dependence in a narrow sense.* Example 7.3.1 is a nice illustration of our viewpoint.

We are indeed interested in dependence in the narrow sense rather than in the broad sense. If we are given another case where E is a relevant document d and t_j is a unique query term, but this time $\mathbf{ifd}_M^d(t_i, t_j) > 0$, $\mathbf{ifd}_M^d(t_i, \bar{t}_j) \leq 0$, $\mathbf{ifd}_M^d(\bar{t}_i, t_j) \leq 0$, $\mathbf{ifd}_M^d(\bar{t}_i, \bar{t}_j) > 0$, then, for a higher value $I_d(\delta_i, \delta_j)$, we would like to answer the first question: 'Yes, definitely!' (How about $\mathbf{ifd}_M^d(\bar{t}_i, \bar{t}_j) > 0$? See Section 7.5).

Consequently, the dependence of terms can be given by the mutual information of terms under some state value, by the consistent/inconsistent mutual information of terms, by the expected mutual information of terms. The different implications of the dependence should be carefully distinguished from one another.

7.4.2 Global and Local Dependence

We often need to find the values of dependence (of terms) that some dependence measure $\zeta(t_i^{\delta_i}, t_j^{\delta_j})$, such as $\mathbf{ifd}_M^E(t_i^{\delta_i}, t_j^{\delta_j})$, can take when variable (t_i, t_j) is restricted to lie in a certain domain.

For a given document collection D, we call $\zeta(t_i^{\delta_i}, t_j^{\delta_i})$ the *global dependence* of t_i and t_j under the state value (δ_i, δ_j) if measure $\zeta(t_i^{\delta_i}, t_j^{\delta_j})$ is derived from the statistical data within D. The global dependence that $\zeta(t_i^{\delta_i}, t_j^{\delta_j})$ takes on the vocabulary $V^D = V$ is precisely what it says, that is, it is all the possible values the measure takes while (t_i, t_j) varies throughout domain $V \times V$.

For a given document set (i.e., information entity) $\Xi^+ \subseteq D$, we call $\zeta(t_i^{\delta_i}, t_j^{\delta_j})$ the *local dependence* of t_i and t_j under the state value (δ_i, δ_j) if measure $\zeta(t_i^{\delta_i}, t_j^{\delta_j})$ is derived from the statistical data within Ξ^+. Again, the local dependence on the set V^{Ξ^+} is precisely all the possible values the measure takes as (t_i, t_j) is allowed to vary throughout domain $V^{\Xi^+} \times V^{\Xi^+}$.

The estimation Method C derives its importance from the fact that its simplicity of computation enables us to have an insight into the dependence of terms. Thus, from the example below, which shows a typical situation, you can see that the local dependence of a certain term pair need not to be equal to the global dependence of this term pair, and need not be the same as another local dependence of this term pair.

Example 7.4.1 Let us return to Example 7.3.1. We now further suppose $D = \{d_1, d_2, ..., d_{10000}\}$. For term t_1 and t_4, we have $F_{\Xi^+}(t_1) = 2$, $F_{\Xi^+}(t_4) = |\Xi^+| = 3$, $F_{\Xi^+}(t_1, t_4) = 2$. Thus, from Theorem 7.2.5, we obtain immediately

$$\begin{aligned} I_{\Xi^+}(\delta_1, \delta_4) &= \sum_{\delta_1, \delta_4 = 1,0} \mathbf{ifd}_M^{\Xi^+}(t_1^{\delta_1}, t_4^{\delta_4}) \\ &= 0.0000 - 0.0000 - 0.0000 + 0.0000 = 0.0000. \end{aligned}$$

Now, we fix $F_{\Xi^+}(t_1) = 2$, $F_{\Xi^+}(t_4) = |\Xi^+| = 3$, $F_{\Xi^+}(t_1, t_4) = 2$, but this time, take $|\Xi^+| = 10$. Then, we find

$$\begin{aligned} I_{\Xi^+}(\delta_1, \delta_4) &= \frac{2}{10} \log \frac{\frac{2}{10}}{\frac{2}{10}\frac{3}{10}} + \frac{2-2}{10} \log \frac{\frac{2-2}{10}}{\frac{2}{10}(1-\frac{3}{10})} \\ &\quad + \frac{3-2}{10} \log \frac{\frac{3-2}{10}}{(1-\frac{2}{10})\frac{3}{10}} + \frac{10-2-3+2}{10} \log \frac{\frac{10-2-3+2}{10}}{(1-\frac{2}{10})(1-\frac{3}{10})} \\ &= \frac{2}{10} \log \frac{10}{3} + 0 \log 0 + \frac{1}{10} \log \frac{10}{24} + \frac{7}{10} \log \frac{10}{8} \\ &\approx 0.2408 - 0.0000 - 0.0875 + 0.1562 = 0.3095. \end{aligned}$$

More dependence values for a variety of sizes of set Ξ^+ are computed and results are listed in the table below.

$\lvert\Xi^+\rvert$	$\mathbf{ifd}_M^{\Xi^+}(t_1, t_4)$	$\mathbf{ifd}_M^{\Xi^+}(t_1, \bar{t}_4)$	$\mathbf{ifd}_M^{\Xi^+}(\bar{t}_1, t_4)$	$\mathbf{ifd}_M^{\Xi^+}(\bar{t}_1, \bar{t}_4)$	$I_{\Xi^+}(\delta_1, \delta_4)$
3	0.0000	0.0000	0.0000	0.0000	0.0000
4	0.1438	0.0000	-0.1014	0.1733	0.2157
5	0.2043	0.0000	<u>-0.1176</u>	<u>0.2043</u>	0.2910
6	0.2310	0.0000	-0.1155	0.2027	0.3182
7	0.2421	0.0000	-0.1089	0.1923	<u>0.3255</u>
8	<u>0.2452</u>	0.0000	-0.1014	0.1798	0.3236
9	0.2441	0.0000	-0.0941	0.1675	0.3175
10	0.2408	0.0000	-0.0875	0.1562	0.3095
15	0.2146	0.0000	-0.0637	0.1145	0.2654
20	0.1897	0.0000	-0.0497	0.0896	0.2296
30	0.1535	0.0000	-0.0343	0.0621	0.1813
50	0.1125	0.0000	-0.0212	0.0384	0.1297
100	0.0701	0.0000	-0.0108	0.0196	0.0789
1000	0.0116	0.0000	-0.0011	0.0020	0.0125
10000	0.0016	0.0000	-0.0001	0.0002	0.0017

$F_{\Xi^+}(t_1) = 2$, $F_{\Xi^+}(t_4) = 3$, $F_{\Xi^+}(t_1, t_4) = 2$

in which, the numbers at the last row are for the global dependence of terms t_1 and t_4 (where $|\Xi^+| = |D| = 10000$), and the numbers underlined are the maximum (in absolute values) for the corresponding measures. ♠

Clearly we are using the above table of (mutual) information quantities to tell us about the behaviour of the individual discrimination measures. There are five different measures that can give us useful information. Each tells us different aspects about the dependences of terms, and so should be given the correct interpretation. This information table can be called the *broad dependence measure table* of terms t_1 and t_4.

Not surprisingly, from the above table, we see that the dependence values (i.e., amount of mutual information) vary as the size of set Ξ^+. As we know, the sample sets with the different sizes should be thought of as different entities, and usually they provide rather different statistical data. Thus, the state distributions, further, the discrimination measures, are very inconsistent from sample set to sample set. Therefore, the dependence of terms is *in reference to* entity Ξ^+, that is, it is a local concept.

Let us now carefully examine the information table above to look at what insight it can give. First, when $|\Xi^+| = 3$, we have $F_{\Xi^+}(t_4) = |\Xi^+|$, i.e., term t_4 occurs in all documents in Ξ^+. In this case, the occurrence of term t_4 would not provide any information about the occurrence of term t_1 in documents in Ξ^+. Thus, these two terms are independent of each other, and $\mathbf{ifd}_M^{\Xi^+}(t_1^{\delta_1}, t_4^{\delta_4}) = 0$ for $\delta_1, \delta_4 = 1, 0$, so $I_{\Xi^+}(\delta_1, \delta_4) = 0$.

Next, we see the individual dependence values in each of the columns are increasing (in absolute values) as the increase in the size of Ξ^+. This is because if terms t_1 and t_4 occur in few of the documents in Ξ^+, and also co-occur in some of these, then it should indicate that these two terms are dependent.

Intuitively, if the size of Ξ^+ is larger, while terms t_1 and t_4 occur only in few of the documents in Ξ^+ (i.e., do not occur in most of the documents in Ξ^+), and meanwhile they co-occur in some of the few documents, then these two terms should be very dependent. That is, the dependence values should become greater as the size of Ξ^+ increases (when $F_{\Xi^+}(t_1)$, $F_{\Xi^+}(t_4)$, $F_{\Xi^+}(t_1, t_4)$ are fixed). However, to our surprise, from the information table above, we see that the dependence values drop greatly when, for instance, $|\Xi^+| = 100$ (again, in absolute values), and almost are equal to zero when $|\Xi^+| = 10000$.

More generally, when $F_{\Xi^+}(t_i, t_j)$, $F_{\Xi^+}(t_i)$ and $F_{\Xi^+}(t_j)$ are fixed, we can find:

$$
\begin{aligned}
I_{\Xi^+}(\delta_i, \delta_j) = & \frac{F_{\Xi^+}(t_i, t_j)}{|\Xi^+|} \log \frac{F_{\Xi^+}(t_i, t_j)}{F_{\Xi^+}(t_i) F_{\Xi^+}(t_j)} |\Xi^+| \\
& + \frac{F_{\Xi^+}(t_i) - F_{\Xi^+}(t_i, t_j)}{|\Xi^+|} \log \frac{F_{\Xi^+}(t_i) - F_{\Xi^+}(t_i, t_j)}{F_{\Xi^+}(t_i)\big(|\Xi^+| - F_{\Xi^+}(t_j)\big)} |\Xi^+| \\
& + \frac{F_{\Xi^+}(t_j) - F_{\Xi^+}(t_i, t_j)}{|\Xi^+|} \log \frac{F_{\Xi^+}(t_j) - F_{\Xi^+}(t_i, t_j)}{\big(|\Xi^+| - F_{\Xi^+}(t_i)\big) F_{\Xi^+}(t_j)} |\Xi^+| \\
& + \frac{|\Xi^+| - F_{\Xi^+}(t_i) - F_{\Xi^+}(t_j) + F_{\Xi^+}(t_i, t_j)}{|\Xi^+|} \times \\
& \quad \log \frac{|\Xi^+| - F_{\Xi^+}(t_i) - F_{\Xi^+}(t_j) + F_{\Xi^+}(t_i, t_j)}{\big(|\Xi^+| - F_{\Xi^+}(t_i)\big)\big(|\Xi^+| - F_{\Xi^+}(t_j)\big)} |\Xi^+| \\
\to & \; 0.0000 - 0.0000 - 0.0000 + 0.0000 = 0.0000 \quad (\text{when } |\Xi^+| \to \infty).
\end{aligned}
$$

The mathematical reason for this is that factor $\frac{1}{|\Xi^+|}$, in the individual measures $\mathbf{ifd}_M^{\Xi^+}(t_i^{\delta_i}, t_j^{\delta_j})$, approaches to zero as $|\Xi^+|$ becomes large, and the last measure $\mathbf{ifd}_M^{\Xi^+}(\bar{t}_i, \bar{t}_j)$ approaches to $1 \times \log 1 = 0$, as $|\Xi^+| \to \infty$. It is clear that, for those term pairs whose statistical data $F_{\Xi^+}(t_i)$, $F_{\Xi^+}(t_j)$ and $F_{\Xi^+}(t_i, t_j)$ are relatively smaller, too large a number $|\Xi^+|$ would overwhelm the

more important statistical information, and thereby weaken and dilute the potential capability of discrimination measures derived from the statistical information. It is not clear at present how to determine an appropriate size of the sample set against a set of term pairs. So far, there is no research about this interesting issue in IR.

7.5 Association Functions

One of the applications of the mutual information of terms, that we shall study in this section, is to establish the concepts of mutual association in the context of IR. The mutual association of a term with another term, with the relevant sample set, or with the query, have long been interesting subjects for query expansion. Thus, this section discusses three concepts: *term-based* association, *set-based* association and *query-based* association, all based on the interpretations and estimation of the mutual information of terms as provided in the previous sections.

In the following two sections, we will consider individual documents as the information entities, and use only the consistent mutual information, which are estimated by using Method A and B. Also, for a given term pair (t_i, t_j), the state value δ_j of term t_j is required to follow the state value δ_i of term t_i, i.e., δ_j is given by δ_i. Therefore, for terms $t_i, t_j \in V^d$, the general expression $\mathbf{ifd}_M^E(t_i^{\delta_i}, t_j^{\delta_i})$ in Eq.(7.18) is specialized as

$$\begin{aligned} \mathbf{ifd}_M^d(t_i, t_j) &= \gamma_d(t_i, t_j) \log \frac{\gamma_d(t_i, t_j)}{p_d(t_i) p_d(t_j)}, \\ \mathbf{ifd}_M^d(\bar{t}_i, \bar{t}_j) &= \big(1 - p_d(t_i) - p_d(t_j) + \gamma_d(t_i, t_j)\big) \log \frac{1 - p_d(t_i) - p_d(t_j) + \gamma_d(t_i, t_j)}{\big(1 - p_d(t_i)\big)\big(1 - p_d(t_j)\big)}, \end{aligned} \tag{7.19}$$

where $\gamma_d(t_i, t_j)$ is given in Eq.(7.5) or Eq.(7.9), and $p_d(t)$ is given in Eq.(7.1).

In addition, all concepts discussed in this section are related to the relevant sample set Ξ^+. Thus, hereafter, we will always assume that $\Xi^+ \neq \emptyset$ and $|V^{\Xi^+}| \geq 2$, and that Ξ^+ is effective, that is, all important relevant information to the query can be contained in Ξ^+.

7.5.1 Term-Based Association $att_M(t_i^{\delta_i}, t_j^{\delta_i})$

In order to estimate the mutual association of terms with the context of the query, it is reasonable for us to think of drawing some pieces of 'useful information' from set Ξ^+. For a given term $t_i \in V^{\Xi^+}$, a piece of useful information in the current method would be: how term t_i is associated with another term $t_j \in V^{\Xi^+} - \{t_i\}$. A term-based association function $att_M(t_i, t_j)$ is thus introduced to measure the quantity of a piece of useful information.

More precisely, in the foregoing, measure $\mathbf{ifd}_M^d(t_i, t_j)$ was the mutual information of terms t_i and t_j concerning document d when they co-occur in d. Thus, we now consider all of documents in Ξ^+, check them one by one, and sum $\mathbf{ifd}_M^d(t_i, t_j)$ for those documents in which terms t_i and t_j co-occur. We can formulate this idea by a function:

$$att_M(t_i, t_j) = \frac{1}{|\Xi^+|} \sum_{d \in \Xi^+} \mathbf{ifd}_M^d(t_i, t_j),$$

which can be regarded as the (*average*) *mutual association* of terms t_i and t_j concerning set Ξ^+ when the proposition (t_i, t_j) is true in some of the documents in Ξ^+, where $\frac{1}{|\Xi^+|}$ is a normalization factor for the size of set Ξ^+.

More generally, a definition of the mutual association of terms may be made, in which only the consistent state values are considered. We are thus led to the following more formal definition.

Definition 7.5.1 For a given term $t_i \in V^{\Xi^+}$, the mutual association of term t_i with another term $t_j \in V^{\Xi^+} - \{t_i\}$ under the state values (δ_i, δ_i) is defined as

$$att_M(t_i^{\delta_i}, t_j^{\delta_i}) = \frac{1}{|\Xi^+|} \sum_{d \in \Xi^+} \mathbf{ifd}_M^d(t_i^{\delta_i}, t_j^{\delta_i}),$$

which is called the *term-based association* of term t_i with term t_j.

Obviously, when terms t_i and t_j do not co-occur in any document in Ξ^+, we have $att_M(t_i^{\delta_i}, t_j^{\delta_i}) = 0$ since, by Definition 7.3.2, $\mathbf{ifd}_M^d(t_i^{\delta_i}, t_j^{\delta_i}) = 0$ for every $d \in \Xi^+$.

Notice that it is important to be careful about notation here. It is easy to confuse $\mathbf{ifd}_M^d(t_i^{\delta_i}, t_j^{\delta_i})$ and $\mathbf{ifd}_M^d(t_i^{\delta_i}, t_j^{\delta_j})$ in the above definition. Again, term t_j is imposed upon the same state value δ_i as term t_i has.

To gain an understanding of term-based association of terms, let us see the simple example below.

Example 7.5.1 Let us return to Example 7.3.1. Recall that $V^{d_1} = \{t_1, t_2, t_3, t_4, t_5\}$, $V^{d_2} = \{t_1, t_4, t_5, t_7\}$, $V^{d_3} = \{t_4, t_7, t_8\}$. Thus, $V^{\Xi^+} = \{t_1, t_2, t_3, t_4, t_5, t_7, t_8\}$, and

$$att_M(t_1, t_2) = \frac{1}{3}\mathbf{ifd}_M^{d_1}(t_1, t_2), \qquad att_M(t_5, t_4) = \frac{1}{3}\left[\mathbf{ifd}_M^{d_1}(t_5, t_4) + \mathbf{ifd}_M^{d_2}(t_5, t_4)\right],$$

$$att_M(t_1, t_8) = 0, \qquad att_M(t_1, t_6) - \text{does not exist since } t_6 \notin V^{\Xi^+},$$

and so on. ♠

7.5.2 Set-Based Association $ats_M(t_i^{\delta_i}, \Xi^+)$

In the foregoing we have discussed the concept of term-based association. However, one may desire to consider the 'overall' association of term t_i with the whole set Ξ^+. A set-based association function $ats_M(t_i, \Xi^+)$ is introduced here to achieve this.

If we accept the assumption (stated rather informally) that the statement 'the mutual association of terms t_i and t_j concerning set Ξ^+' is equivalent to 'the mutual association of term t_i with set Ξ^+ by means of term t_j', then, for a given term $t_i \in V^{\Xi^+}$, quantity $att_M(t_i, t_j)$ actually provides a piece of association information of term t_i with set Ξ^+ by means of a single term t_j when t_i and t_j co-occur in some documents in Ξ^+.

Function $ats_M(t_i, \Xi^+)$ is designed by extending the term-based association function, i.e., by considering the summation of quantities $att_M(t_i, t_j)$ for all terms $t_j \in V^{\Xi^+} - \{t_i\}$. A more formal definition, which involves the consistent state values, is given as follows.

Definition 7.5.2. For a given term $t_i \in V^{\Xi^+}$, the mutual association of term t_i with set Ξ^+ under the state value (δ_i, δ_i) is defined as

$$ats_M(t_i^{\delta_i}, \Xi^+) = \sum_{t_j \in V^{\Xi^+} - \{t_i\}} att_M(t_i^{\delta_i}, t_j^{\delta_i}),$$

which is called the *set-based association* of term t_i with set Ξ^+.

It is easily found that function $ats_M(t_i^{\delta_i}, \Xi^+)$ is simply a pooling of all possible pieces of the association information $att_M(t_i^{\delta_i}, t_j^{\delta_i})$, satisfying $t_j \neq t_i$, together. The following example illustrates the idea involved.

Example 7.5.2. Let us return to Example 7.4.1. Recall that $V^{\Xi^+} = \{t_1, t_2, t_3, t_4, t_5, t_7, t_8\}$. So we can give the table below, in which, d_m $(m = 1, 2, 3)$ in each cell expresses the fact that terms t_i and t_j co-occur in document d_m, whereas an empty cell expresses that there is no document (in Ξ^+) in which terms t_i and t_j co-occur.

Table 7.5.1 Documents in which t_i and t_j $(\in V^{\Xi^+})$ co-occur

$t_i \backslash t_j$	t_1	t_2	t_3	t_4	t_5	t_7	t_8
t_1	-	d_1	d_1	d_1, d_2	d_1, d_2	d_2	
t_2	d_1	-	d_1	d_1	d_1		
t_3	d_1	d_1	-	d_1	d_1		
t_4	d_1, d_2	d_1	d_1	-	d_1, d_2	d_2, d_3	d_3
t_5	d_1, d_2	d_1	d_1	d_1, d_2	-	d_2	
t_7	d_2			d_2, d_3	d_2	-	d_3
t_8				d_3		d_3	-

Then, for instance, we have,

$$
\begin{aligned}
ats_M(t_1, \Xi^+) &= \sum_{t_j \in V^{\Xi^+} - \{t_1\}} att_M(t_1, t_j) \\
&= \frac{1}{3} \left[\mathbf{ifd}_M^{d_1}(t_1, t_2) + \mathbf{ifd}_M^{d_1}(t_1, t_3) + \mathbf{ifd}_M^{d_1}(t_1, t_4) + \mathbf{ifd}_M^{d_2}(t_1, t_4) + \right. \\
&\qquad \left. \mathbf{ifd}_M^{d_1}(t_1, t_5) + \mathbf{ifd}_M^{d_2}(t_1, t_5) + \mathbf{ifd}_M^{d_2}(t_1, t_7) \right], \\
ats_M(t_2, \Xi^+) &= \sum_{t_j \in V^{\Xi^+} - \{t_2\}} att_M(t_2, t_j) \\
&= \frac{1}{3} \left[\mathbf{ifd}_M^{d_1}(t_2, t_1) + \mathbf{ifd}_M^{d_1}(t_2, t_3) + \mathbf{ifd}_M^{d_1}(t_2, t_4) + \mathbf{ifd}_M^{d_1}(t_2, t_5) \right], \\
ats_M(t_8, \Xi^+) &= \sum_{t_j \in V^{\Xi^+} - \{t_8\}} att_M(t_8, t_j) \\
&= \frac{1}{3} \left[\mathbf{ifd}_M^{d_3}(t_8, t_4) + \mathbf{ifd}_M^{d_3}(t_8, t_7) \right]. \quad \spadesuit
\end{aligned}
$$

Generally, we have the following theorem:

Theorem 7.5.1 For a given term $t_i \in V^{\Xi^+}$, we have

$$
ats_M(t_i^{\delta_i}, \Xi^+) = \frac{1}{|\Xi^+|} \sum_{t_j \in V^d - \{t_i\}; d \in \Xi^+} \mathbf{ifd}_M^d(t_i^{\delta_i}, t_j^{\delta_i}).
$$

From the above theorem (for the proof see Section 10.4), it follows that we can give up writing $t_j \in V^{\Xi^+} - \{t_i\}; d \in \Xi^+$, and write $t_j \in V^d - \{t_i\}; d \in \Xi^+$ instead. Obviously, the latter is clearer for describing the domain over which the sum can be carried out than the former.

7.5.3 Query-Based Association $atq_M(t_i^{\delta_i}, q)$

We have discussed the concepts of term-based and set-based association; both of them are considered without directly involving the query. Now, we come to the heart of this section — defining the concept of query-based association.

Under the hypothesis that query terms in set $V^q \cap V^{\Xi^+}$ are good ones with respect to the query itself, our aim is to judge other good terms from set $V^{\Xi^+} - V^q = V^{\Xi^+} - (V^{\Xi^+} \cap V^q)$. Thus, for the first component of term pairs, we limit terms considered to those that occur in at least one relevant sample document but are not query terms, i.e., $t_i \in V^{\Xi^+} - V^q$.

Next, what should we think about the limitation of the second component of term pairs? As mentioned, terms that occur in some document would more or less contain information related to the document itself. Because the documents may be long, and include different information content, many of them may not be very relevant to the query. Thus, it is unlikely that every term $t_j \in V^{\Xi^+} - \{t_i\}$ would be related to the query. It is clear that an effective score function should be constructed by combining only those pieces of association information $att_M(t_i, t_j)$, in which, terms t_j are related to the query. In other words, we should not simply apply function $ats_M(t_i, \Xi^+)$, the mutual association of term t_i with set Ξ^+, to estimate the mutual association of term t_i with the query. Instead, we can adopt a conservative but more effective way that considers only those terms occurring in the query, i.e., $t_j \in V^q \cap V^{\Xi^+} \subseteq V^{\Xi^+} - \{t_i\}$, since we certainly know that those terms are good ones under the hypothesis. In a word, the extent of the mutual association of term t_i with the query should be measured by pooling 'valuable' pieces of association information by means of good query terms $t_j \in V^q \cap V^{\Xi^+}$, rather than all terms $t_j \in V^{\Xi^+} - \{t_i\}$.

The above paragraph may give you an idea: what do we mean by the statement that term t_i has query-based association with a given query? The concept of term-based association derives its importance from the fact that it provides a means to define the concept of query-based association. We can thus construct a query-based association function $atq_M(t_i, q)$, which computes the summation of $att_M(t_i, t_j)$ for all query terms satisfying $t_j \in V^q \cap V^{\Xi^+}$. We therefore make the following formal definition which considers only the consistent state values.

Definition 7.5.3 For a given term $t_i \in V^{\Xi^+} - V^q$, the mutual association of term t_i with query q under the state value (δ_i, δ_i) is defined as

$$atq_M(t_i^{\delta_i}, q) = \sum_{t_j \in V^q \cap V^{\Xi^+}} \varrho(t_j)\ att_M(t_i^{\delta_i}, t_j^{\delta_i}),$$

which is called the *query-based association* of term t_i with query q, where scale factor $\varrho(t_j) \geq 0$ measures the significance of term t_j in representing query q.

We clarify the meaning of the foregoing idea by considering the example below.

Example 7.5.3 (Example 7.5.1 continued). Consider our example of the term-based association. We now further assume $V^q = \{t_2, t_5, t_6\}$. Notice that $t_j \in V^q \cap V^{\Xi^+} = \{t_2, t_5\}$. So we have Table 7.5.2, in which, the terms in parentheses are query terms $t_j \in V^q \cap V^{\Xi^+}$.

Then, for instance, we obtain

$$atq_M(t_1, q) = \sum_{t_j \in V^q \cap V^{\Xi^+}} \varrho(t_j)\ att_M(t_1, t_j)$$

Table 7.5.2 Documents in which t_i and t_j ($\in V^q$) co-occur

$t_i \backslash t_j$	t_1	(t_2)	t_3	t_4	(t_5)	t_7	t_8
t_1	-	d_1			d_1, d_2		
(t_2)		-					
t_3		d_1	-		d_1		
t_4		d_1		-	d_1, d_2		
(t_5)					-		
t_7					d_2	-	
t_8							-

$$= \frac{1}{3} \left[\varrho(t_2)\ \mathbf{ifd}_M^{d_1}(t_1, t_2) + \varrho(t_5)\ \mathbf{ifd}_M^{d_1}(t_1, t_5) + \varrho(t_5)\ \mathbf{ifd}_M^{d_2}(t_1, t_5) \right],$$

$$atq_M(t_3, q) = \sum_{t_j \in V^q \cap V^{\Xi^+}} \varrho(t_j)\ att_M(t_3, t_j)$$

$$= \frac{1}{3} \left[\varrho(t_2)\ \mathbf{ifd}_M^{d_1}(t_3, t_2) + \varrho(t_5)\ \mathbf{ifd}_M^{d_1}(t_3, t_5) \right],$$

$$atq_M(t_8, q) = \sum_{t_j \in V^q \cap V^{\Xi^+}} \varrho(t_j)\ att_M(t_8, t_j) = 0. \quad \spadesuit$$

There are two main differences between the set-based and query-based associations: (a) For $ats_M(t_i^{\delta_i}, \Xi^+)$, the domains of terms are given without any limitation, i.e., $t_i, t_j \in V^{\Xi^+}$ (where $t_j \neq t_i$); whereas for $atq_M(t_i^{\delta_i}, q)$, the domains are limited to $t_i \in V^{\Xi^+} - V^q$ and $t_j \in V^q \cap V^{\Xi^+}$. (b) For $ats_M(t_i^{\delta_i}, \Xi^+)$, terms $t_j \in \Xi^+ - \{t_i\}$ are treated as equally important, and assigned the same weight 1; whereas for $atq_M(t_i^{\delta_i}, q)$, terms $t_j \in V^q \cap V^{\Xi^+}$ are assigned higher (generally unequal) weights $\varrho(t_j) \geq 0$, and terms $t_j \in \Xi^+ - V^q$ are assigned weight 0 and thus thrown away. These two differences allow the computation of the query-based association to be reduced greatly. From Table 7.5.2, we can clearly see that the non-empty cells in the table above are far fewer than the non-empty cells in the table of Table 7.5.1. Thus, the total computation of the mutual association of terms with the query should not be excessive even though the size of V^{Ξ^+} is larger.

Similar to the set-based association, we have the corresponding theorem for the query-based association below.

Theorem 7.5.2 For a given term $t_i \in V^{\Xi^+} - V^q$, we have

$$atq_M(t_i^{\delta_i}, q) = \frac{1}{|\Xi^+|} \sum_{t_j \in V^q \cap V^d; d \in \Xi^+} \varrho(t_j)\ \mathbf{ifd}_M^d(t_i^{\delta_i}, t_j^{\delta_i}). \tag{7.20}$$

The proof is similar to the one given in Theorem 7.5.1.

Up to now we have formally introduced three concepts of the mutual association. We shall see that with these concepts the association score functions given in the next section are intuitive and simple.

7.6 Score Functions

This section proposes two score functions for judging good terms. The functions are constructed based on the concept of query-based association given in the last section.

From Theorem 7.5.2 we can see that individual items of function $atq_M(t_i^{\delta_i}, q)$ are summed over a combined domain. From the definition of measure $\mathbf{ifd}_M^d(t_i^{\delta_i}, t_j^{\delta_i})$, the combined domain, more clearly, should be written as $t_i, t_j \in V^d; t_j \in V^q \cap V^d; d \in \Xi^+$. In order to have an insight into the combined domain and its nature, we need to introduce a further piece of notation — we need define the notion of association set.

7.6.1 Association Set

For a given term $t_i \in V^{\Xi^+} - V^q$, to calculate its query-based association, a key point is to find all possible term pairs (t_i, t_j), each of which satisfies the condition that t_i and t_j co-occur in $d \in \Xi^+$, and that t_j also occurs in q. Let $(t_i, t_j)_d$ represent $t_i, t_j \in V^d$, we thus can introduce the following definition.

Definition 7.6.1 Given a query q and its relevant sample set Ξ^+, for a given term $t_i \in V^{\Xi^+} - V^q$, define

$$\mho_{t_i}^{\Xi^+} = \{(t_i, t_j)_d \mid t_j \in V^q \text{ and } d \in \Xi^+\},$$

which is called the *association set* of term t_i with query q concerning set Ξ^+.

According to the above definition, the following aspects are obvious but important:

- From the notation of $(t_i, t_j)_d$, we have

$$\begin{aligned} \mho_{t_i}^{\Xi^+} &= \{(t_i, t_j)_d \mid t_j \in V^q \text{ and } t_j \in V^d \text{ and } d \in \Xi^+\} \\ &= \cup_{t_j \in V^q \cap V^d; d \in \Xi^+} \{(t_i, t_j)_d\}. \end{aligned} \tag{7.21}$$

- For every element $(t_i, t_j)_d \in \mho_{t_i}^{\Xi^+}$, its first component is the given term t_i under consideration; its second component t_j is always some query term satisfying $t_j \in V^q \cap V^d \subseteq V^q \cap V^{\Xi^+}$.

- Since $t_i \in V^{\Xi^+} - V^q$ and $t_j \in V^q \cap V^{\Xi^+}$, term t_i under consideration will never be the same as term t_j, and never be a query term.

- $\mho_{t_i}^{\Xi^+} = \emptyset$ if term t_i and none of the query terms $t_j \in V^q \cap V^{\Xi^+}$ co-occur in any documents in Ξ^+.

- $\mho_{t_i}^{\Xi^+}$ does not exist if term $t_i \notin V^{\Xi^+} - V^q$.

- For a given collection D and a given scheme of the representation of documents over D, $\mho_{t_i}^{\Xi^+}$ is uniquely determined by term t_i, query q and set Ξ^+.

It can be seen that the notion of association set is useful in describing a mutual association phenomenon of term t_i with query q. It is also helpful for computation: for a given query q and its Ξ^+, set $\mho_{t_i}^{\Xi^+}$ is exactly the domain over which the individual items of function

$atq_M(t_i, q)$ are summed. In fact, comparing Eq.(7.20) with Eq.(7.21), it can be immediately written down

$$atq_M(t_i^{\delta_i}, q) = \frac{1}{|\Xi^+|} \sum_{(t_i,t_j)_d \in \mho_{t_i}^{\Xi^+}} \varrho(t_j)\mathbf{ifd}_M^d(t_i^{\delta_i}, t_j^{\delta_i}). \tag{7.22}$$

It avoids reference to the combined domain $t_i, t_j \in V^d; t_j \in V^q \cap V^d; d \in \Xi^+$, and instead speaks of a single domain $(t_i, t_j)_d \in \mho_{t_i}^{\Xi^+}$.

Notice that the size of the association set can be easily expressed as

$$|\mho_{t_i}^{\Xi^+}| = \sum_{d \in \Xi^+} |V^q \cap V^d|,$$

i.e., it is the sum of the numbers of query terms in the individual documents over $d \in \Xi^+$. In particular, we have the following three special cases:

- If $|\Xi^+| = |\{d\}| = 1$, then $\mho_{t_i}^{\Xi^+} = \mho_{t_i}^{\{d\}} = \mho_{t_i} = \cup_{t_j \in V^q \cap V^d}\{(t_i, t_j)_d\}$. Thus, $|\mho_{t_i}| = |V^q \cap V^d|$, i.e., the number of query terms in d.
- If $|V^q| = |\{t\}| = 1$, then $\mho_{t_i}^{\Xi^+} = \cup_{d \in \Xi^+}\{(t_i, t)_d\}$. Thus, $|\mho_{t_i}^{\Xi^+}| = \sum_{d \in \Xi^+} |\{t\} \cap V^d|$, i.e., the number of documents in which t occurs.
- If $|\Xi^+| = |\{d\}| = 1$ and $|V^q| = |\{t\}| = 1$, then $\mho_{t_i}^{\Xi^+} = \mho_{t_i} = \{(t_i, t)_d\}$. Thus, $|\mho_{t_i}| = |\{t\} \cap V^d|$, i.e., 1 if t occurs in d, 0 otherwise.

Informally speaking, it is clear that the extent of the mutual association of term t_i with the query would depend relatively on the size of $\mho_{t_i}^{\Xi^+}$, that is, the number of possible term pairs in $\mho_{t_i}^{\Xi^+}$. The fewer term pairs the set $\mho_{t_i}^{\Xi^+}$ has, the less chance term t_i has mutual information with good query terms, the less the mutual association of term t_i is with the query as a whole. By 'relatively' here we mean that the length of each document $d \in \Xi^+$ and the size of set Ξ^+ should be taken into account. The size of $\mho_{t_i}^{\Xi^+}$ is likely to be greater for longer documents than shorter ones, and for a larger sample set than a smaller one. This problem can be managed in some way by, for instance, normalizing the length of documents (e.g., using the probability distributions to represent documents), and normalizing the size of set Ξ^+ (e.g., using factor $\frac{1}{|\Xi^+|}$).

In order to give you some idea of what association sets look like, we consider the following example.

Example 7.6.1 Let us return to the example for the query-based association concept (cf. Examples 7.4.3). Let $t_i \in V^{\Xi^+} - V^q$ and $t_j \in V^q \cap V^{\Xi^+}$. Then we can give association sets $\mho_{t_i}^{\Xi^+}$ for all terms $t_i \in V^{\Xi^+} - V^q$ in the resultant table below.

Table 7.6.1 Association sets for terms $t_i \in V^{\Xi^+} - V^q$

$t_i \backslash t_j$	t_2	t_5	$\mho_{t_i}^{\Xi^+}$
t_1	d_1	d_1, d_2	$\mho_{t_1}^{\Xi^+} = \{(t_1, t_2)_{d_1}, (t_1, t_5)_{d_1}, (t_1, t_5)_{d_2}\}$
t_3	d_1	d_1	$\mho_{t_3}^{\Xi^+} = \{(t_3, t_2)_{d_1}, (t_3, t_5)_{d_1}\}$
t_4	d_1	d_1, d_2	$\mho_{t_4}^{\Xi^+} = \{(t_4, t_2)_{d_1}, (t_4, t_5)_{d_1}, (t_4, t_5)_{d_2}\}$
t_7	none	d_2	$\mho_{t_7}^{\Xi^+} = \{(t_7, t_5)_{d_2}\}$
t_8	none	none	$\mho_{t_8}^{\Xi^+} = \emptyset$

Comparing this table with Table 7.5.2, we can clearly see that each element of $\mho_{t_i}^{\Xi^+}$ corresponds to some d_m in one of the non-empty cells for term t_i. Thus, the results of calculating $atq_M(t_1, q)$, $atq_M(t_3, q)$ and $atq_M(t_8, q)$ by using Eq.(7.22) (taking $\delta_i = 1$) will certainly be in agreement with the calculation performed in Example 7.5.3. Also, from looking at this table, we can notice that it is reasonable for each element, $(t_i, t_j)_d$, of $\mho_{t_i}^{\Xi^+}$ to be attached a specific document d rather than expressed as a subscriptless (t_i, t_j): two elements, for instance, $(t_1, t_5)_{d_1}$ and $(t_1, t_5)_{d_2}$ of $\mho_{t_1}^{\Xi^+}$ would otherwise be indistinguishable. ♠

7.6.2 Score Functions $score_{M_1}(t_i)$ and $score_{M_2}(t_i)$

Assume that $V^{\Xi^+} - V^q$ constitutes a source of candidate terms. Now we are ready to talk about the construction of the score function, which in fact integrates all concepts and ideas from the previous sections.

Notice that the construction of the score function is rather simple. It is a direct application of the concept of query-based association. There may be a variety of ways to define $\varrho(t_j)$, and estimate $P_d(\delta_i, \delta_j)$ and $P_d(\delta)$ for computing $\mathbf{ifd}_M^d(t_i^{\delta_i}, t_j^{\delta_i})$. These ways would construct the different score functions. As an example, we show below one way by using the estimates $P_d(\delta_i, \delta_j)$ and $P_d(\delta)$ given by Methods A and B, and define $\varrho(t_j)$ as follows.

Consider a given query q. Assume that

$$\varrho(t_j) = p_q(t_j) = \frac{f_q(t)}{\sum_{t' \in V^q} f_q(t')} = \frac{f_q(t_j)}{||q||} \qquad (t_j \in V^q)$$

is an *a priori* probability of the proposition t_j being true in query q.

Consider the statistical information of the co-occurrence of terms. For each candidate term $t_i \in V^{\Xi^+} - V^q$, with Definition 7.4.3 and Eq.(7.22), the association score function may be defined by

$$score_{M_1}(t_i) = atq_M(t_i, q) = \frac{1}{|\Xi^+|} \sum_{(t_i,t_j)_d \in \mho_{t_i}^{\Xi^+}} \frac{f_q(t_j)}{||q||} \mathbf{ifd}_M^d(t_i, t_j),$$

in which, $\frac{1}{|\Xi^+|}$ and $||q||$ are basically just scale factors normalizing set Ξ^+ and query q, respectively. Notice that the scale factors are independent of all elements $(t_i, t_j)_d \in \mho_{t_i}^{\Xi^+}$. Thus, we can eliminate the factors and obtain a completely equivalent score function, which, by Eq.(7.19), can be further written as:

$$score_{M_1}(t_i) = \sum_{(t_i,t_j)_d \in \mho_{t_i}^{\Xi^+}} f_q(t_j) \times \gamma_d(t_i, t_j) \times \log \frac{\gamma_d(t_i, t_j)}{p_d(t_i)p_d(t_j)},$$

which is called the (mutual) association *score* of term t_i with query q concerning set Ξ^+.

We can see, as function $score_I(t)$, that the score function is the summation of the product of three essential factors — frequency $f_q(t_j)$ of query term t_j, probability $P_d(\delta_i = 1, \delta_j = 1) = \gamma_d(t_i, t_j)$ of state value $(1, 1)$, and the mutual information $i_d(H_1 : H_2|(\delta_i = 1, \delta_j = 1)) = \log \frac{\gamma_d(t_i,t_j)}{p_d(t_i)p_d(t_j)}$ of terms t_i and t_j concerning relevant document d under the state value $(1, 1)$ — over the association set $\mho_{t_i}^{\Xi^+}$ of term t_i.

The idea of using statistical information of co-occurrence of terms leads us next to the idea of using statistical information of 'none-occurrence' of terms (neither of the terms occur

in some document). In fact, it is fairly natural to conceive of the matter that, if two terms are closely related to the same topic, then they should have identical state values, namely, either they co-occur or neither of them occurs, in some relevant documents. In fact, for measuring the mutual association of terms with the query, the statistical information of the none-occurrence of terms, which might be equally important as the one of the co-occurrence of terms, should also be taken into consideration.

Under such a consideration, for each candidate term $t_i \in V^{\Xi+} - V^q$, an alternative way to constructing the association score function would be formulated by

$$score_{M_2}(t_i) = \sum_{\delta_i=1,0} atq_M(t_i^{\delta_i}, q) = \frac{1}{|\Xi^+|} \sum_{(t_i,t_j)_d \in \mho_{t_i}^{\Xi+}} \frac{f_q(t_j)}{||q||} \big[\mathbf{ifd}_M^d(t_i,t_j) + \mathbf{ifd}_M^d(\bar{t}_i,\bar{t}_j)\big].$$

Similar to $score_{M_1}(t_i)$, by eliminating $\frac{1}{|\Xi^+|}$ and $||q||$, we have an equivalent score function, which, by Eq.(7.19), can be further written as:

$$score_{M_2}(t_i) = \sum_{(t_i,t_j)_d \in \mho_{t_i}^{\Xi+}} f_q(t_j)\big[\gamma_d(t_i,t_j)\log\frac{\gamma_d(t_i,t_j)}{p_d(t)p_d(t_j)} + \\ + \big(1-p_d(t_i)-p_d(t_j)+\gamma_d(t_i,t_j)\big)\log\frac{1-p_d(t_i)-p_d(t_j)+\gamma_d(t_i,t_j)}{\big(1-p_d(t_i)\big)\big(1-p_d(t_j)\big)}\big].$$

Some details about the computation of scores of terms using two score functions for Methods A and B, respectively, can be found in Section 10.6 (see Examples A and B).

7.6.3 About Positive Scores

Recall that, by Corollaries 7.3.1 and 7.3.2, we have $\gamma_d(t_i,t_j) > p_d(t)p_d(t_j)$, $\mathbf{ifd}_M^d(t_i,t_j) > 0$ and $\mathbf{ifd}_M^d(\bar{t}_i,\bar{t}_j) > 0$, for arbitrary terms $t_i, t_j \in V^d$ and every $d \in \Xi^+$. Because $f_q(t_j) > 0$ for all query terms $t_j \in V^q \cap V^{\Xi+}$, we thus have $score_{M_1}(t_i) > 0$ and $score_{M_2}(t_i) > 0$ for every term $t_i \in V^{\Xi+} - V^q$.

Thus, the scores can be regarded as the measure of the extent of the mutual association of the candidate terms with the context of the query. The terms selected should be those which obtain higher (positive) scores. The higher the score terms obtain, the stronger they are mutually associated with the query.

7.6.4 Relationship of Score Functions

It may be interesting to think of the relationship between these two score functions mathematically. We now attempt to explain that they might not be *equivalent*. By equivalent we mean here that they give the same order, that is, for arbitrary terms t_1 and t_2, $score_{M_1}(t_1) \leq score_{M_1}(t_2)$ implies $score_{M_2}(t_1) \leq score_{M_2}(t_2)$, and vice versa. In other words, we want to show that order $score_{M_2}(t_1) \leq score_{M_2}(t_2)$ may not guarantee the same order $score_{M_1}(t_1) \leq score_{M_1}(t_2)$. Our final theorem is established for this purpose, its proof can be found in Section 10.4.

Theorem 7.6.1 Given two candidate terms $t_1, t_2 \in V^{\Xi+} - V^q$, if $score_{M_2}(t_1) \leq score_{M_2}(t_2)$, then there exists a function $\nu(t_1,t_2)$, such that

$$score_{M_1}(t_1) \leq score_{M_1}(t_2) + \nu(t_1,t_2),$$

where $\nu(t_1, t_2)$ may not be always equal to zero.

It is important to realize that equivalence requires $\nu = 0$ and that $\nu \neq 0$ may lead to non-equivalence. This can be illustrated by a very simple example. Suppose $score_{M_1}(t_1) = 0.5$ and $score_{M_1}(t_2) = 0.3$; $score_{M_2}(t_1) = 0.2$ and $score_{M_2}(t_2) = 0.3$; $\nu(t_1, t_2) = 0.4 \neq 0$. Then,

$$score_{M_1}(t_1) < score_{M_1}(t_2) + \nu(t_1, t_2).$$

However, it is clear that these two score functions are not equivalent since

$$score_{M_2}(t_1) < score_{M_2}(t_2) \quad \text{and} \quad score_{M_1}(t_1) > score_{M_1}(t_2).$$

In contrast to function $score_{M_1}(t_i)$, function $score_{M_2}(t_i)$ takes into account the consistent mutual information of terms which, by Theorem 7.6.1, incorporates some additional information into scores of terms. Therefore, $score_{M_2}(t_i)$ might be more accurate at estimating the mutual association of terms with the query than $score_{M_1}(t_i)$. But a major advantage of $score_{M_1}(t_i)$ is that it is simple to compute.

7.6.5 A Few Points of Discussion

There are a some interesting points to make about the score functions given in the current section.

- For some query term $t' \notin V^{\Xi^+}$ (i.e., $t' \notin V^q \cap V^{\Xi^+}$), the method proposed in the current chapter will immediately discard it. As mentioned, if $t' \notin V^{\Xi^+}$ then it does not occur in any sample documents, and is not considered as a good term under the hypothesis. In fact, in this case, we also have $\mathbf{ifd}^d_M(t_i, t') = \mathbf{ifd}^d_M(\bar{t}_i, \bar{t}') = 0$ for every $d \in \Xi^+$ when $t_i \in V^{\Xi^+} - V^q$.

- Functions $score_{M_1}(t)$ and $score_{M_2}(t)$ do not assign scores for any query terms. They just judge good terms among candidate terms $t_i \in V^{\Xi^+} - V^q$, whereas all query terms $t_j \in V^q \cap V^{\Xi^+}$ are immediately considered as good ones with respect to the query itself under the hypothesis.

- In constructing the score functions, we intentionally disregard the inconsistent mutual information of terms. That is, they are constructed by considering the consistent mutual information rather than the expected mutual information. This is because, as shown in Corollaries 7.3.1 and 7.3.2, the signs of the consistent mutual information are always positive, whereas the signs of the inconsistent mutual information are always non-positive. The expected mutual information is calculated based on the sum of the individual items over the state space, and thus the amounts of information given by the individual items offset one another (see Examples A and B given in Section 10.6). In practice, we are concerned usually with occurrence or/and none-occurrence of terms. Therefore, it might be a sensible way to construct the score functions using only the consistent mutual information.

- For a given query q and its set Ξ^+, let us now analyse the complexity for computing scores for all candidate terms. Notice that for each $t_i \in V^{\Xi^+} - V^q$, its association set $\mathcal{U}^{\Xi^+}_{t_i}$ consists of elements $(t_i, t_j)_d$ satisfying $t_j \in V^{\Xi^+} \cap V^q$ and $d \in V^{\Xi^+}$. Thus, it can easily be seen that the association sets of individual candidate terms will never intersect.

that is, for two different terms $t_i, t_i' \in V^{\Xi^+} - V^q$ we always have $\mho_{t_i}^{\Xi^+} \cap \mho_{t_i'}^{\Xi^+} = \emptyset$ (see Example 7.5.2, for instance). Denote

$$\begin{aligned} G &= \bigcup_{t_i \in (V^{\Xi^+} - V^q)} \mho_{t_i}^{\Xi^+}, \\ G_1 &= (V^{\Xi^+} - V^q) \times (V^{\Xi^+} \cap V^q) \times \Xi^+, \\ G_2 &= V^{\Xi^+} \times V^{\Xi^+} \times \Xi^+, \\ G_3 &= V \times V \times D, \end{aligned}$$

Obviously, each element $(t_i, t_j)_d \in G$ corresponds to one computation of $\mathbf{ifd}_M^d(t_i, t_j)$. Thus, we need total $|G|$ number of computations of $\mathbf{ifd}_M^d(t_i, t_j)$ for function $score_{M_1}(t)$, and $2|G|$ for function $score_{M_2}(t)$. In practice, the size of G is much smaller than the size of G_1, and is inconsiderable compared with the sizes of G_2 and G_3. For instance, from Table 7.5.2 and Table 7.6.1, we have

$$\begin{aligned} |G| &= 9, \\ |G_1| &= 5 \times 2 \times 3 = 30, \\ |G_2| &= 8 \times 8 \times 3 = 192, \\ |G_3| &\to \infty \quad (\text{when } n = |V| \to \infty \text{ and/or } N = |D| \to \infty). \end{aligned}$$

Consequently, the total computation involved in our methods is not expensive.

7.7 Extension

In the last two sections, we expounded the concepts of the mutual association and the constructions of the score functions on the premise that the information entities were individual documents, and that the state distributions were estimated by using Methods A or B. We point out that all discussions given in the last two sections may be applicable to a variety of information entities, and to the different estimation methods of the state distributions. The following discussion is made to support this viewpoint. We shall start by considering a special case where there is only one document in the relevant sample set.

7.7.1 A Special Case

Notice that we did not put any restriction on the size of set Ξ^+ for the discussions given in Sections 7.5 and 7.6. Thus, suppose now $|\Xi^+| = |\{d\}| = 1$. Obviously, in this case, the concepts of the mutual association and the construction of the score functions can be described in a simpler way.

From Definition 7.5.1, for terms $t_i, t_j \in V^{\Xi^+} = V^d$, the mutual association of term t_i with term t_j under the state values (δ_i, δ_i) can be written as

$$att_M(t_i^{\delta_i}, t_j^{\delta_i}) = \frac{1}{|\Xi^+|} \sum_{d \in \Xi^+} \mathbf{ifd}_M^d(t_i^{\delta_i}, t_j^{\delta_i}) = \mathbf{ifd}_M^d(t_i^{\delta_i}, t_j^{\delta_i}).$$

From Definition 7.5.3, for a given term $t_i \in V^{\Xi^+} - V^q = V^{\{d\}} - V^q = V^d - V^q$, the mutual association of term t_i with query q under the state value (δ_i, δ_i) can be written as

$$atq_M(t_i^{\delta_i}, q) = \sum_{t_j \in V^q \cap V^d} \varrho(t_j) tba(t_i^{\delta_i}, t_j^{\delta_i}) = \sum_{t_j \in V^q \cap V^d} \varrho(t_j) \mathbf{ifd}_M^d(t_i^{\delta_i}, t_j^{\delta_i}).$$

Also, for a given term $t_i \in V^d - V^q$, the expression of the association set $\mho_{t_i}^{\Xi^+}$ of term t_i can be simplified to

$$\mho_{t_i} = \cup_{t_j \in V^q \cap V^d} \{(t_i, t_j)_d\}.$$

Therefore, we can express an alternative form of the query-based association function:

$$atq_M(t_i^{\delta_i}, q) = \sum_{(t_i, t_j)_d \in \mho_{t_i}} \varrho(t_j) \mathbf{ifd}_M^d(t_i^{\delta_i}, t_j^{\delta_i}),$$

and the score functions:

$$score_{M_1}(t_i) = \sum_{(t_i, t_j)_d \in \mho_{t_i}} f_q(t_j) \mathbf{ifd}_M^d(t_i, t_j),$$

$$score_{M_2}(t_i) = \sum_{(t_i, t_j)_d \in \mho_{t_i}} f_q(t_j) \left[\mathbf{ifd}_M^d(t_i, t_j) + \mathbf{ifd}_M^d(\bar{t}_i, \bar{t}_j)\right],$$

where $\mathbf{ifd}_M^d(t_i, t_j)$ and $\mathbf{ifd}_M^d(\bar{t}_i, \bar{t}_j)$ are given in Eq.(7.19).

7.7.2 Extension to Other Information Entities

Having discussed the special case, it is now very easy to extend our method to other information entities. Recall that we mentioned that an entity is in our method a document, and that any superentity or subentity can be thought of as a new entity, i.e., a new larger or smaller single document. Thus, for the new entity, still denoted by E, we have the same discussion as the one given in the special case just above. Therefore, the general forms for the corresponding expressions given in the special case can be written out as follows.

From Definition 7.5.1, for terms $t_i, t_j \in V^E$, the mutual association of term t_i with term t_j under the state values (δ_i, δ_j) is

$$att_M(t_i^{\delta_i}, t_j^{\delta_i}) = \mathbf{ifd}_M^E(t_i^{\delta_i}, t_j^{\delta_i}).$$

From Definition 7.5.3, for a given term $t_i \in V^E - V^q$, the mutual association of term t_i with query q under the state value (δ_i, δ_i) is

$$atq_M(t_i^{\delta_i}, q) = \sum_{t_j \in V^q \cap V^E} \varrho(t_j) \mathbf{ifd}_M^E(t_i^{\delta_i}, t_j^{\delta_i}).$$

Let $(t_i, t_j)_E$ represent $t_i, t_j \in V^E$. Thus, for a given term $t_i \in V^E - V^q$, the association set of term t_i can be expressed as

$$\mho_{t_i} = \cup_{t_j \in V^q \cap V^E} \{(t_i, t_j)_E\}.$$

Consequently, we have the query-based association function:

$$atq_M(t_i^{\delta_i}, q) = \sum_{(t_i,t_j)_E \in \mho_{t_i}} \varrho(t_j)\mathbf{ifd}_M^E(t_i^{\delta_i}, t_j^{\delta_i}).$$

and the score functions:

$$score_{M_1}(t_i) = \sum_{(t_i,t_j)_E \in \mho_{t_i}} f_q(t_j)\mathbf{ifd}_M^E(t_i, t_j),$$

$$score_{M_2}(t_i) = \sum_{(t_i,t_j)_E \in \mho_{t_i}} f_q(t_j)\big[\mathbf{ifd}_M^E(t_i, t_j) + \mathbf{ifd}_M^E(\bar{t}_i, \bar{t}_j)\big],$$

where $\mathbf{ifd}_M^E(t_i, t_j)$ and $\mathbf{ifd}_M^E(\bar{t}_i, \bar{t}_j)$ are the general expressions of the discrimination measures given in Eq.(7.18). Obviously, when functions $\gamma_E(t_i, t_j)$ and $\psi_E(t_i)$ are given, measures $\mathbf{ifd}_M^E(t_i^{\delta_i}, t_j^{\delta_i})$, where $\delta_i = 1, 0$, can be specialized, and all general forms above can be determined. We can thus write the specializations for our three estimation methods as follow.

For Methods A and B

(1) When $E = \Xi^+$, the general expression Eq.(7.18) becomes

$$\mathbf{ifd}_M^{\Xi^+}(t_i, t_j) = \gamma_{\Xi^+}(t_i, t_j) \log \frac{\gamma_{\Xi^+}(t_i, t_j)}{p_{\Xi^+}(t_i)p_{\Xi^+}(t_j)},$$

$$\mathbf{ifd}_M^{\Xi^+}(\bar{t}_i, \bar{t}_j) = \big(1 - p_{\Xi^+}(t_i) - p_{\Xi^+}(t_j) + \gamma_{\Xi^+}(t_i, t_j)\big) \log \frac{1 - p_{\Xi^+}(t_i) - p_{\Xi^+}(t_j) + \gamma_{\Xi^+}(t_i, t_j)}{\big(1 - p_{\Xi^+}(t_i)\big)\big(1 - p_{\Xi^+}(t_j)\big)},$$

where functions $\gamma_{\Xi^+}(t_i, t_j)$ and $p_{\Xi^+}(t)$ are given in Eq.(7.7) or Eq.(7.10).

(2) When $E = d_0$, the general expression Eq.(7.18) is

$$\mathbf{ifd}_M^{d_0}(t_i, t_j) = \gamma_{d_0}(t_i, t_j) \log \frac{\gamma_{d_0}(t_i, t_j)}{p_{d_0}(t_i)p_{d_0}(t_j)},$$

$$\mathbf{ifd}_M^{d_0}(\bar{t}_i, \bar{t}_j) = \big(1 - p_{d_0}(t_i) - p_{d_0}(t_j) + \gamma_{d_0}(t_i, t_j)\big) \log \frac{1 - p_{d_0}(t_i) - p_{d_0}(t_j) + \gamma_{d_0}(t_i, t_j)}{\big(1 - p_{d_0}(t_i)\big)\big(1 - p_{d_0}(t_j)\big)},$$

where functions $\gamma_{d_0}(t_i, t_j)$ and $p_{d_0}(t)$ are given in Eq.(7.8) or Eq.(7.11).

For Method C

There is only one case $E = \Xi^+$. The general expression Eq.(7.18), is written as

$$\mathbf{ifd}_M^{\Xi^+}(t_i, t_j) = \gamma_{\Xi^+}(t_i, t_j) \log \frac{\gamma_{\Xi^+}(t_i, t_j)}{\phi_{\Xi^+}(t_i)\phi_{\Xi^+}(t_j)},$$

$$\mathbf{ifd}_M^{\Xi^+}(\bar{t}_i, \bar{t}_j) = \big(1 - \phi_{\Xi^+}(t_i) - \phi_{\Xi^+}(t_j) + \gamma_{\Xi^+}(t_i, t_j)\big) \log \frac{1 - \phi_{\Xi^+}(t_i) - \phi_{\Xi^+}(t_j) + \gamma_{\Xi^+}(t_i, t_j)}{\big(1 - \phi_{\Xi^+}(t_i)\big)\big(1 - \phi_{\Xi^+}(t_j)\big)},$$

where functions $\gamma_{\Xi^+}(t_i, t_j)$ and $\phi_{\Xi^+}(t)$ are given in Eq.(7.13) and Eq.(7.14), respectively.

Notice that, by Corollary 7.3.3, the signs of the consistent mutual information $\mathbf{ifd}_M^{\Xi^+}(t_i^{\delta_i}, t_j^{\delta_i})$, where $\delta_i = 1, 0$, can be positive or negative since Method C cannot guarantee that relation $\gamma_{\Xi^+}(t_i, t_j) > \phi_{\Xi^+}(t_i)\phi_{\Xi^+}(t_j)$ holds for arbitrary terms $t_i, t_j \in V^{\Xi^+}$. Thus, we should be aware that the scores of term t_i,

$$score_{M_1}(t_i) = \sum_{(t_i,t_j)_{\Xi^+} \in \mho_{t_i}} f_q(t_j)\mathbf{ifd}_M^{\Xi^+}(t_i, t_j),$$

$$score_{M_2}(t_i) = \sum_{(t_i,t_j)_{\Xi^+} \in \mho_{t_i}} f_q(t_j)\left[\mathbf{ifd}_M^{\Xi^+}(t_i, t_j) + \mathbf{ifd}_M^{\Xi^+}(\bar{t}_i, \bar{t}_j)\right],$$

are likely to be negative.

Obviously, terms selected should be those which obtain the higher *positive* scores. The restriction that each score should be positive is imposed so that the total information quantity contained in term t_i (i.e., the algebraic sum of $f_q(t_j)\mathbf{ifd}_M^{\Xi^+}(t_i^{\delta_i}, t_j^{\delta_i})$ over $\mho_{t_i}$) should still support dependent hypothesis H_1. If the score of t_i is negative, then the total information quantity in t_i would support independent hypothesis H_2.

Some details about the computation of scores of terms using the above score functions for Method C can be found in Section 10.6 (see Example C).

7.8 Summary

This chapter focuses on discrimination using mutual information of terms. The formalism of the dependence discrimination measures is based on the concept of the expected mutual information. The notion of the amount of mutual information contained in a given term pair is formally interpreted.

¶ The mathematical methods for the estimation of term state distributions are developed. Three specific estimation methods are considered: using term co-occurrence data, using conditional probabilities, and using document frequency data. Some properties of the estimated state distributions are studied which are important for guiding practical applications. Then, a unified method is suggested and a general framework is established for tackling a variety of estimations of term state distributions.

¶ The dependence discrimination measures are formally defined corresponding to four state values of term pairs. Some relationships between the measures are revealed:

- A single relation between $\gamma_E(t_i, t_j)$ and $\psi_E(t_i) \cdot \psi_E(t_j)$ can entirely determine the signs of all the dependence discrimination measures.
- The signs of the consistent mutual information are always the same, so are the signs of the inconsistent mutual information.
- The signs of the consistent mutual information are always opposite to the signs of the inconsistent mutual information.

These relations are important, they underpin the method proposed in this chapter. Particularly, the relations of the dependence discrimination measures estimated using three specific methods are carefully discussed.

¶ The concept of the dependence of terms is analysed by clarifying the difference between broad and narrow dependence, and between global and local dependence. It is pointed out that a term very dependent on another term t_j (even $t_j \in V^q$ is a unique query term) may not imply that it is one we desire. It is also pointed out that the local dependence of a term pair need not to be equal to the global dependence of the term pair, nor the same as another local dependence of the term pair.

¶ The mutual association of a term with another term, with the relevant sample set, or with the query, are addressed. Three basic concepts: term-based association, set-based association and query-based association, are introduced, and their difference and relation are shown.

¶ Two score functions are proposed, which directly apply the concept of query-based association, for the judgement of good terms. One uses statistical information of co-occurrence of terms; the other is as the first, but incorporating statistical information of 'none-occurrence' of terms. The relationship between these two score functions is analysed, and the conclusion is that they may not be equivalent.

¶ All discussion in this chapter may be applicable to a variety of information entities, and to different estimation methods of term state distributions.

Chapter 8

Experimental Results

In this chapter, we concentrate on investigating to what extent each relevance discrimination measure contributes to improvement of retrieval performance. We evaluate the average retrieval performances of the expanded queries obtained from our methods, and compare the performances with that of the original queries without query expansion, and with that of the expanded queries obtained from the reduced Rocchio formula.

We point out that information retrieval is a complex procedure and, from an empirical point of view, it is unlikely a single technique will be effective for all retrieval problems. The effectiveness of the query expansion will be dependent on the ability to use several retrieval techniques (such as the method of weighting expanded query terms) in concert.

In Section 8.1, we introduce a reweighting function for terms of the expanded queries. In Section 8.2, we describe briefly the $\mathcal{I}f\mathcal{D}$ methodology. In Sections 8.3 and 8.4, we concentrate on investigating retrieval effectiveness resulting from the estimation of term probability distributions. In Sections 8.5 and 8.6, we focus on investigating retrieval effectiveness of the discrimination measures. In Section 8.7, we are concerned with the optimal size of sample set and number of expansion terms. In Section 8.8, our experimental results are discussed.

8.1 Weighting Function for Terms of Expanded Query

There are several points which should be considered in designing any reweighting function for terms of the expanded query: (1) The original query terms appearing in the top-ranked documents should be important and properly emphasized; (2) The association scores of terms may be an important factor in indicating the importance of terms with respect to the context of the query, and should be incorporated into weights of expanded query terms; (3) Weights of query terms and scores of selected terms should be adjusted to the same scale. Based on these three points, we can propose a reweighting function as follows.

Assume that m_w is the maximum weight among weights of the original query terms and that m_s is the maximum score among scores of selected terms, that is,

$$m_w = \max\left\{w_q(t)|\ t \in V^q \cap V^{\Xi}\right\} \quad \text{and} \quad m_s = \max\left\{score_q(t)|\ t \in S^q\right\}.$$

Let t_w be a term maximizing $w_q(t)$, and t_s a term maximizing $score_q(t)$, that is,

$$t_w = arg\max\left\{w_q(t)|\ t \in V^q \cap V^{\Xi}\right\} \quad \text{and} \quad t_s = arg\max\left\{score_q(t)|\ t \in S^q\right\},$$

that is, $w_q(t_w) = m_w$ and $score(t_s) = m_s$.

Also, take numbers L_w and L_s, such that,

$$m_w \in [10^{L_w}, 10^{L_w+1}) \quad \text{and} \quad m_s \in [10^{L_s}, 10^{L_s+1}),$$

where 10^{L_w} and 10^{L_s} are called the *significant figures*[1] of m_w and m_s, respectively.

It is reasonable to believe that the original query carries some useful information regarding the user's information needs. Thus it may be appropriate not to make too much change to the original query. For this reason, we take advantage of the original query by incorporating term weights into the expanded query, as defined by Eq.(8.1) below. As with any method of query expansion, it is important for us to have a relatively good original query.

A key point in the design of a reweighting function is to ensure that new term weights neither override the original term weights, nor have an negligible effect during the next retrieval iteration. To achieve this aim, we introduce the following simple piecewise function:

$$rew_{IfD}(t) = \begin{cases} w_q(t) + spw_q(t) & \text{when } t \in S^q \cap V^q \\ spw_q(t) & \text{when } t \in S^q - V^q \\ w_q(t) & \text{when } t \in V^q - S^q \end{cases} \tag{8.1}$$

for all $t \in V^{q'} = S^q \cup V^q$, where

$$spw_q(t) = 10^{L_w - L_s} \times score(t) \qquad (t \in S^q)$$

is referred to as a *supplementary weight* of term t, in which, $10^{L_w - L_s}$ is called a *shifting factor* of the decimal point.

The idea behind the reweighting function is the following. We are only interested in terms in $V^{q'} = S^q \cup V^q$ over which $rew_{IfD}(t)$ is defined. First of all, let us keep the original weight $w_q(t)$ unchanged for terms $t \in V^q - S^q$. Then, for each selected term $t \in S^q$, a supplementary weight $spw_q(t)$ is produced by adjusting its score to be of the same scale as the weights of the original terms, that is, by proportionately increasing or decreasing the score by means of consistently shifting the decimal point for each of them. The supplementary weight is assigned to terms in $S^q - V^q$ immediately, and added to the original weight $w_q(t)$ to highlight terms in $S^q \cap V^q$.

It is interesting to notice that the supplementary weight satisfies a constraint:

$$spw_q(t_s) = 10^{L_w - L_s} \times m_s \in [10^{L_w}, 10^{L_w+1}).$$

The constraint states that the maximum supplementary weight $spw_q(t_s)$ should fall in the same significant interval as the maximum weight m_w by multiplying the maximum score m_s by the shifting factor $10^{L_w - L_s}$. There exists one and only one shifting factor under the constraint. It is easily verifiable that the shifting factor with the form $10^{L_w - L_s}$ will satisfy the constraint, such that, $spw_q(t_s) = 10^{L_w - L_s} \times m_s$ lies precisely in *significant interval*[2] $[10^{L_w}, 10^{L_w+1})$ of m_w. The direction and distance (the number of digits) to be shifted are

[1] For any real number x, there exists an integer L satisfying $x \in [10^L, 10^{L+1})$. Number 10^L is called the *significant figure* of x, digit L the *significant digit* of x, and $[10^L, 10^{L+1})$ the *significant interval* of x. For instance, for $x = 0.0765 \in [10^{-2}, 10^{-2+1}) = [0.01, 0.1)$, 10^{-2} is its significant figure, 2 its significant digit, and $[0.01, 0.1)$ its significant interval.

[2] See preceding footnote.

determined by the shifting factor $10^{L_w - L_s}$, which depends only on two maximums, m_w and m_s, under the constraint.

Consequently, when $L_s = L_w$ ($L_w - L_s$ is zero in this case), that is, both m_s and m_w have the same significant interval, we need not shift the decimal point, and we retain the original scores as the supplementary weights for all terms $t \in S^q$. When $L_s \neq L_w$ ($L_w - L_s$ can be positive or negative), that is, two significant intervals of m_s and m_w are not completely overlapping, the shifting factor $10^{L_w - L_s}$ is used to carry out the shifting of the decimal point on the scores of terms, and the supplementary weights of all terms $t \in S^q$ can be obtained from their corresponding modified scores.

A simple example given below illustrates how our reweighting function works.

Example 8.1.1 From Table 8.1.1, it can be easily seen that

$$m_w = w_q(t_3) = 0.0765 = 7.65 \times 10^{-2} \in [10^{-2}, 10^{-1}),$$
$$m_s = score(t_1) = 6.6317 \times 10^{-8} \in [10^{-8}, 10^{-7}).$$

Thus, we have $L_w - L_s = (-2) - (-8) = 6$. Take the supplementary weight

$$spw_q(t) = 10^6 \times score(t) \quad (t \in S^q),$$

which satisfies

$$spw_q(t_s) = 10^6 \times m_s = 0.066317 = 6.6317 \times 10^{-2} \in [10^{-2}, 10^{-1}).$$

Table 8.1.1 Reweighting terms in the expanded query

V^q	$w_q(t)$	S^q	$score(t)$	$spw_q(t)$	$rew_{IfD}(t)$
t_1	0.0534	t_1	6.6317×10^{-8}	0.066317	0.119717
		t_2	4.1003×10^{-8}	0.041003	0.041003
t_3	0.0765	t_3	3.7103×10^{-8}	0.037103	0.113603
		t_4	1.0604×10^{-8}	0.010604	0.010604
		t_5	8.2038×10^{-9}	0.008204	0.008204
t_6	0.0227				0.022700

This example shows intuitively how it is possible for the reweighting function to achieve our aim that supplementary weights of terms neither override the original term weights, nor have a negligible effect in the next retrieval iteration. ♠

In practice, we almost always have $t_s \in S^q \cap V^q$. Therefore, the design of function $rew_{IfD}(t)$ is in effect based chiefly on consideration of the 'most important' term, t_w, of the query and the 'most associated' term, t_s, with the query (they are frequently the same one). Of course, more importantly, the function provides an effective way to incorporate the scores (i.e., the information on the power of discrimination of terms with respect to the original query) into the new weights of the expanded query terms, particularly, when the two values $w_q(t)$ and $score(t)$ are different in scale.

Notice that, for a given query q, the maximums m_w and m_s are fixed after the score functions are applied to the query. Thus, the shifting factor $10^{L_w - L_s}$ can be uniquely determined with respect to the query. However, the factor may be very different from query to query, from collection to collection (even for the same query), and further, from model to model.

Notice also that we do not introduce any additional parameters in Eq.(8.1). This is because the parameters have to be entirely determined experimentally: they are thus dependent on the document representations, query representations and score functions.

8.2 Overview of the $\mathcal{I}f\mathcal{D}$ Methodology

In order to investigate to what extent our methods of query expansion improve retrieval performance, we have carried out a number of experiments with TREC data [215]. The components of $\mathcal{I}f\mathcal{D}$ used in our experiments include: database, vocabulary, query expansion process and baselines. These components are presented below.

8.2.1 Database

Our experiments use two collections from the TREC *ad hoc* data: AP90 (the Associated Press newswire, 1990) and FT (the Financial Times 1991-1994). The statistics for these two collections can be found in Table 8.2.1.

Table 8.2.1 Document collection statistics

Collection	Number of documents	Mean number of terms / document
AP90	78,321	478.4
FT	210,158	412.7

Table 8.2.2 Topic set statistics

	Min	Max	Mean
TREC-4 (201-250)	8	33	16.3
description	8	33	16.3
TREC-7 (351-400)	31	114	57.6
title	1	3	2.5
description	5	34	14.3
narrative	14	92	40.8

Each document is formed by extracting from collections AP90 and FT. Words are strings of alphanumeric character. No stop words were removed, and no word stemming was performed.

Regarding queries for the experiments, we use two sets of queries, which are automatically derived from the corresponding two groups of 50 natural language *topics*. The two groups of topics are TREC-4 (201-250) and TREC-7 (351-400). The number of words in topics, including stop words, is shown in Table 8.2.2.

For TREC topics, the title field consists of words considered to best describe the topics. The description field is a one sentence description of the topic area, which might not contain all the words of the title field. The narrative field gives a concise description of what makes a document relevant. A typical example of a topic is shown as follows.

For TREC-4, each of the queries (201-250) is produced from the corresponding description of the topic, which is the only field (denoted by *desc-only*). For TREC-7, each of the queries (351-400) is produced, respectively, from the corresponding title field (denoted by *title-only*), both title and description fields (denoted by *title+desc*), and the full text of the topic (denoted by *full-text*).

< *top* >

< *num* > Number: 354

< *title* > Journalist risks

< *desc* > Description:
Identify instances where a journalist has been put at risk (e.g., killed, arrested or taken hostage) in the performance of his work.

< *narr* > Narrative:
Any document identifying an instance where a journalist or correspondent has been killed, arrested or taken hostage in the performance of his work is relevant.

< */top* >

In subsequent sections, for the purpose of discussing the experiments, we always suppose that the TREC-4 queries (201-250) are used to retrieve against collection AP90, and the TREC-7 queries (351-400) against collection FT.

8.2.2 Vocabulary

It has been recognized since the earliest days of IR [119] that many of the most frequently occurring words in English (such as, 'and', 'of', 'the', 'to', etc., called as common function words) are worthless as indexing terms. These words make up a large fraction of the text of most documents: the ten most frequently occurring words in English typically account for 20% to 30% of the words in a document [59]. Eliminating such words from consideration early in automatic indexing processing, saves huge amounts of space in indexes, and does not substantially damage retrieval effectiveness. A list of common function words, filtered out during automatic indexing processing, is called a *stoplist*. A frequently used stoplist of 250 words can be found in [207]. A stoplist of 425 words drawn from a broad range of literature in English was given by [59]. Fox [57] discussed the derivation of a stoplist, which is specially constructed by using the lexical analysis. In $\mathcal{IFD}$, a stoplist [207] is used to delete common function words from document and query texts.

As Fox [58] stated that most numbers are not good discriminators, and should be removed from further consideration as indexing terms. However, certain numbers in some kinds of databases may be useful. A solution for such a problem is to allow terms including numbers but do not begin with a number. Also, he stated that breaking hyphenated terms into their constituents may lose the specificity of a hyphenated phrase. Breaking up hyphenated terms increases recall but decreases precision. Following his statements, $\mathcal{IFD}$ removes all numbers that do not begin with a number. However, $\mathcal{IFD}$ breaks hyphenated terms into their constituents. The retrieval with hyphenated terms will be considered as a further work.

Stemming may be useful for the entire retrieval process, particularly, to alleviate deficiencies in the lexicon. $\mathcal{IFD}$ uses a stemming algorithm [137] for suffix-stripping to extract word stem forms. Word stems are called *terms* in this thesis.

As we have stated repeatedly, the document frequencies, $F_D(t)$, of terms can be used as an important factor in determining whether a term should be considered as an indexing term. A term with a very high value $F_D(t)$ tends to have a poor power of discrimination, and should be dropped from the vocabulary. A term with a very low value $F_D(t)$ should

also be dropped from the vocabulary. There are normally many very infrequent terms in the collection, and many of them are strange codes or misspellings. The terms with very high or very low document frequencies should be regarded as '*bad*' terms. All others are regarded as '*not bad*' terms, and are therefore deemed 'useful' indexing terms. However, it is not necessarily the case that these 'not bad' terms are *good* ones, that is, they may or may not be informative or associated with a given query.

It is desirable that a system exhibits both high recall by retrieving all documents that are relevant, and also high precision by rejecting all documents that are non-relevant. As is well-known, the recall-preferred weighting function appears to use more general, high-frequency terms that occur in many documents of the collection. Such terms may be expected to pull out many documents, including many of the relevant documents. The precision-preferred weighting function, however, seems to support more specific, very low-frequency terms that are capable of isolating the few relevant documents from the mass of non-relevant ones.

Thus, the removal of terms with very high document frequencies would result in decreasing the retrieval recall by making it impossible to accept some relevant documents. The deletion of many terms with very low document frequencies, on the other hand, would tend to diminish retrieval precision by making it impossible to reject many non-relevant documents. In practice, compromises are normally made to achieve a reasonable recall level, without at the same time producing unreasonably poor precision [164]. The different recall and precision requirements favour the combined use of a variety of term weighting factors that contain both recall- and precision-enhancing components. We have discussed this issue in Section 3.7.

Following the studies given in [169, 170, 171, 172] (see Section 2.4) $\mathcal{I}f\mathcal{D}$ removes those terms that appear in more than ten percent of the size of the collection or in less than three documents. That is, we have

$$V = \{\, t \mid 2 < F_D(t) < 0.1 \times |D| \,\}.$$

By dropping these '*bad*' terms, the size of the vocabulary is greatly reduced. For instance, in the FT collection, roughly a quarter of terms occur in only one or two documents. We note, however, that some experiments have shown that retrieval performance decreases slightly when very frequent and infrequent terms are not used as indexing terms [17, 141].

In this thesis, content units, which are used for content representations of documents and queries, are treated as *single terms*. In most early experiments, quite effective retrieval was achieved using single-term content representations [167, 207]. Salton & Buckley [164] reviewed some past studies and pointed out:

> "*Ultimately, however, sets of single terms cannot provide complete identification of document content. For this reason, many enhancements in content analysis and text indexing procedures have been proposed over the years in an effort to generate complex text representations. ... It was evident that the construction and identification of complex text representations was inordinately difficult. ... The overwhelming evidence is that the judicious use of single-term identifiers is preferable to the incorporation of more complex entities extracted from the texts themselves ... the performance of the retrieval system with complex identifiers (such as, term phrases) will differ only marginally from the results obtainable with single-term indexing.*"

8.2.3 Query Expansion Process

For the initial retrieval, we use the Okapi weighting scheme (BM25) [155] to calculate initial term weights $w_d^*(t)$ and $w_q^*(t)$, $t \in V$, for documents and queries, respectively. That is,

$$w_d^*(t) = \frac{(1.2+1)f_d(t)}{1.2 \times \left[(1-0.75) + 0.75\frac{||d||}{ave(D)}\right] + f_d(t)},$$

$$w_q^*(t) = \frac{(1000+1)f_q(t)}{1000+f_q(t)} \times \log \frac{|D| - F_D(t) + 0.5}{F_D(t) + 0.5},$$

where $ave(D)$ is the average length of documents in collection D. This weighting scheme has widely been realized to be able to produce good initial retrieval performance.

Obviously, an n-tuple $M_d = [w_d^*(t_1), w_d^*(t_2), ..., w_d^*(t_n)]$ can also be seen as an n-dimensional vector $v_d = (w_d^*(t_1), w_d^*(t_2), ..., w_d^*(t_n))$ in the Euclidean space. Thus, the *inner product*,

$$sim(d,q) = v_d \cdot v_q = \sum_{t \in V} w_d^*(t) \cdot w_q^*(t),$$

is used as a decision function in our experiments to compute the similarity between document d and query q.

Based on the initial retrieval, for each original query q, the sample set Ξ consisting of top-ranked documents. In pseudo-relevance feedback, in order to avoid potential harm caused by a large amount of information noise in too large a sample set, 10 top-ranked documents are used in our experiments (i.e., taking $|\Xi| = 10$). In this case the candidate terms come from V^{Ξ}. In relevance feedback, however, we use 100 top-ranked documents (i.e., taking $|\Xi| = 100$). In this case the candidate terms come from V^{Ξ^+}, where $\Xi^+ = \Xi \cap R$ and R is provided from TREC data.

Then our score functions, $score(t)$ and $score^*(t)$, see Example 3.6.2, for instance, are applied to score and rank candidate terms. 30 additional expansion terms are selected (i.e., taking $|E^q| = |S^q - V^q| = 30$). Our reweighting function $rew_{IfD}(t)$ is then performed over terms of the expanded query (i.e., terms $t \in V^{q'} = S^q \cup V^q$). The whole collection then goes through the second retrieval iteration with respect to the expanded query, and documents are re-ranked using the similarity measure

$$sim(d,q') = v_d \cdot v_{q'} = \sum_{t \in V} w_d^*(t) \cdot rew_{IfD}(t).$$

Finally, the results are presented to the user.

The $\mathcal{IFD}$ system is implemented in the Java programming language. $\mathcal{IFD}$ incurs an initial time cost by indexing the whole collection and by performing the initial ranking with respect to a set of original queries. This one-off cost can be relatively large (several hours in our implementation) because I/O procedures are not optimized. The time necessary for performing solely query expansion is negligible (a few seconds). As the collection is stored as an inverted file, the computation of the term probability distributions is straightforward. Our query expansion methods are not complex, and the time required for the second ranking is proportional to the number of terms in the expanded queries, and is comparable to standard query expansion methods, such as [150].

8.2.4 Three Benchmarks

In order to evaluate the retrieval effectiveness of $\mathcal{IFD}$ at satisfying user information needs, experiments have been conducted in terms of the standard evaluation measures using the TREC evaluation tools [200]. This subsection presents the average retrieval performance of the original queries, and that of the expanded queries using a reduced Rocchio formula. These results are adopted as benchmarks in our experimental studies demonstrated in the subsequent sections.

In our experiments, all queries are considered equally since all users are assumed to be equally important. Hereafter, the standard evaluation measures used in all the result tables in this chapter are Average Precision (non-interpolated) over the set of 50 queries (denoted by A-P), Precision at 5 and 10 documents (denoted by P@5 and P@10, respectively) and R-Precision (i.e., precision at $|R|$ documents, denoted by R-P).

Performance of the Original Queries

Table 8.2.3 gives statistics in terms of the standard evaluation measures when the $\mathcal{IFD}$ system, with the Okapi weighting scheme, retrieves documents with respect to the original queries. The statistics that apply to the set of queries 201-250 (desc-only) retrieving collection AP90 can be found in column 2; columns 3 to 5 pertain to the set of queries 351-400 (full-text, desc+title, title-only, respectively) retrieving collection FT.

It should be noticed that, for queries 351-400 against collection FT, there are large performance differences between the run that uses the full-text queries, the run that uses the desc+title queries, and the run that uses the title-only queries. It is clear that better performance is achieved when retrieval is based on the full-text queries. This is most likely because long queries are usually more detailed and may thus contain more useful information than the short ones. The decreases in performance using desc+title or title-only queries are rather marked.

Table 8.2.3 Performances of the original queries

	AP90	FT		
	desc-only	full-text	desc+title	title-only
P@5	0.3667	0.4542	0.3583	0.3106
P@10	0.3167	0.3604	0.2792	0.2532
A-P	0.2682	0.2697	0.2212	0.1973
R-P	0.3041	0.2925	0.2537	0.2154

For convenience, we will use *benchmark-1* as an abbreviation to mean 'the performance of the original queries'.

Performances of the Expanded Queries

A well known method of query expansion is to use the Rocchio formula Eq.(2.1) which has been shown to achieve a good retrieval performance [16, 165]. A reduced version of the formula is expressed as

$$rew_{Roc}(t) = \alpha w_q^*(t) + \frac{\beta}{|\Xi|} \sum_{d \in \Xi} \frac{w_d^*(t)}{\sqrt{\sum_{t \in V^d} [w_d^*(t)]^2}}.$$

In practice, the reduced formula can be used for both ranking candidate terms and reweighting expanded query terms [28, 29, 130, 179, 196]. It is employed in our experiments with parameter setting $\alpha = \beta = 1$.

(a) By Pseudo-Relevance Feedback

Table 8.2.4 gives statistics in terms of the standard evaluation measures. The retrieval with the expanded queries (expanding the original queries 201-250) against collection AP90 can be found in column two. Columns 3 to 5 pertain to the different expanded queries (expanding the different parts of the original queries 351-400) against collection FT.

The results in column 2 show that, for collection AP90, ranking with the expanded queries using the reduced Rocchio formula achieves better performance than ranking with the original queries at all the evaluation points (see Table 8.2.3, column 2).

The results in columns 3 to 5 also show that, for collection FT, the different parts of the original queries produce different expanded queries. As one might expect, the three runs from the different expanded queries obtain rather inconsistent average retrieval performance. It is clear that the run based on title-only queries works worse than others. However, comparing with Table 8.2.3 (columns 3 to 5), it can readily be seen that, for the three runs, the ranking with the expanded queries achieves better performances than ranking with the original queries at most evaluation points. It seems that the best performance corresponds to the run based on the expanded queries derived from full-text queries.

Table 8.2.4 Performances of the pseudo-relevance feedback queries

	AP90	FT		
	desc-only	full-text	desc+title	title-only
P@5	0.3958	0.4708	0.3583	0.3106
P@10	0.3521	0.3729	0.2896	0.2532
A-P	0.2965	0.2891	0.2361	0.2198
R-P	0.3274	0.3110	0.2718	0.2411

For convenience, we will use *benchmark-2* as an abbreviation to mean 'the performances of the expanded queries obtained from pseudo-relevance feedback using the reduced Rocchio formula'.

(b) By Relevance Feedback

Similarly, Table 8.2.5 gives statistics from investigations based on relevance-feedback. The results clearly show that ranking with the expanded queries achieves substantial improvement in performances compared with the original queries (see Table 8.2.3), and with the expanded queries derived from pseudo-relevance feedback (see Table 8.2.4), for both collections AP90 and FT. The level of improvement is marked at all the evaluation points.

Table 8.2.5 Performances of the relevance feedback queries

	AP90	FT		
	desc-only	full-text	desc+title	title-only
P@5	0.6500	0.5875	0.6255	0.6917
P@10	0.5604	0.4646	0.4915	0.5583
A-P	0.5251	0.3983	0.4307	0.5422
R-P	0.5311	0.4113	0.4316	0.5575

For convenience, we will use *benchmark-3* as an abbreviation to mean 'the performances of the expanded queries obtained from relevance feedback using the reduced Rocchio formula'.

8.3 Effects of the Different Probability Estimation Schemes (by Pseudo-Relevance Feedback)

For the experiments given in this section and the next, the selection of good terms for the expanded queries uses only function $score_I^*(t)$, which does not incorporate weights of terms of the original queries into term scores (see Example 3.6.2 given in Section 3.6). In this way, we can gain an insight into how purely discrimination information of terms can be used to select good terms.

Recall that, in Section 3.7, we discussed some schemes for estimating the term probability distributions. We gave some examples of estimations, which generate the 'different' score functions. The estimation schemes used in our experiments are listed in Table 8.3.0. The main purpose of this section and the next is to compare average retrieval performance of query expansion using these different estimation schemes.

Table 8.3.0 The different probability estimation schemes

Schemes	$d \in \Xi$	$d \in D$
scheme-1	$\chi_\Xi(d) = sim(d,q), \quad w_d(t) = p_d^*(t)$	$\chi_D(d) = 1, \quad w_d(t) = w_d^*(t)$
scheme-2	$\chi_\Xi(d) = sim(d,q), \quad w_d(t) = w_d^*(t)$	$\chi_D(d) = 1, \quad w_d(t) = w_d^*(t)$
scheme-3	$\chi_\Xi(d) = 1, \quad w_d(t) = w_d^*(t)$	$\chi_D(d) = 1, \quad w_d(t) = w_d^*(t)$
scheme-4	$\chi_\Xi(d) = sim(d,q), \quad w_d(t) = f_d(t)$	$\chi_D(d) = 1, \quad w_d(t) = f_d(t)$
scheme-5	$\chi_\Xi(d) = 1, \quad w_d(t) = f_d(t)$	$\chi_D(d) = 1, \quad w_d(t) = f_d(t)$

In which, $f_d(t)$ is the frequency of term t within document d; $w_d^*(t)$ is the weight of term t (concerning document d) derived from the Okapi weighting scheme; $p_d^*(t) = \frac{w_d^*(t)}{\sum_{t \in V^d} w_d^*(t)}$ is the probability density of term t in V^d; and $sim(d,q)$ is the cosine similarity measure between document d and query q. Thus, for instance, from the discussions given in Section 3.7, for scheme-4 , we have

$$P_\Xi(t) = \frac{w_\Xi(t)}{\sum_{t \in V^\Xi} w_\Xi(t)} = \frac{\sum_{d \in \Xi} \chi_\Xi(d) w_d(t)}{\sum_{t \in V^\Xi} \left(\sum_{d \in \Xi} \chi_\Xi(d) w_d(t) \right)} = \frac{\sum_{d \in \Xi} sim(d,q) f_d(t)}{\sum_{t \in V^\Xi} \left(\sum_{d \in \Xi} sim(d,q) f_d(t) \right)},$$

$$P_D(t) = \frac{w_D(t)}{\sum_{t \in V} w_D(t)} = \frac{\sum_{d \in D} \chi_D(d) w_d(t)}{\sum_{t \in V} \left(\sum_{d \in D} \chi_D(d) w_d(t) \right)} = \frac{\sum_{d \in D} f_d(t)}{\sum_{t \in V} \left(\sum_{d \in D} f_d(t) \right)},$$

which satisfies that $P_\Xi(t) > 0$ for every $t \in V^\Xi$ and $P_D(t) > 0$ for every $t \in V$.

The following two subsection focus on analysing average retrieval performance based on two different ways of weighting terms of the expanded queries. Our reweighting function $rew_{IfD}(t)$ and the reduced Rocchio formula $rew_{Roc}(t)$ will be considered, respectively.

8.3.1 Weighting Terms of Expanded Query by Function $rew_{IfD}(t)$

Table 8.3.1 gives the average retrieval performances in terms of standard evaluation measures when the different estimation schemes in Table 8.3.0 are considered. The reweighting function

$rew_{IfD}(t)$ is used to weight terms of the expanded queries, expanding the original queries 201-250. Collection AP90 is retrieved with respect to the expanded queries. The statistics that apply to scheme-1 can be found in column 2, column 3 pertains to scheme-2, and so on. From the experimental results, it can be seen that:

- The five estimation schemes exhibit different average retrieval performances. Scheme-4 and scheme-5 gain similar performances. Scheme-5 works slightly better than others.
- The rankings with the expanded queries obtained from the five estimation schemes achieve better performances than *benchmark-1*. The improvements are shown at all the evaluation points.
- The performances of the expanded queries obtained from the five estimation schemes are similar to *benchmark-2*. Scheme-5 shows a slight further performance improvement compared with *benchmark-2* at all the evaluation points.

Table 8.3.1 Performances of the estimation schemes on TREC-4 (desc-only)

	benchmark-1	benchmark-2	scheme-1	scheme-2	scheme-3	scheme-4	scheme-5
P@5	0.3667	0.3958	0.3792	0.3917	0.3792	[0.4167]	[0.4167]
P@10	0.3167	0.3521	0.3458	[0.3604]	0.3583	0.3510	0.3532
A-P	0.2682	0.2965	[0.3068]	0.2920	0.2910	0.3030	0.3045
R-P	0.3041	0.3274	0.3348	0.3107	0.3206	0.3323	[0.3350]

Table 8.3.2 Performances of the estimation schemes on TREC-7 (full-text)

	benchmark-1	benchmark-2	scheme-1	scheme-2	scheme-3	scheme-4	scheme-5
P@5	0.4542	0.4708	0.4542	[0.5000]	0.4875	0.4792	0.4792
P@10	0.3604	0.3729	0.3771	[0.3875]	[0.3875]	0.3646	0.3667
A-P	0.2697	0.2891	0.2692	[0.2955]	0.2776	0.2849	0.2724
R-P	0.2925	0.3110	0.2909	[0.3216]	0.2997	0.3113	0.3010

Table 8.3.3 Performances of the estimation schemes on TREC-7 (desc+title)

	benchmark-1	benchmark-2	scheme-1	scheme-2	scheme-3	scheme-4	scheme-5
P@5	0.3583	0.3583	0.3500	0.3792	0.3708	[0.3875]	0.3833
P@10	0.2792	0.2896	0.2896	0.2854	0.2854	[0.2917]	0.2896
A-P	0.2212	0.2361	0.2044	0.2170	0.2179	[0.2359]	0.2184
R-P	0.2537	0.2718	0.2416	0.2439	0.2502	[0.2792]	0.2463

Table 8.3.4 Performances of the estimation schemes on TREC-7 (title-only)

	benchmark-1	benchmark-2	scheme-1	scheme-2	scheme-3	scheme-4	scheme-5
P@5	0.3106	0.3106	0.2766	0.3191	0.3277	[0.3362]	0.3234
P@10	0.2532	0.2532	0.2426	0.2574	0.2553	[0.2596]	0.2574
A-P	0.1973	0.2198	0.1977	0.2070	0.2022	[0.2305]	0.2285
R-P	0.2154	0.2411	0.2333	0.2192	0.2188	[0.2490]	0.2483

Similarly, Tables 8.3.2–8.3.4 give the average retrieval performances when the expanded queries, expanding the original queries 351-400, are used to retrieve collection FT for the different parts of the original queries. From the experimental results, it can be seen that:

- The five estimation schemes exhibit different average retrieval performances. Scheme-2 works better than others for the full-text queries. Scheme-4 works better than others for desc+title or title-only queries. Scheme-4 and scheme-5 have similar performance gains at most evaluation points.

- The rankings with the expanded queries obtained from the five estimation schemes achieve better performances than *benchmark-1*. The improvements are shown at most evaluation points. On close inspection, Scheme-2 gains a marked performance improvement for the full-text queries. Scheme-4 gains better performances for all the different parts of the queries, and the improvements at most evaluation points are marked.

- Scheme-2 shows a further performance improvement compared with *benchmark-2* for the full-text queries. Scheme-4 shows further performance improvements compared with *benchmark-2* for desc+title or title-only queries.

In the subsequent discussions, all the experiments based on pseudo-relevance feedback use scheme-4 to estimate the term probability distributions. We chose scheme-4 because the results presented above suggest that it may achieve a relatively better performance for all the different parts of queries. In particular, it appears to work well for desc+title or title-only queries (which is to be desired), even though it may not give the best performance for the full-text queries.

8.3.2 Weighting Terms of Expanded Query by Formula $rew_{Roc}(t)$

The Rocchio formula can be used for both ranking candidate terms and reweighting expanded query terms. In order to compare average retrieval performance of our reweighting function with that of the reduced Rocchio formula, we repeat the experiments given in the last subsection using $rew_{Roc}(t)$ to reweight expanded query terms, in place of $rew_{IfD}(t)$.

Table 8.3.5 gives the retrieval performances when the query expansion is applied to queries (201-250) for retrieving collection AP90. The experimental results demonstrate that:

- The five estimation schemes exhibit different average retrieval performances, but the differences are small.

- The rankings with the expanded queries obtained from the five estimation schemes achieve much better performances than *benchmark-1*. The improvements are shown at all the evaluation points.

- The performances of the expanded queries obtained from the five estimation schemes are similar to *benchmark-2*.

- Clearly, the reduced Rocchio formula (for only reweighting terms) performs poorer than our reweighting function for scheme-4 and scheme-5. This indicates that our reweighting function is effective (see Table 8.3.1).

Similarly, Tables 8.3.6–8.3.8 give the retrieval performances when the query expansion is applied to queries 351-400 (full-text, title+desc and title-only, respectively) for retrieving collection FT. The experimental results demonstrate that:

- The five estimation schemes exhibit different average retrieval performances, but the differences are small.

- The rankings with the expanded queries obtained from the five estimation schemes achieve better performances than *benchmark-1*. The improvements are shown at almost all the evaluation points.

- Scheme-2 shows slight further performance improvements compared with *benchmark-2* for all the different parts of the queries. Scheme-4 shows a slight further performance improvement compared with *benchmark-2* for the title-only queries.

- The reduced Rocchio formula (for only reweighting terms) shows a poorer performance, for the full-text queries, compared with our reweighting function for scheme-2 (see Tables 8.3.2–8.3.4). Also, for scheme-4, it performs less well than our reweighting function for all the different parts of queries. This indicates that our reweighting function is effective.

Table 8.3.5 Performances of the estimation schemes on TREC-4 (desc-only)

	benchmark-1	benchmark-2	scheme-1	scheme-2	scheme-3	scheme-4	scheme-5
P@5	0.3667	0.3958	0.3917	0.3917	0.3917	[0.4000]	0.3917
P@10	0.3167	0.3521	0.3521	[0.3542]	0.3521	0.3500	0.3500
A-P	0.2682	0.2965	[0.3033]	0.2990	0.2993	0.2991	0.2923
R-P	0.3041	0.3274	0.3191	0.3246	0.3174	[0.3310]	0.3173

Table 8.3.6 Performances of the estimation schemes on TREC-7 (full-text)

	benchmark-1	benchmark-2	scheme-1	scheme-2	scheme-3	scheme-4	scheme-5
P@5	0.4542	0.4708	0.4708	[0.4792]	0.4750	0.4708	0.4708
P@10	0.3604	0.3729	0.3729	[0.3771]	0.3750	0.3625	0.3667
A-P	0.2697	0.2891	0.2844	0.2907	[0.2912]	0.2845	0.2902
R-P	0.2925	0.3110	0.3123	[0.3136]	0.3126	0.3098	0.3128

Table 8.3.7 Performances of the estimation schemes on TREC-7 (desc+title)

	benchmark-1	benchmark-2	scheme-1	scheme-2	scheme-3	scheme-4	scheme-5
P@5	0.3583	0.3583	0.3625	[0.3792]	[0.3792]	0.3708	0.3708
P@10	0.2792	0.2896	0.2771	0.2813	0.2792	0.2852	[0.2854]
A-P	0.2212	0.2361	0.2296	[0.2360]	0.2347	0.2345	0.2347
R-P	0.2537	0.2718	0.2724	[0.2739]	0.2735	0.2704	0.2689

Table 8.3.8 Performances of the estimation schemes on TREC-7 (title-only)

	benchmark-1	benchmark-2	scheme-1	scheme-2	scheme-3	scheme-4	scheme-5
P@5	0.3106	0.3106	0.2979	0.3106	0.3106	[0.3234]	0.3191
P@10	0.2532	0.2532	0.2468	0.2574	0.2553	[0.2617]	0.2596
A-P	0.1973	0.2198	0.2088	0.2211	0.2072	0.2263	[0.2287]
R-P	0.2154	0.2411	0.2418	0.2452	0.2239	[0.2483]	0.2480

8.4 Effects of the Different Probability Estimation Schemes (by Relevance Feedback)

This section presents the results of experiments which use relevance information to estimate the term probability distributions. As with the discussions given in the last section, the selection of good terms uses function $score^*_I(t)$, and the reweighting of terms uses two functions $rew_{IfD}(t)$ and $rew_{Roc}(t)$. The experiments may give the effective performance limit of our methods because they use the complete relevance information.

8.4.1 Weighting Terms of Expanded Query by Function $rew_{IfD}(t)$

From the experimental results in Tables 8.4.1–8.4.4, it can be seen that:

- The five estimation schemes obtain different average retrieval performances, and the differences are marked. It is clear that scheme-3 shows the best performance compared with others.

- The rankings with the expanded queries generated using relevance information achieve dramatically better performances than *benchmark-1*. The improvements are shown at all the evaluation points, for all the different parts of the queries, for the five estimation schemes.

- The performances of the expanded queries obtained from the five estimation schemes, using relevance information, are significantly better than *benchmark-2*. The further improvements are shown at all the evaluation points for all the different parts of the queries.

- The performances of the expanded queries obtained from scheme-3 are markedly better than *benchmark-3*. The further improvements are shown at all the evaluation points for all the different parts of the queries. The performances of the expanded queries obtained from the other four schemes, however, are not consistently better than *benchmark-3*.

Table 8.4.1 Performances of the estimation schemes on TREC-4 (desc-only)

	benchmark-1	benchmark-2	benchmark-3	scheme-1	scheme-2	scheme-3	scheme-4	scheme-5
P@5	0.3667	0.3958	0.6500	0.7261	0.7261	[0.7500]	0.6652	0.6708
P@10	0.3167	0.3521	0.5604	0.5870	0.5870	[0.6146]	0.5239	0.5417
A-P	0.2682	0.2965	0.5251	0.5124	0.5202	[0.5834]	0.4550	0.4968
R-P	0.3041	0.3274	0.5311	0.5039	0.5193	[0.5808]	0.4856	0.5221

Table 8.4.2 Performances of the estimation schemes on TREC-7 (full-text)

	benchmark-1	benchmark-2	benchmark-3	scheme-1	scheme-2	scheme-3	scheme-4	scheme-5
P@5	0.4542	0.4708	0.5875	0.6522	0.6478	[0.6625]	0.6348	0.5792
P@10	0.3604	0.3729	0.4646	0.5196	0.5196	[0.5458]	0.4826	0.4563
A-P	0.2697	0.2891	0.3983	0.4300	0.4284	[0.4993]	0.3909	0.3803
R-P	0.2925	0.3110	0.4113	0.4385	0.4447	[0.5106]	0.4031	0.3830

Table 8.4.3 Performances of the estimation schemes on TREC-7 (desc+title)

	benchmark-1	benchmark-2	benchmark-3	scheme-1	scheme-2	scheme-3	scheme-4	scheme-5
P@5	0.3583	0.3583	0.6255	0.6791	0.7070	[0.7234]	0.6465	0.6213
P@10	0.2792	0.2896	0.4915	0.5256	0.5326	[0.5809]	0.4930	0.4809
A-P	0.2212	0.2361	0.4307	0.4016	0.4298	[0.5366]	0.3925	0.4049
R-P	0.2537	0.2718	0.4316	0.4091	0.4574	[0.5415]	0.4126	0.3967

Table 8.4.4 Performances of the estimation schemes on TREC-7 (title-only)

	benchmark-1	benchmark-2	benchmark-3	scheme-1	scheme-2	scheme-3	scheme-4	scheme-5
P@5	0.3106	0.3106	0.6917	0.6829	0.6927	[0.7375]	0.6634	0.6128
P@10	0.2532	0.2532	0.5583	0.5195	0.5366	[0.6062]	0.4780	0.4787
A-P	0.1973	0.2198	0.5422	0.4260	0.4573	[0.5986]	0.3936	0.4341
R-P	0.2154	0.2411	0.5575	0.4378	0.4744	[0.5894]	0.4135	0.4352

In subsequent sections, all the experiments based on relevance feedback will uniformly use scheme-3 to estimate the term probability distributions.

8.4.2 Weighting Terms of Expanded Query by Formula $rew_{Roc}(t)$

As in Subsection 8.3.2, the performance results presented here are obtained by incorporating formula $rew_{Roc}(t)$ into our methods for reweighting expanded query terms. The experimental results in Tables 8.4.5–8.4.8 demonstrate that:

- The five estimation schemes exhibit different average retrieval performances, and some of the differences are marked.
- The rankings with the expanded queries generated using relevance information achieve dramatically better performances than *benchmark-1* for the five estimation schemes.
- The performances of the expanded queries obtained from the five estimation schemes, using relevance information, are significantly better than *benchmark-2*.
- The performances of the expanded queries obtained from the five schemes are consistently poorer than *benchmark-3*. The performance decreases are marked.
- Obviously, the reduced Rocchio formula (for only reweighting terms) performs poorer than our reweighting function consistently for the five schemes (see Tables 8.4.1–8.4.4). This is the case at all the evaluation points for all the different parts of the queries. The performance decreases are significant. These demonstrate again that our reweighting function is effective.

Table 8.4.5 Performances of the estimation schemes on TREC-4 (desc-only)

	benchmark-1	benchmark-2	benchmark-3	scheme-1	scheme-2	scheme-3	scheme-4	scheme-5
P@5	0.3667	0.3958	0.6500	0.6208	0.6208	0.6417	0.6375	[0.6542]
P@10	0.3167	0.3521	0.5604	0.4979	0.5125	[0.5458]	0.5188	0.5354
A-P	0.2682	0.2965	0.5251	0.4452	0.4500	[0.5148]	0.4650	0.5075
R-P	0.3041	0.3274	0.5311	0.4582	0.4647	0.5327	0.4757	[0.5328]

Table 8.4.6 Performances of the estimation schemes on TREC-7 (full-text)

	benchmark-1	benchmark-2	benchmark-3	scheme-1	scheme-2	scheme-3	scheme-4	scheme-5
P@5	0.4542	0.4708	0.5875	0.5708	0.5667	0.5750	[0.5792]	0.5667
P@10	0.3604	0.3729	0.4646	0.4333	0.4354	[0.4500]	0.4438	0.4438
A-P	0.2697	0.2891	0.3983	0.3587	0.3580	0.3769	0.3612	[0.3870]
R-P	0.2925	0.3110	0.4113	0.3722	0.3731	0.3749	0.3689	[0.3957]

Table 8.4.7 Performances of the estimation schemes on TREC-7 (desc+title)

	benchmark-1	benchmark-2	benchmark-3	scheme-1	scheme-2	scheme-3	scheme-4	scheme-5
P@5	0.3583	0.3583	0.6255	0.5617	0.5574	0.5957	0.5617	[0.6043]
P@10	0.2792	0.2896	0.4915	0.4234	0.4170	[0.4681]	0.4213	0.4660
A-P	0.2212	0.2361	0.4307	0.3373	0.3383	0.3956	0.3388	[0.4019]
R-P	0.2537	0.2718	0.4316	0.3610	0.3648	0.4053	0.3642	[0.4164]

Table 8.4.8 Performances of the estimation schemes on TREC-7 (title-only)

	benchmark-1	benchmark-2	benchmark-3	scheme-1	scheme-2	scheme-3	scheme-4	scheme-5
P@5	0.3106	0.3106	0.6917	0.5787	0.5745	[0.6500]	0.5617	0.6426
P@10	0.2532	0.2532	0.5583	0.4404	0.4340	[0.5250]	0.4462	0.5043
A-P	0.1973	0.2198	0.5422	0.3644	0.3653	0.5214	0.3654	[0.5216]
R-P	0.2154	0.2411	0.5575	0.3848	0.3814	0.5278	0.3821	[0.5491]

8.5 Effects of the Different Discrimination Measures (by Pseudo-Relevance Feedback)

In this section and the next, we investigate experimentally the effects on performance of applying 'relevance' discrimination measures to AQE. The investigation of 'dependence' discrimination measures is considered further work.

Two strategies for scoring candidate terms will be considered in the experimental investigations. One scores terms by using only the discrimination information. Another calculates term scores by incorporating query term weights into the scores. Two sets of experiments corresponding to the two strategies are presented. The following techniques are adopted in the experiments discussed in this section.

- Experiments are carried out with pseudo-relevance feedback;
- Term probability distributions are estimated using scheme-4;
- Candidate terms are selected using two sets of functions
  - without considering query term weights: $score^*_I(t)$, $score^*_J(t)$, $score^*_K(t)$
  - considering query term weights: $score_I(t)$, $score_J(t)$, $score_K(t)$;
- Expanded query terms are reweighted using $rew_{IfD}(t)$.

8.5.1 Without Considering Weights of Query Terms

In the first set of experiments, we employ the strategy that scores terms using only the discrimination information without involving query term weights. That is, we use $score^*_I(t)$, $score^*_J(t)$ and $score^*_K(t)$ (see Example 3.6.2) to compute scores of the candidate terms. From the experimental results in Tables 8.5.1–8.5.4, it can be seen that:

- $score^*_I(t)$ and $score^*_J(t)$ exhibit similar performances when used for the different parts of the queries.
- The rankings with the expanded queries obtained from $score^*_I(t)$ and $score^*_J(t)$ achieve better performances than *benchmark-1*. The improvements are shown at all the evaluation points for all the different parts of the queries. The ranking with the expanded queries obtained from $score^*_K(t)$ achieves a better performance than *benchmark-1* at most evaluation points, with some exceptions.

  Of particular note is that improvements for the three score functions are most noticeable for precision at-5. This experimentally verifies that our query expansion methods are effective precision devices.
- The performances of the expanded queries obtained from $score^*_I(t)$ and $score^*_J(t)$ are similar to *benchmark-2*. This is the case at most evaluation points. However, the performance of the expanded queries obtained from $score^*_K(t)$ is poorer than *benchmark-2* at most evaluation points.

Table 8.5.1 Performances of the score functions on TREC-4 (desc-only)

	benchmark-1	benchmark-2	$score^*_I(t)$	$score^*_J(t)$	$score^*_K(t)$
P@5	0.3667	0.3958	[0.4167]	[0.4167]	0.3917
P@10	0.3167	0.3521	[0.3510]	0.3500	0.3458
A-P	0.2682	0.2965	[0.3030]	0.3025	0.2871
R-P	0.3041	0.3274	[0.3323]	0.3312	0.3008

Table 8.5.2 Performances of the score functions on TREC-7 (full-text)

	benchmark-1	benchmark-2	$score^*_I(t)$	$score^*_J(t)$	$score^*_K(t)$
P@5	0.4542	0.4708	0.4792	[0.4833]	0.4542
P@10	0.3604	0.3729	0.3646	0.3688	[0.3813]
A-P	0.2697	0.2891	0.2849	0.2864	[0.2928]
R-P	0.2925	0.3110	[0.3113]	[0.3113]	0.3107

Table 8.5.3 Performances of the score functions on TREC-7 (desc+title)

	benchmark-1	benchmark-2	$score^*_I(t)$	$score^*_J(t)$	$score^*_K(t)$
P@5	0.3583	0.3583	[0.3875]	[0.3875]	0.3708
P@10	0.2792	0.2896	[0.2917]	[0.2917]	[0.2917]
A-P	0.2212	0.2361	0.2359	[0.2384]	0.2194
R-P	0.2537	0.2718	[0.2792]	0.2791	0.2488

Table 8.5.4 Performances of the score functions on TREC-7 (title-only)

	benchmark-1	benchmark-2	$score^*_I(t)$	$score^*_J(t)$	$score^*_K(t)$
P@5	0.3106	0.3106	0.3362	0.3319	[0.3404]
P@10	0.2532	0.2532	[0.2596]	[0.2596]	[0.2596]
A-P	0.1973	0.2198	[0.2305]	0.2302	0.1917
R-P	0.2154	0.2411	[0.2490]	0.2481	0.2111

8.5.2 Considering Weights of Query Terms

Table 8.5.5 Performances of the score functions on TREC-4 (desc-only)

	benchmark-1	benchmark-2	$score_I(t)$	$score_J(t)$	$score_K(t)$
P@5	0.3667	0.3958	[0.4042]	[0.4042]	0.3958
P@10	0.3167	0.3521	0.3500	[0.3521]	0.3417
A-P	0.2682	0.2965	[0.2985]	0.2983	0.2846
R-P	0.3041	0.3274	[0.3254]	0.3252	0.2972

Table 8.5.6 Performances of the score functions on TREC-7 (full-text)

	benchmark-1	benchmark-2	$score_I(t)$	$score_J(t)$	$score_K(t)$
P@5	0.4542	0.4708	[0.4667]	[0.4667]	0.4542
P@10	0.3604	0.3729	0.3667	0.3688	[0.3750]
A-P	0.2697	0.2891	0.2808	0.2805	[0.2883]
R-P	0.2925	0.3110	[0.3083]	0.3082	0.2964

Table 8.5.7 Performances of the score functions on TREC-7 (desc+title)

	benchmark-1	benchmark-2	$score_I(t)$	$score_J(t)$	$score_K(t)$
P@5	0.3583	0.3583	0.3875	[0.3917]	0.3625
P@10	0.2792	0.2896	[0.2917]	0.2896	[0.2917]
A-P	0.2212	0.2361	[0.2355]	0.2348	0.2274
R-P	0.2537	0.2718	[0.2632]	0.2620	0.2547

Table 8.5.8 Performances of the score functions on TREC-7 (title-only)

	benchmark-1	benchmark-2	$score_I(t)$	$score_J(t)$	$score_K(t)$
P@5	0.3106	0.3106	[0.3277]	[0.3277]	0.3149
P@10	0.2532	0.2532	[0.2702]	0.2681	0.2489
A-P	0.1973	0.2198	[0.2240]	0.2227	0.2086
R-P	0.2154	0.2411	[0.2515]	0.2499	0.2391

In the second set of experiments, we use an alternative strategy which incorporates query term weights into the term scores. It combines the discrimination information with query term weights. That is, we employ $score_I(t)$, $score_J(t)$ and $score_K(t)$ to compute scores of the candidate terms. From the experimental results in Tables 8.5.5–8.5.8, it can be seen that:

- $score_I(t)$ and $score_J(t)$ exhibit different performances, but the differences are small.
- The rankings with the expanded queries obtained from the three score functions achieve better performances than *benchmark-1* at all the evaluation points. The ranking with the expanded queries obtained from $score_K(t)$ achieves a better performance than *benchmark-1* at most evaluation points, with some exceptions.
- The performances of the expanded queries obtained from the three score functions are poorer than *benchmark-2* at many evaluation points.
- Clearly, the three functions $score_I(t)$, $score_J(t)$ and $score_K(t)$ perform consistently poorer than the corresponding three functions $score_I^*(t)$, $score_J^*(t)$ and $score_K^*(t)$ (see Tables 8.5.1–8.5.4). This is the case at most evaluation points.

8.6 Effects of the Different Discrimination Measures (by Relevance Feedback)

In this section, we continue to investigate experimentally the performances achieved by applying the 'relevance' discrimination measures to AQE. As in the last section, two strategies for scoring candidate terms will be considered, and two sets of experiments corresponding to these two strategies are carried out. The following techniques are adopted in the experiments given in this section.

- Experiments are carried out with relevance feedback;
- Term probability distributions are estimated using scheme-3;
- Candidate terms are selected using two sets of functions
  - without considering query term weights: $score_I^*(t)$, $score_J^*(t)$, $score_K^*(t)$
  - considering query term weights: $score_I(t)$, $score_J(t)$, $score_K(t)$;
- Expanded query terms are reweighted using $rew_{IfD}(t)$.

8.6.1 Without Considering Weights of Query Terms

As in the first set of experiments (Subsection 8.5.1), we use only the discrimination information, i.e., $score_I^*(t)$, $score_J^*(t)$ and $score_K^*(t)$, to compute scores of candidate terms. From the experimental results shown in Tables 8.6.1–8.6.4, it can be seen that:

- $score_I^*(t)$ and $score_J^*(t)$ exhibit similar performances when used for the different parts of queries. On close inspection, $score_J^*(t)$ is slightly better than $score_I^*(t)$ for desc-only queries for AP90, whereas $score_I^*(t)$ is slightly better than $score_J^*(t)$ for all the different parts of the queries for FT. Both of them are much better than $score_K^*(t)$ at almost all the evaluation points. The three functions are consistently most effective when used for desc+title or title-only queries.

- The rankings with the expanded queries generated using relevance information achieve dramatically better performances than *benchmark-1*. The improvements are shown at all the evaluation points, for all the different parts of the queries, for the three score functions.

  Of particular note is that improvements at high precision points are largely increased consistently for the three score functions. This experimentally verifies again that our query expansion methods are effective precision devices.

- The performances of the expanded queries obtained from the three score functions, using relevance information, are significantly better than *benchmark-2*. The further improvements are shown at all the evaluation points, for all the different parts of the queries.

- The performances of expanded queries obtained from $score_I^*(t)$ and $score_J^*(t)$, using relevance information, are markedly better than *benchmark-3*. This is the case at all the evaluation points for all the different parts of the queries. The performance of expanded queries obtained from $score_K^*(t)$, using relevance information, is markedly better than *benchmark-3* at almost all the evaluation points.

Table 8.6.1 Performances of the score functions on TREC-4 (desc-only)

	benchmark-1	benchmark-2	benchmark-3	$score_I^*(t)$	$score_J^*(t)$	$score_K^*(t)$
P@5	0.3667	0.3958	0.6500	0.7500	[0.7625]	0.6708
P@10	0.3167	0.3521	0.5604	0.6146	[0.6229]	0.5687
A-P	0.2682	0.2965	0.5251	0.5834	[0.5853]	0.5421
R-P	0.3041	0.3274	0.5311	0.5808	[0.5825]	0.5413

Table 8.6.2 Performances of the score functions on TREC-7 (full-text)

	benchmark-1	benchmark-2	benchmark-3	$score_I^*(t)$	$score_J^*(t)$	$score_K^*(t)$
P@5	0.4542	0.4708	0.5875	[0.6625]	0.6542	0.6458
P@10	0.3604	0.3729	0.4646	[0.5458]	0.5375	0.5000
A-P	0.2697	0.2891	0.3983	[0.4993]	0.4950	0.4607
R-P	0.2925	0.3110	0.4113	[0.5106]	0.5019	0.4555

Table 8.6.3 Performances of the score functions on TREC-7 (desc+title)

	benchmark-1	benchmark-2	benchmark-3	$score_I^*(t)$	$score_J^*(t)$	$score_K^*(t)$
P@5	0.3583	0.3583	0.6255	[0.7234]	0.7149	0.6766
P@10	0.2792	0.2896	0.4915	0.5809	[0.5851]	0.5553
A-P	0.2212	0.2361	0.4307	[0.5366]	0.5339	0.4792
R-P	0.2537	0.2718	0.4316	[0.5415]	0.5391	0.4864

Table 8.6.4 Performances of the score functions on TREC-7 (title-only)

	benchmark-1	benchmark-2	benchmark-3	$score_I^*(t)$	$score_J^*(t)$	$score_K^*(t)$
P@5	0.3106	0.3106	0.6917	[0.7375]	0.7250	0.7333
P@10	0.2532	0.2532	0.5583	[0.6062]	0.6042	0.5937
A-P	0.1973	0.2198	0.5422	[0.5986]	0.5937	0.5552
R-P	0.2154	0.2411	0.5575	[0.5894]	0.5876	0.5450

8.6.2 Considering Weights of Query Terms

Analogously to the experiments described in Subsection 8.5.2, we incorporate query term weights into term scores, i.e., we use $score_I(t)$, $score_J(t)$ and $score_K(t)$ to select good terms. From the experimental results in Tables 8.6.5–8.6.8, it can be seen that:

- $score_I(t)$ and $score_J(t)$ exhibit similar performances when used for the different parts of the queries. When compared with $score_K(t)$, both of them show significantly better performances for desc-only queries for AP90, and better performances for full-text queries for FT overall. $score_K(t)$ works better for desc+title or title-only queries. The three score functions are consistently most effective when used for desc+title or title-only queries, at all the evaluation points.

- The rankings with the expanded queries generated using relevance information achieve dramatically better performances than *benchmark-1* for the three score functions.

- The performances of the expanded queries obtained from the three score functions, using relevance information, are significantly better than *benchmark-2*.

- The performances of expanded queries obtained from the three score functions, using relevance information, are markedly better than *benchmark-3*. This is the case at most evaluation points, except for the title-only queries. In contrast, they show consistently much poorer performances for the title-only queries.

- Obviously, the three functions $score_I(t)$, $score_J(t)$ and $score_K(t)$ perform consistently poorer than the corresponding three functions $score_I^*(t)$, $score_J^*(t)$ and $score_K^*(t)$ (see Tables 8.6.1–8.6.4). This is the case at all the evaluation points for all the different parts of the queries. Notice that the performance decreases are significant when used for desc+title or title-only queries.

Table 8.6.5 Performances of the score functions on TREC-4 (desc-only)

	benchmark-1	benchmark-2	benchmark-3	$score_I(t)$	$score_J(t)$	$score_K(t)$
P@5	0.3667	0.3958	0.6500	0.7375	[0.7417]	0.6667
P@10	0.3167	0.3521	0.5604	[0.6042]	0.6000	0.5521
A-P	0.2682	0.2965	0.5251	[0.5688]	0.5647	0.5255
R-P	0.3041	0.3274	0.5311	[0.5759]	0.5727	0.5290

Table 8.6.6 Performances of the score functions on TREC-7 (full-text)

	benchmark-1	benchmark-2	benchmark-3	$score_I(t)$	$score_J(t)$	$score_K(t)$
P@5	0.4542	0.4708	0.5875	0.5750	0.5833	[0.6083]
P@10	0.3604	0.3729	0.4646	0.4563	[0.4604]	0.4583
A-P	0.2697	0.2891	0.3983	[0.4290]	0.4288	0.4031
R-P	0.2925	0.3110	0.4113	0.4434	[0.4473]	0.4127

Table 8.6.7 Performances of the score functions on TREC-7 (desc+title)

	benchmark-1	benchmark-2	benchmark-3	$score_I(t)$	$score_J(t)$	$score_K(t)$
P@5	0.3583	0.3583	0.6255	0.6417	0.6417	[0.6500]
P@10	0.2792	0.2896	0.4915	0.5125	0.5167	[0.5396]
A-P	0.2212	0.2361	0.4307	0.4776	0.4758	[0.5014]
R-P	0.2537	0.2718	0.4316	0.4884	0.4909	[0.5183]

Table 8.6.8 Performances of the score functions on TREC-7 (title-only)

	benchmark-1	benchmark-2	benchmark-3	$score_I(t)$	$score_J(t)$	$score_K(t)$
P@5	0.3106	0.3106	0.6917	0.6458	0.6375	[0.6917]
P@10	0.2532	0.2532	0.5583	0.4979	0.4937	[0.5417]
A-P	0.1973	0.2198	0.5422	0.4929	0.4957	[0.5159]
R-P	0.2154	0.2411	0.5575	0.5173	[0.5207]	0.5153

8.7 Effects of Other Aspects on Performance

The retrieval effectiveness of AQE depends on several factors. This section experimentally investigates two aspects: the optimal size of the sample set and, the optimal number of expansion terms. Two sets of experiments are given in this section, and the following techniques are adopted in the experiments.

- Experiments are carried out with pseudo-relevance feedback;
- Term probability distributions are estimated using scheme-4;
- Candidate terms are selected using $score_I^*(t)$;
- Expanded query terms are reweighted using $rew_{IfD}(t)$.

In order to simulate a realistic retrieval, we used the desc-only queries (201-250) and desc+title queries (351-400) in our experiments.

8.7.1 The Size of Sample Set

Considering the size of sample set as a factor may be useful for AQE. If, for instance, the number of sample documents is set too high, it will yield more candidate terms than is necessary and so waste subsequent processing effort. If the number is too low, the number of relevant documents may be insufficient to yield a good set of candidate terms.

In many relevance feedback investigations, for instance, [72, 74, 76, 77, 180, 187, 188, 189, 210], the number of top-ranked documents presented to the user has typically been around 10 or 20. Of these documents only those judged relevant are used for the subsequent feedback iteration and query expansion search. Here, the sample size of interest is that of the set of documents judged relevant.

Harper [76], Martin [125], White [217] and Efthimiadis [52] suggested a sample size of 5 relevant documents. Sparck Jones [188] used a sample of 3-4 relevant documents, and in another experiment she used 1-3 relevant documents [189]. Harper suggested that at least one relevant document is needed [76]. These experiments showed that a small sample of relevant documents could be an adequate basis for reweighting terms. However, it is believed that the

larger the size of sample set, the better the probability estimation should be. The problem of selecting an optimal sample size is still an open IR research issue.

In the absence of user assessment of document relevance, an alternative method which treats the top-ranked documents as relevant is used. The top-ranked documents then become the sample documents, on which reweighting terms is based. This technique is well known and has been used by Salton and Sparck Jones in early experiments and more recently in TREC [15, 19, 54].

We carried out some experiments in order to see how the retrieval performance was affected by changes in the size of sample set. In the results given in Tables 8.7.1 and 8.7.2, the size of the sample set varies from 3, 6 up to 100, and the maximum value of each measure is displayed in square brackets. The results show that the retrieval performance increases as the size increases from 3 to 10 on the whole, and tends to gradually drop when 50 or more documents are considered. It can easily be seen that the average performance is even poorer than that of the retrieval with the original queries at some evaluation points when 50 or more pseudo-relevant documents are used.

Table 8.7.1 Performance vs. the size of sample set on TREC-4 (desc-only)

	3	6	10	20	30	50	80	100
P@5	0.4125	0.4083	[0.4167]	0.4125	0.4083	0.3750	0.4000	0.3792
P@10	0.3479	0.3354	[0.3510]	0.3396	0.3333	0.3271	0.3375	0.3229
A-P	0.2981	0.2971	[0.3030]	0.2965	0.2953	0.2802	0.2789	0.2681
R-P	0.3125	0.3189	0.3323	0.3200	[0.3303]	0.3140	0.3009	0.2983

Table 8.7.2 Performance vs. the size of sample set on TREC-7 (desc+title)

	3	6	10	20	30	50	80	100
P@5	[0.3875]	0.3750	[0.3875]	0.3833	0.3792	0.3708	0.3667	0.3583
P@10	[0.3125]	0.2958	0.2917	0.2896	0.2833	0.2774	0.2688	0.2667
A-P	0.2432	[0.2453]	0.2359	0.2302	0.2317	0.2210	0.2141	0.1940
R-P	0.2739	0.2781	[0.2792]	0.2693	0.2605	0.2508	0.2456	0.2183

8.7.2 The Number of Expansion Terms

The number of expansion terms is also regarded as a factor affecting retrieval performance.

Table 8.7.3 Performance vs. the number of expansion terms on TREC-4 (desc-only)

	10	20	30	50	80	100	150	200
P@5	0.4083	0.4083	[0.4167]	[0.4167]	0.4125	0.4083	0.4042	0.4042
P@10	0.3479	0.3479	[0.3521]	0.3500	0.3437	0.3458	0.3479	0.3511
A-P	0.2994	0.3009	[0.3030]	0.3015	0.3024	0.3016	0.3024	0.3010
R-P	[0.3366]	0.3245	0.3293	0.3159	0.3202	0.3222	0.3252	0.3243

Table 8.7.4 Performance vs. the number of expansion terms on TREC-7 (desc+title)

	10	20	30	50	80	100	150	200
P@5	0.3708	0.3750	[0.3875]	0.3875	0.3792	0.3792	0.3750	0.3750
P@10	0.2917	[0.2958]	0.2917	0.2917	0.2896	0.2917	0.2938	0.2938
A-P	0.2322	0.2353	0.2359	0.2388	0.2399	0.2399	0.2403	[0.2410]
R-P	0.2633	0.2685	0.2692	[0.2719]	0.2718	0.2711	0.2710	0.2713

For instance, in Harman's experiments [72, 74], the performance increased as the number of relevance feedback terms increased up to 20 terms, then gradually degraded with the

addition of further terms. In the experiments of Magennis & Van Rijsbergen [122], the peak performance came at only 6 terms and decreased gradually at higher numbers. Different results were reported by Buckley *et al.* [18]: the performance continued to increase as the number was increased, and never degraded. Some explanations for this difference are discussed in [19]. In some of their experiments, the number of expansion terms were set to 300 to 500 [15, 19]. In Qiu's experiments [140], 400 expansion terms were used to generate the expanded queries. It is therefore difficult to predict how query expansion performance will vary with the cut-off in the context of feedback experiments.

We performed some experiments in order to see how the retrieval performance was affected by changes in the number of expansion terms. As shown in Tables 8.7.3 and 8.7.4, the number of expansion terms varies from 10 to 200, and the maximum value of each measure is displayed in square brackets. On the whole, the performance appears slightly better in the range of 20 to 50 expansion terms. However, the variations in performances with the numbers of expansion terms do not appear significant.

8.8 Discussion of Experimental Results

There are several interesting points to make about the experimental results given in this chapter.

* It is interesting to observe that the score functions presented in this thesis and the reduced Rocchio formula generate different sets, E^q, of expansion terms. This difference indicates that these methods are not equivalent (i.e., the functions produce different candidate term orders). For example, consider query 211:

  > *How effective* are the driving* while intoxicated* (DWI*) regulations*? Has the number of deaths* caused by DWI been significantly* lowered*? Why are not penalties* as harsh for DWI drivers* as for the sober* driver.*

  First, '*harsh*' is dropped as it does not occur in any document in the collection (i.e., $harsh \notin V$). Set V^q consists of those terms marked with an asterisk. The sets, $S^q = E^q \cup V^q$ (and $|S^q| = |E^q| + |V^q|$), of selected terms corresponding to each of these score functions are shown in Table 8.8.1. The set, E^q, of expansion terms consists of those terms without asterisks ($|E^q| = 30$).

Table 8.8.1 The sets of selected terms for the different score functions

$score^*_I(t)$	$score^*_J(t)$	$score^*_K(t)$	$score_{Roc}(t)$
drunken	drunken	drive*	driver*
drive*	drive*	driver*	drive*
intox*	intox*	drunken	intox*
driver*	driver*	test	vehicl
dwi*	dwi*	vehicl	drunken
hazelwood	hazelwood	judg	car
reprimand	reprimand	justic	test
checkpoint	checkpoint	arrest	sign
motorist	clifford	michigan	judg
clifford	motorist	motorist	arrest
sober*	sober*	stop	regul*

$score^*_I(t)$	$score^*_J(t)$	$score^*_K(t)$	$score_{Roc}(t)$
vehicl	trooper	suprem	dwi*
trooper	vehicl	trooper	hit
madson	madson	sober*	sober*
michigan	michigan	injuri	alcohol
sobrieti	sobrieti	hit	requir
privaci	privaci	ship	convict
test	drunk	argu	william
drunk	test	sign	stop
justic	justic	car	drunk
intrus	intrus	privaci	struck
prouti	prouti	refus	motorist
arrest	crusad	truck	direct
crusad	breathalyz	wit	fine
jude	arrest	alcohol	person
suprem	injuri	direct	result
injuri	suprem	drunk	school
breathalyz	argu	prosecut	wit
argu	jude	passeng	trooper
stop	aground	violat	hazelwood
alcohol	alcohol	dismiss	conduct
aground	prosecut	rest	note
prosecut	stop	score	safeti
skipper	skipper	effect*	refus
hit	score	regul*	feet
effect*	regul*	lower*	traffic
regul*	effect*	death*	lower*
lower*	lower*	penalti*	death*
penalti*	penalti*	intox*	effect*
death*	significantli*	dwi*	penalti*
significantli*	death*	significantli*	significantli*

The expansion terms in this table are obtained from retrieving collection AP90 with respect to query 211; the experiments are carried out using pseudo-relevance feedback; the estimations of the term probability distributions use scheme-4.

✳ An important feature of Rocchio's method is that it emphasizes those terms which have higher frequencies of occurrence in the sample documents. Particularly, when each sample document contains at least one query term, it in effect emphasizes those terms which have higher frequencies of co-occurrence with the query terms in the sample documents.

However, all our methods consider not only the frequencies of co-occurrence of terms with the query terms in the sample documents as a feature inherent in Rocchio's method, but also the power of discrimination of terms by the divergence of the term probability distributions from one another.

For instance, recall in Chapter 3 that we discussed the discrimination measure $\mathbf{ifd}_I(t) = P_\Xi(t) \log \frac{P_\Xi(t)}{P_D(t)}$ (where $t \in V$). Obviously, this measure is directly proportional to probability $P_\Xi(t)$. Thus, it is clear that a greater probability $P_\Xi(t)$ would result in a higher frequency of occurrence of term t and, further, result in a higher frequency of co-occurrence of term t, with the query terms. Not only this, but measure $\mathbf{ifd}_I(t)$ also

considers the divergence of probability $P_D(t)$ of term t from probability $P_\Xi(t)$. Thus, it incorporates more (discrimination) information into function $score^*_I(t) = \mathbf{ifd}_I(t)$ than function $score_{Roc}(t)$. The superiority of our methods over the Rocchio method is apparent in our results.

✳ For the five estimation schemes of the term probability distributions, when the weighting of expanded query terms uses our reweighting function, and

- the experiments are carried out with pseudo-relevance feedback (see Tables 8.3.1–8.3.4):
  - ① Scheme-2 shows a markedly better performance than *benchmark-1* for the full-text queries. Scheme-4 shows better performances than *benchmark-1* for all the different parts of the queries, and the improvements at most evaluation points are marked.
  - ② Scheme-2 shows a performance increase compared with *benchmark-2* for the full-text queries. Scheme-4 shows performance increases for desc+title or the title-only queries.
- the experiments are carried out with relevance feedback (see Tables 8.4.1–8.4.4):
  - ① Scheme-3 achieves dramatically better performances than *benchmark-1*. This is the case at all the evaluation points for all the different parts of the queries.
  - ② Scheme-3 obtains significantly better performances than *benchmark-2*. The further improvements are shown at all the evaluation points for all the different parts of the queries.
  - ③ Scheme-3 gains markedly better performances than *benchmark-3*. The further improvements are shown at all the evaluation points for all the different parts of the queries.

✳ For the three score functions, when they are constructed based on the discrimination measures without incorporating query term weights, the weighting of expanded query terms uses our reweighting function, and

- the experiments are carried out with pseudo-relevance feedback (see Tables 8.5.1–8.5.4):
  - ① $score^*_I(t)$ and $score^*_J(t)$ show better performances than *benchmark-1* at all the evaluation points. $score^*_K(t)$ shows better performances than *benchmark-1* at most evaluation points, with some exceptions.
  - ② $score^*_I(t)$ and $score^*_J(t)$ show similar performances to *benchmark-2* at most evaluation points. $score^*_K(t)$ shows poorer performances than *benchmark-2* at most evaluation points.
- the experiments are carried out with relevance feedback (see Tables 8.6.1–8.6.4):
  - ① The three score functions achieve dramatically improved performances compared with *benchmark-1*. This is the case at all the evaluation points for all the different parts of the queries.
  - ② The three score functions obtain significantly further improved performances compared with *benchmark-2*. This is the case at all the evaluation points for all the different parts of the queries.

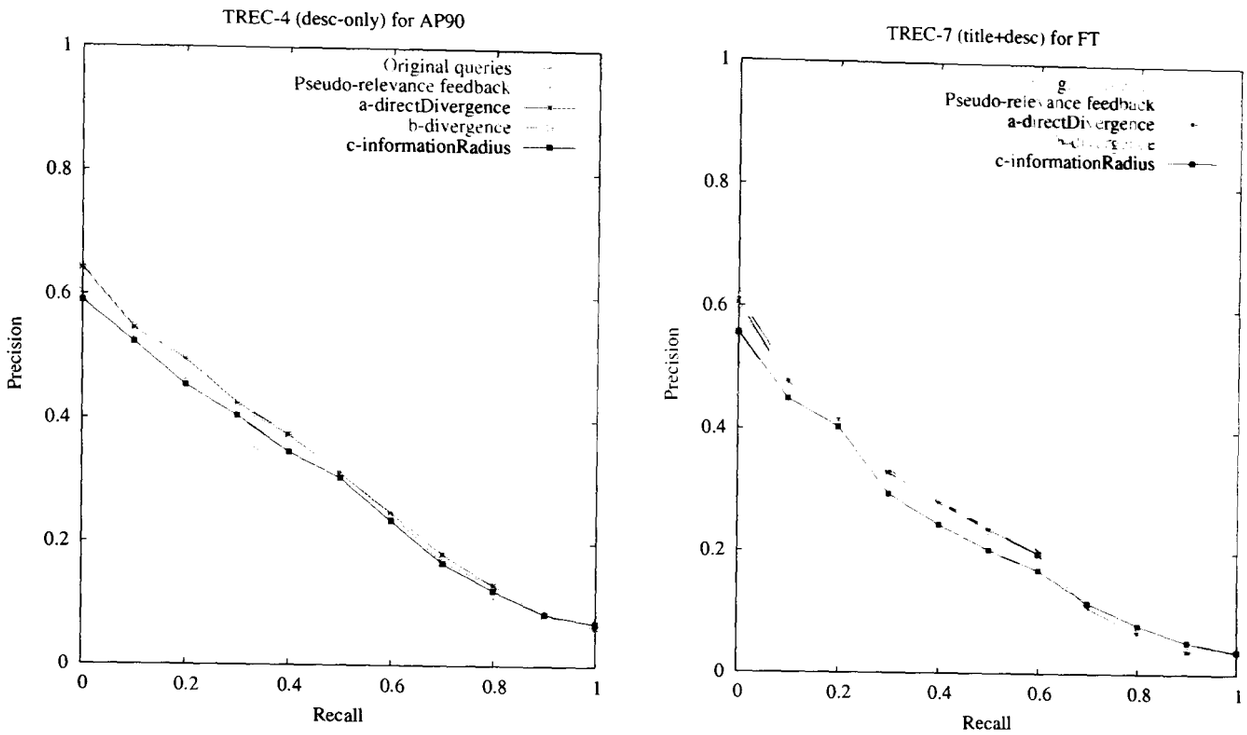

Figure 8.1: Performances of the score functions (pseudo-relevance feedback)

③ $score_I^*(t)$ and $score_J^*(t)$ gain markedly further improved performances compared with *benchmark-3* at all the evaluation points, for all the different parts of the queries. $score_K^*(t)$ gains greatly further improved performances compared with *benchmark-3* at most evaluation points, with few exceptions.

The experimental results of the three score functions are graphically shown with recall-precision curves in Figure 8.1 and Figure 8.2 for pseudo-relevance and relevance feedbacks, respectively. Also, only the curves for desc-only queries (201-250) and desc+title queries (351-400) are shown.

✳ Of particular note is that the evaluation measures at high precision points are mainly responsible for the performance improvements obtained from the three score functions: the measures of precision at-5 and at-10 are greatly increased. This is readily understandable because our query expansion methods are both recall and precision devices. This experimentally verifies that our methods are effective in improving retrieval performance.

✳ An interesting finding in our experiments is that the three functions $score_I^*(t)$, $score_J^*(t)$ and $score_K^*(t)$ exhibit consistently better performances compared with the corresponding three functions $score_I(t)$, $score_J(t)$ and $score_K(t)$. The performance increases are significant when they are used for desc+title or title-only queries on relevance feedback.

The better performances lead us to think it might be inappropriate to incorporate query term weights into the term scores for selecting good terms. This is likely to be because the information of query terms has already been incorporated into the discrimination measures, and using the information repeatedly may result in the decreased performances.

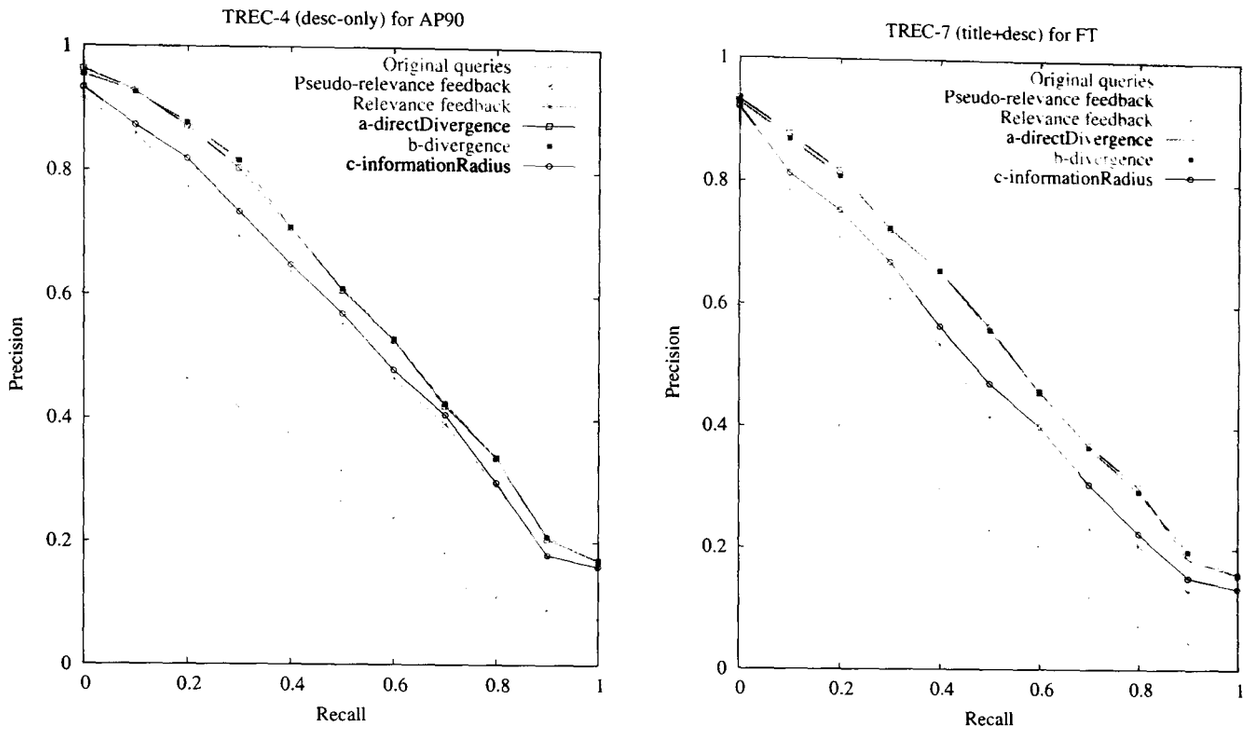

Figure 8.2: Performances of the score functions (relevance feedback)

✳ Another interesting finding in our experiments is that there is an apparent performance difference in using the two reweighting functions. Weighting expanded query terms using $rew_{IfD}(t)$ works better than using $rew_{Roc}(t)$ on pseudo-relevance feedback for scheme-4, and significantly better than using the formula on relevance feedback for scheme-3.

The better performances suggest that treating the discrimination information of terms as an important factor in weighting expanded query terms, as is done by our reweighting function, may help increase retrieval performance.

✳ Some past studies, [25, 94, 214] for instance, have shown that shorter queries may have greater performance gains. This is reinforced by our experimental results, particularly in the case of relevance feedback.

As we know, the original queries formed from the TREC topics are normally long and detailed. They are generally formulated carefully and are more elaborate than the queries users usually provide to the systems. In a realistic retrieval situation, however, it is unlikely users formulate such queries [75, 191]. Thus, it is desirable that query expansion can be particularly effective for shorter queries.

✳ It appears that the performance improvements arise from the information in the relevant sample documents, and that shorter queries show relatively greater performance gains. These observations give rise to the suggestion that we might obtain improved performances even without using query term weights to reweight expanded query terms. In practice, the suggestion is not borne out. We carried out some experiments, which ignored the weights, $w_q(t)$, of terms in our reweighting function $rew_{IfD}(t)$ in Eq.(8.1), and the retrieval performances were greatly decreased (the results are not shown in this thesis).

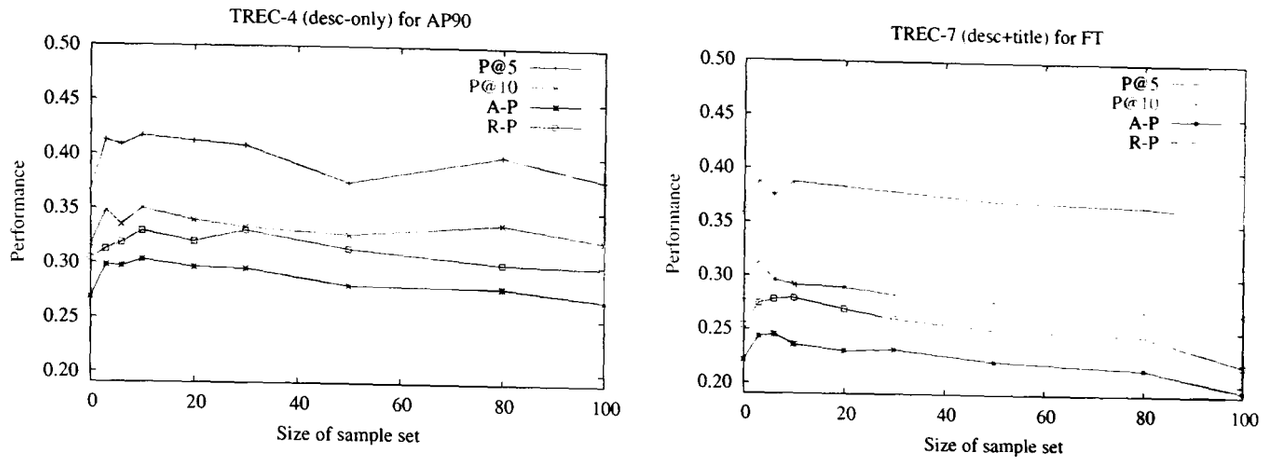

Figure 8.3: Performance vs. the size of sample set

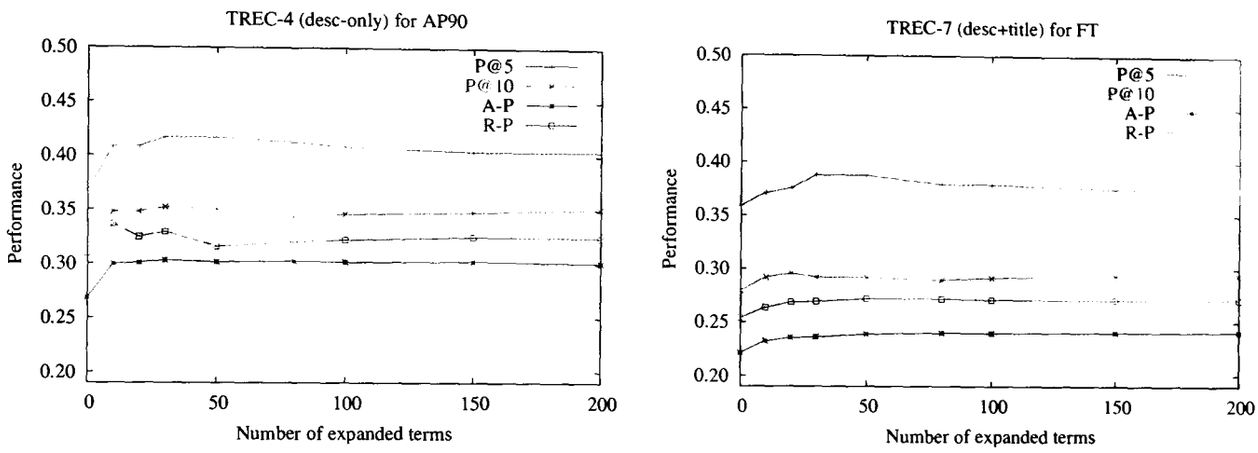

Figure 8.4: Performance vs. the number of expansion terms

* For the expanded queries derived from $score^*_I(t)$, $score^*_J(t)$ and $score^*_K(t)$ on relevance-feedback, the title-only queries work markedly better than the desc+title queries, which in turn work markedly better than the full-text queries (see Tables 8.6.2–8.6.4). These contrast strongly with the corresponding results obtained from pseudo-relevance feedback (see Tables 8.5.2–8.5.4). It seems that the query expansions obtained from these three functions suits shorter queries on relevance feedback, whereas they are more effective for longer queries on pseudo-relevance feedback.

* For the experimental investigation into the size of sample set, the results of Tables 8.7.1 and 8.7.2 are graphically shown in Figure 8.3. This view of the data suggests that 5 to 20 may be an appropriate range for the size of sample set for pseudo-relevance feedback.

  For the experimental investigation into the number of expansion terms, the results of Tables 8.7.3 and 8.7.4 are graphically shown in Figure 8.4. The graphs suggest that 10 to 40 may be an appropriate range for the number of expansion terms used on pseudo-relevance feedback.

  It is worth mentioning that using a relatively small number of expansion terms may be important to reduce response time, especially for large document collections. Harman's experiments [72, 74] showed that adding only well-selected feedback terms (e.g., 20

expansion terms) was better than adding all the candidate terms. She argued [74] that a large-scale system had a response time, and that the size of the collection had less impact on the response time than the number of query terms. The cost of adding 200 rather than 20 terms is therefore significant in respect to the response time.

* As some studies in earlier literature have shown, query expansion often has a negative effect on retrieval effectiveness, regardless of the source of candidate terms, unless relevance feedback is employed [153, 181].

  Our experimental results demonstrate that relevance feedback, with our score functions, performs significantly better than pseudo-relevance feedback (even though the complete relevance information over the collection is not available). Thus, if partial relevance information over the sample set (with a reasonable size) is available, we believe that our query expansion methods will be able to produce performance improvement in a practical IR setting.

* Any query expansion method, on pseudo-relevance feedback, may behave very differently depending on the performance of the initial retrieval run. There is ample evidence to indicate that improvement in retrieval effectiveness does not occur unless the sample set is a good one (including enough relevant documents). The negative effect of using a poor sample set for query expansion is well-known.

  A good initial performance will bring more relevant documents up to the top-ranked sample set. On pseudo-relevance feedback, there are enough relevant documents in the sample set for a good initial performance, which is likely to be further improved as a consequence of query expansion. Some studies, [23, 230] for instance, suggest that, rather than expanding all queries, one should only expand those which result in sufficient relevant documents in the sample set from the initial run. Thus, if we can know whether the sample set is a good one or not, we would be able to do a better job using query expansion on pseudo-relevance feedback. Further study of selective query expansion is needed.

Chapter 9

Summary and Further Work

This thesis is intended to give a unified account of the discrimination information of terms. It has demonstrated how the formal model $\mathcal{I}f\mathcal{D}$ deals with some basic retrieval concepts, and how the mathematical analysis is supported by empirical evidence drawn from substantial performance experiments. In this chapter, we summarize the contributions of this thesis. Some suggestions for directions of further study are then made. Finally, conclusions are drawn in general.

9.1 Summary

We have described the basic principle and idea on which the measurement of discrimination information of terms and the judgement of good terms are based. Some outstanding problems of applying information measures to AQE have been discussed. A formal model, $\mathcal{I}f\mathcal{D}$, has been established. An AQE procedure has been designed and implemented, and evaluated on two standard TREC collections. We summarize some major points of our studies from the following four aspects: the exploration of discrimination measures, the definition of association concepts, the construction of score functions and the establishment of an experimental environment.

9.1.1 Explored Discrimination Measures

- Put forward two criteria on the divergence measures.

  In Section 3.3, we put forward two criteria that underly the methodology introduced in this thesis. We pointed out that, in the context of IR, it is important for any divergence measure to satisfy these two criteria. Under these two criteria, the extent to which terms contribute to the expected divergences can be measured, and the divergence measures can be independent of the addition or removal of terms unrelated to the relevance classification. We stated that terms whose distribution is concentrated in the set of relevant sample documents make more contribution to the expected divergence and, therefore, should be interpreted as statistically containing more discrimination information than others.

- Defined a relevance discrimination measure based on directed divergence.

In Chapter 3, we studied the application of the basic concept of the directed divergence $I(P_R, P_{\bar{R}})$ to AQE. The discrimination factor $i(H_1 : H_2|t)$ was carefully examined, and was regarded as a measure of the amount of information contained in term t for discrimination in favour of relevant hypothesis H_1 against non-relevant hypothesis H_2. The discrimination measure $\mathbf{ifd}_I(t)$, which forms a basis for the methods proposed in this thesis, was formally defined. Consequently, the amount of information of terms was regarded as the power of terms to discriminate two opposite relevance hypotheses.

- Defined a relevance discrimination measure based on divergence.

In Chapter 4, we discussed a formal method of AQE based on the basic concept of divergence $J(P_R, P_{\bar{R}})$. We pointed out that a necessary condition that must be satisfied in application of the divergence is that the two term probability distributions $P_R(t)$ and $P_{\bar{R}}(t)$ should be absolutely continuous with respect to one another. Usually, the condition cannot be satisfied when we derive the distributions from different document sets. In fact, an open mathematical problem remained when applying divergence to query expansion. In Chapter 4, a possible way of solving the problem was suggested, and the solution was carefully discussed in a general form. Then, a modified discrimination measure was formally defined. Mathematical discussions on the existence of the modified discrimination measure were made by providing two typical methods of modifying the term probability distributions.

- Defined the relevance discrimination measure based on information radius.

In Chapter 5, the basic concept of the information radius $K(\lambda_1, \lambda_2; P_R, P_{\bar{R}})$ was developed as a device for formalizing the discrimination measure for AQE. An easily understood account of the concept of information radius was given. The meaning of applying information radius to measure the amount of information of terms was analysed and interpreted. We pointed out that information radius was well-defined in comparison with both $I(P_R, P_{\bar{R}})$ and $J(P_R, P_{\bar{R}})$, therefore, it might be effective to apply the information radius to query expansion in the situation where $P_R(t) \ll P_{\bar{R}}(t)$ and/or $P_{\bar{R}}(t) \ll P_R(t)$ do not hold.

- Discussed the appropriateness of using Jensen difference as a divergence measure.

Chapter 6 studied the applications of the basic concept of entropy, or entropy increase, to AQE by introducing the more general concept of Jensen difference. Three typical entropy functions were discussed, and the appropriateness of using entropy functions as divergence measures were carefully investigated. We made the following claims.

- The concavity of $H_{Sh}(P)$ is particularly useful in IR application: it provides a natural measure of the divergence between distributions $P_R(t)$ and $P_{\bar{R}}(t)$. The entropy increase $J_{H_{Sh}}(\lambda_1, \lambda_2; P_R, P_{\bar{R}})$ turns out to be the information radius $K(\lambda_1, \lambda_2; P_R, P_{\bar{R}})$, and thus the expression of its items is the discrimination measure $\mathbf{ifd}_K(t)$.
- The entropy increase measure $J_{H_{Re}}(\lambda_1, \lambda_2; P_R, P_{\bar{R}})$ does not possess Criterion 2, so it should not be an appropriate divergence measure of term distributions.
- The entropy increase measure $J_{H_{HC}}(\lambda_1, \lambda_2; P_R, P_{\bar{R}})$ possesses Criterion 2. However, the expression of its items might not be a suitable discrimination measure of terms since the expression cannot give the relationship between $P_R(t)$ and $P_{\bar{R}}(t)$ when, for instance, when $\alpha = 2$.

✠ Defined the dependence discrimination measures based on expected mutual information.

Chapter 7 focused on the formalism of the discrimination measures based on the basic concept of the expected mutual information $I_E(\delta_i, \delta_j)$. The amount of mutual information contained in a term pair (t_i, t_j) was formally interpreted. Each of the discrimination measures was corresponded to a specific state value of occurrence of the term pair. Some relationships inherent in the measures, which are important in practical IR applications, were studied. The method proposed in Chapter 7 not only cover the method EMIM given in [206, 207] as a special case, but also suggest a unified formalism for defining and estimating the mutual information of terms within a general probabilistic framework.

✠ Shown some important characteristics of the divergence measures.

The divergence measures discussed in Chapters 3, 4 and 5 addressed, in different ways, the issue of how to make estimates of the apparent difference between the term probability distributions derived from the relevant and non-relevant document sets, respectively. We showed that the strength of these divergence measures lies in their ability to provide rational estimates of the difference, and thus to capture semantic relations between terms. We claimed that the divergence measures have the following important characteristics.

- $I(P_R, P_{\bar{R}})$, $J(P_R, P_{\bar{R}})$ and $K(\lambda_1, \lambda_2; P_R, P_{\bar{R}})$ emphasize the importance of those terms with variant probabilities within sets R and $\bar{R}$, and remove the dependence on terms with invariant probabilities over both R and $\bar{R}$, since the terms would not provide profitable information for the relevance classification.
- If two term distributions overlap over some sub-domain $\Gamma \subset V^R \cap V^{\bar{R}}$, i.e., $P_R(t) = P_{\bar{R}}(t)$ for all $t \in \Gamma$, then values $I(P_R, P_{\bar{R}})$, $J(P_R, P_{\bar{R}})$ and $K(\lambda_1, \lambda_2; P_R, P_{\bar{R}})$ drop sharply. If $P_R(t) = P_{\bar{R}}(t)$ for all $t \in V$, then $I(P_R, P_{\bar{R}}) = J(P_R, P_{\bar{R}}) = K(\lambda_1, \lambda_2; P_R, P_{\bar{R}}) = 0$.
- When two term distributions $P_R(t)$ and $P_{\bar{R}}(t)$ are completely disjoint, i.e., $V^R \cap V^{\bar{R}} = \emptyset$, $K(\lambda_1, \lambda_2; P_R, P_{\bar{R}})$ is reduced to the entropy of its *a priori* probability distribution. In this case, $I(P_R, P_{\bar{R}})$ and $J(P_R, P_{\bar{R}})$ do not exist.
- $I(P_R, P_{\bar{R}})$ requires $V^R \subseteq V^{\bar{R}}$; $J(P_R, P_{\bar{R}})$ requires $V^R = V^{\bar{R}}$; $K(\lambda_1, \lambda_2; P_R, P_{\bar{R}})$ does not place any requirement on the relation between V^R and $V^{\bar{R}}$.
  In other words, $I(P_R, P_{\bar{R}})$ must satisfy $P_R(t) \ll P_{\bar{R}}(t)$ for $t \in V$; $J(P_R, P_{\bar{R}})$ must satisfy $P_R(t) \ll P_{\bar{R}}(t)$ and $P_{\bar{R}}(t) \ll P_R(t)$ for $t \in V$; it is unnecessary for $K(\lambda_1, \lambda_2; P_R, P_{\bar{R}})$ to satisfy the absolute continuity since $P_R(t)$ and $P_{\bar{R}}(t)$ are both absolutely continuous with respect to the composite distribution $P_\Sigma(t) = \lambda_1 P_R(t) + \lambda_2 P_{\bar{R}}(t)$.
- $I(P_R : P_{\bar{R}})$ is not symmetric in arguments $P_R(t)$ and $P_{\bar{R}}(t)$; $J(P_R, P_{\bar{R}})$ is symmetric in arguments $P_R(t)$ and $P_{\bar{R}}(t)$; $K(\lambda_1, \lambda_2; P_R, P_{\bar{R}})$ is not symmetric in arguments $P_R(t)$ and $P_{\bar{R}}(t)$, nor in arguments λ_1 and λ_2 (a symmetric information radius measure can be easily introduced by setting $\lambda_1 = \lambda_2 = \frac{1}{2}$).
- In the application of $K(\lambda_1, \lambda_2; P_R, P_{\bar{R}})$, *a priori* probability distribution $P_\lambda = \{\lambda_1, \lambda_2\}$ must be provided beforehand. The choice of P_λ depends on a specific model itself. There is no need to have an *a priori* probability distribution for the applications of $I(P_R, P_{\bar{R}})$ and $J(P_R, P_{\bar{R}})$

- Investigated some properties of the relevance discrimination measures.

  We investigated the relevance discrimination measures and revealed some important relationships between them, which underpin the methods proposed in this thesis:

  - $\mathbf{ifd}_I(t)$ can be positive or negative.
  - $\mathbf{ifd}_J(t) = \mathbf{ifd}_{I_{12}}(t) + \mathbf{ifd}_{I_{21}}(t)$ are always non-negative, and its two sub-items are opposite in sign, i.e., $\mathbf{ifd}_{I_{12}}(t) \cdot \mathbf{ifd}_{I_{21}}(t) \leq 0$.
  - $\mathbf{ifd}_K(t) = \lambda_1 \mathbf{ifd}_{I_{1\Sigma}}(t) + \lambda_2 \mathbf{ifd}_{I_{2\Sigma}}(t)$ can be positive or negative, and its two sub-items are opposite in sign, i.e., $\mathbf{ifd}_{I_{1\Sigma}}(t) \cdot \mathbf{ifd}_{I_{2\Sigma}}(t) \leq 0$.

  We pointed out that whether term t supports positively the relevant hypothesis H_1 depends on the relationship between distributions $P_R(t)$ and $P_{\bar{R}}(t)$, rather than on the sign of the discrimination measures (see also Table 6.3.1 in Section 6.3):

  - If $P_R(t) > P_{\bar{R}}(t)$,
    term t contributes quantity $\mathbf{ifd}_I(t) > 0$ for supporting H_1;
    term t contributes quantity $\mathbf{ifd}_J(t) > 0$ for supporting H_1;
    and if $\mathbf{ifd}_K(t) > 0$, term t contributes $\mathbf{ifd}_K(t)$ for supporting H_1.
  - If $P_R(t) < P_{\bar{R}}(t)$,
    term t contributes quantity $\mathbf{ifd}_I(t) < 0$ for supporting H_1;
    term t contributes quantity $\mathbf{ifd}_J(t) > 0$ for supporting H_2;
    and if $\mathbf{ifd}_K(t) > 0$, term t contributes $\mathbf{ifd}_K(t)$ for supporting H_2.

- Investigated some properties of the dependence discrimination measures.

  We formally defined the dependence discrimination measures corresponding to four state values of a term pair. Some relationships between the measures were revealed:

  - A single relation between $\gamma_E(t_i, t_j)$ and $\psi_E(t_i) \cdot \psi_E(t_j)$ can entirely determine all signs of $\mathbf{ifd}_M^E(t_i^{\delta_i}, t_j^{\delta_j})$ for $\delta_i, \delta_j = 1, 0$;
  - The signs of $\mathbf{ifd}_M^E(t_i, t_j)$ and $\mathbf{ifd}_M^E(\bar{t}_i, \bar{t}_j)$ are always the same, so are the signs of $\mathbf{ifd}_M^E(t_i, \bar{t}_j)$ and $\mathbf{ifd}_M^E(\bar{t}_i, t_j)$;
  - The signs of $\mathbf{ifd}_M^E(t_i, t_j)$ and $\mathbf{ifd}_M^E(\bar{t}_i, \bar{t}_j)$ are always opposite to the signs of $\mathbf{ifd}_M^E(t_i, \bar{t}_j)$ and $\mathbf{ifd}_M^E(\bar{t}_i, t_j)$.

  We pointed out that whether term t supports positively the dependent hypothesis H_1 depends on the relation between $\gamma_E(t_i, t_j)$ and $\psi_E(t_i) \cdot \psi_E(t_j)$, i.e., between $P_E(\delta_i = 1, \delta_j = 1)$ and $P_E(\delta_i = 1) \cdot P_E(\delta_j = 1)$:

  - If $P_E(\delta_i = 1, \delta_j = 1) > P_E(\delta_i = 1) \cdot P_E(\delta_j = 1)$,
    state value $(1, 1)$ contributes quantity $\mathbf{ifd}_M^E(t_i, t_j) > 0$ for supporting H_1;
    state value $(1, 0)$ contributes quantity $\mathbf{ifd}_M^E(t_i, \bar{t}_j) \leq 0$ for supporting H_1;
    state value $(0, 1)$ contributes quantity $\mathbf{ifd}_M^E(\bar{t}_i, t_j) \leq 0$ for supporting H_1;
    state value $(0, 0)$ contributes quantity $\mathbf{ifd}_M^E(\bar{t}_i, \bar{t}_j) > 0$ for supporting H_1.

- If $P_E(\delta_i = 1, \delta_j = 1) < P_E(\delta_i = 1) \cdot P_E(\delta_j = 1)$,
  state value $(1,1)$ contributes quantity $\mathbf{ifd}_M^E(t_i, t_j) \leq 0$ for supporting H_1;
  state value $(1,0)$ contributes quantity $\mathbf{ifd}_M^E(t_i, \bar{t}_j) \geq 0$ for supporting H_1;
  state value $(0,1)$ contributes quantity $\mathbf{ifd}_M^E(\bar{t}_i, t_j) \geq 0$ for supporting H_1;
  state value $(0,0)$ contributes quantity $\mathbf{ifd}_M^E(\bar{t}_i, \bar{t}_j) \leq 0$ for supporting H_1.

✠ Studied the estimation of the relevance discrimination measures.

The estimation of the term probability distributions, such as $P_{\Xi^+}(t)$ and $P_D(t)$, is crucial for effectively identifying potentially good terms from many others. In Section 3.7, some estimation methods were elaborated to embody the arguments of the relevance discrimination measures.

We showed that the term probability distribution derived from some document set, say set Ξ^+, can be estimated based on representation M_{Ξ^+}, which in turn can be estimated based on the representation M_d. Some factors should be combinatorially considered to form term weights for representing individual documents in the different documents sets of interest. The different combination schemes produce different estimations of the distributions.

✠ Studied the estimation of the dependence discrimination measures.

In Section 7.2, mathematical methods for estimating the term state distributions were developed, and three typical estimation examples were described. We pointed out that

- The marginal state distributions can be estimated based on a non-negative function, which may or may not be a term probability distribution.
- For Method A, $P_d(\delta_i, \delta_j)$ is a probability distribution if it has $p_d(t_i) \geq \gamma_d(t_i, t_j)$ and $p_d(t_j) \geq \gamma_d(t_i, t_j)$ for $t_i, t_j \in V^d$. In IR applications, it should not be a problem to satisfy these two inequalities.
- For Methods B and C, $P_d(\delta_i, \delta_j)$ is formalized by using conditional probabilities, it is naturally a probability distribution.

More importantly, a unified expression was suggested for tackling a variety of estimations of the joint and marginal state distributions embodying the arguments of the dependence discrimination measures.

✠ Analysed some properties of the estimation Methods A, B and C.

Some properties of the state distributions derived from the estimation Methods A, B and C were analysed:

- With Methods A and B, the signs of the consistent mutual information are always positive, and the signs of the inconsistent mutual information are always non-positive. Thus, their uses can assert that terms co-occurring in some document are more or less statistically dependent on one another.
- With Method C, the signs of the consistent mutual information are always opposite to the signs of the inconsistent mutual information. However, the signs of the consistent/inconsistent mutual information can be positive or negative. Therefore, if Method C is used, we cannot be sure that terms co-occurring in some sample set can be statistically dependent on each other.

- With Method C, a single relation between $\gamma_{\Xi+}(t_i, t_j)$ and $\phi_{\Xi+}(t_i)\phi_{\Xi+}(t_j)$ in the first item of $I_{\Xi+}(\delta_i, \delta_j)$ can infer the signs of discrimination measures $\mathbf{ifd}_M^{\Xi+}(t_i^{\delta_i}, t_j^{\delta_j})$ for $\delta_i, \delta_j = 1, 0$, and then determine whether terms t_i and t_j are statistically dependent under its individual state values. It is important to understand, however, that the inference and determination cannot be made from the relation between n_{11} and $n_{1\cdot}n_{\cdot 1}$ in the first item of $emim_{\Xi+}(\delta_i, \delta_j)$ since it is always non-positive.

✠ Discussed dependence of terms in a broad/narrow sense.

In Section 7.4, the notion of dependence of terms was discussed by clarifying the difference between broad and narrow dependence. We pointed out that a term very dependent on term t_j may not imply that it is the one that we are definitely interested in, even though $t_j \in V^q$ is a unique good query term.

We also pointed out that the implications of dependence for the individual measures $\mathbf{ifd}_M^E(t_i^{\delta_i}, t_j^{\delta_j})$, where $\delta_i, \delta_j = 1, 0$, are very different. Each measure corresponds to a specific state value, and it is the state value that supports the dependence.

✠ Demonstrated possible extensions to our methods.

We pointed out that the general methods proposed in this thesis can be applicable to any quantitative document representations. This means that they can be applied either to a variety of representation schemes (i.e., the weighting functions), or to a variety of representation techniques (e.g., full-text, abstract, etc). They can also be applied to the different estimation methods of the probability distributions.

We pointed out that all discussions given in Sections 7.5 and 7.6 may be applicable to a variety of information entities, such as, the sample set, local context, abstract, summary, passages, etc. All we need to do is to estimate the state distributions using the statistical data within the corresponding information entities.

9.1.2 Defined Association Concepts

✠ Put forward the Generalized Association Hypothesis.

We argued that the Association Hypothesis given in [207] (p.134) is an important underlying hypothesis in IR. Based on the Association Hypothesis, we put forward the Generalized Association Hypothesis. An essential difference between these two hypotheses is that the latter requires the association of a term with a group of 'good' terms. We claimed that this generalization is necessary for almost all methods of query expansion: an expansion term should be associated with all good query terms. We pointed out that query terms treated independently of the specific query context may cause a thorny problem in that expansion terms will be related to inappropriate meanings of query terms.

✠ Defined the concepts of association of terms with the query.

The association functions were defined as query-context-related in order to avoid increasing 'query ambiguity' caused by the ambiguity of individual query terms.

- We defined the concepts of association of terms with the context of the query in the sense of directed divergence, $atq_I(t, q)$, in Section 3.5, in the sense of divergence,

$atq_J(t,q)$, in Section 4.4, and in the sense of information radius, $atq_K(t,q)$, in Section 5.4. With function $\mathcal{Q}(t)$, the statistical information of terms appearing in the relevant sample documents (especially, that of all good query terms) was incorporated into the association functions.

A typical problem in IR is that terms often have multiple meanings. We pointed out that an effective way of handling such a problem is to use the information contained in the relevant sample documents to automatically disambiguate the ambiguous query terms. This is because terms drawn from the relevant sample documents are likely to be more related to the query context than others. Thus, they will be capable of disambiguating the query and possessing the potential power of discrimination on relevance.

- In Section 7.5, we defined the concepts of mutual association in the sense of mutual information of terms. The mutual association of a term with another term $att_M(t_i^{\delta_i}, t_j^{\delta_i})$, with the relevant sample set $ats_M(t_i^{\delta_i}, \Xi^+)$, with the query $atq_M(t_i^{\delta_i}, q)$, was discussed, and the relationships were shown. With function $\varrho(t)$, the statistical information contained in all good query terms was incorporated into the association functions.

  We pointed out that function $atq_M(t_i, q)$ takes comprehensive consideration of the association of term t_i with the context of query q. The consideration is based on the frequencies of co-occurrence of term t_i with all important query terms. Consequently, when term t_i is associated with a specific query term but not related to the query context, it is unlikely to be strongly associated with other query terms. Thus, the (total) association of term t_i with the query should be rather low. In other words, other good query terms may help to avoid the selection of term t_i as an expansion term. This may be an effective way of preventing some undesirable matches as it combinatorially considers all possible information contained in the query.

9.1.3 Constructed Score Functions

- Proposed a series of score functions.

In Sections 3.6, 4.5 and 5.5, the constructions of score functions for selecting good terms were described. We showed the general form of score functions. The general form indicates that the mathematical definition of association scores involve three essential factors: the significance of a term concerning the query, the importance of a term concerning the relevant sample set, and the discrimination information of a term concerning two opposite relevance hypotheses.

In Section 7.6, two further score functions were proposed. One uses statistical information of co-occurrence of terms, another is the same but incorporates statistical information of 'none-occurrence' of terms.

We showed that all the score functions developed in this thesis have the following characteristics:

- They are constructed based on the discrimination measures which are information-theoretic in character.

- They are rather consistent in form, and mathematically interpreted meaningfully and intuitively.
- They consider two essential aspects: (a) the informativeness of terms and, (b) the association of terms with the query context.
- They are capable of revealing some semantic relations between terms even though the informativeness and association are derived statistically.
- They may or may not rely on the relevance assessments provided by the user, i.e., they can be applied to both relevance and pseudo-relevance feedback.
- They are constructed dynamically during query processing and applied to enhance only the query.
- They are constructed automatically.
- They are implemented easily.
- They are not expensive computationally.
- They are effective experimentally.

✠ Elaborated domain reduction of the score functions.

As we have emphasized repeatedly, in order to speak of the discrimination information of terms, the arguments of the divergence measure, i.e., the term probability distributions, $P_{\Xi+}(t)$ and $P_D(t)$ (in relevance feedback), $P_{\Xi}(t)$ and $P_D(t)$ (in pseudo-relevance feedback), should be defined over the same probability space. Thus, when the discrimination measures are defined on domain V, we cannot simply say that the contributions made by terms to divergence come only from terms $t \in V^{\Xi^+}$ (or $t \in V^{\Xi}$).

However, in practice, we are indeed interested only in feedback terms. Thus, it is necessary to discuss the issue of domain reduction. In Sections 3.6, 4.5 and 5.5, we made a thorough analysis of the issue for functions $score_I(t)$, $score_J(t)$ and $score_K(t)$, respectively, and showed that it is reasonable for these three score functions to consider only feedback terms.

✠ Analysed higher positive scores.

We carefully analysed the signs of each score function. We made the following claims.

- A higher positive $score_I(t)$ can immediately infer that term t is positively associated with the query. This is because, when $t \in V^{\Xi^+}$, a positive score always indicates that term t contributes quantity $\mathrm{ifd}_I(t) > 0$ for supporting H_1.
- A higher positive $score_J(t)$ cannot infer that term t is positively associated with the query. This is because, when $t \in V^{\Xi^+}$, $\mathbf{ifd}_{I_{21}}(t)$ can be positive or negative. If $\mathbf{ifd}_{I_{21}}(t) > 0$ (and $\mathbf{ifd}_{I_{12}}(t) < 0$), then $\mathrm{ifd}_J(t) > 0$ indicates that the algebraic sum is dominated by sub-item $\mathbf{ifd}_{I_{21}}(t)$, and term t contributes quantity $\mathbf{ifd}_J(t)$ for supporting H_2.
- A higher positive $score_K(t)$ cannot infer that term t is positively associated with the query. This is because, when $t \in V^{\Xi^+} \cap V^{\Xi^-}$, it may have $\mathbf{ifd}_{I_{2\Sigma}}(t) > 0$ (and $\mathbf{ifd}_{I_{1\Sigma}}(t) < 0$), thus $\mathbf{ifd}_K(t) > 0$ indicates that the weighted algebraic sum is dominated by sub-item $\mathbf{ifd}_{I_{2\Sigma}}(t)$, and term t contributes quantity $\mathbf{ifd}_K(t)$ for supporting H_2.

We pointed out that, in practical retrieval situations, usually $P_{\Xi+}(t) > P_D(t)$ for all terms $t \in V^{\Xi+}$. Thus, for each term $t \in V^{\Xi+}$, we have $\mathbf{ifd}_{I_{12}}(t) > 0$ (and $\mathbf{ifd}_{I_{21}}(t) < 0$), and for terms with higher positive values $score_J(t)$ can immediately be inferred to contain information associated with the query. However, it may not be true that $P_{\Xi+}(t) > P_{\Xi-}(t)$ for $t \in V^{\Xi+}$ when $score_K(t)$ is used. Therefore, both conditions $P_{\Xi+}(t) > P_{\Xi-}(t)$ and $score_K(t) > 0$ have to be simultaneously verified for selecting good terms.

- Examined the relationships of the score functions.

  - In Section 4.5, $score_J(t) = score_{I_{12}}(t) + score_{I_{21}}(t)$ takes into account simultaneously opposite relevance information contained in term t. In contrast, $score_I(t) = score_{I_{12}}(t)$, discussed in Section 3.6, offers only positive association of terms with the query, but ignores negative association inherent in term t when it also appears in non-relevant documents.
  - In Section 5.5, $score_K(t) = \lambda_1 score_{I_{1\Sigma}}(t) + \lambda_2 score_{I_{2\Sigma}}(t)$ offers not only positive associations of terms with the query, but also negative associations inherent in terms when they appear in non-relevant documents.
  - In Section 7.6, the relationship between $score_{M_1}(t)$ and $score_{M_2}(t)$ was also analysed, which showed that they may not be equivalent. In contrast to $score_{M_1}(t)$, $score_{M_2}(t)$ incorporates both the consistent and inconsistent mutual information of terms into scores.

9.1.4 Presented Experimental Results

- Experimentally studied the effectiveness of estimating of term probability distributions.

  With pseudo-relevance feedback and with $rew_{IfD}(t)$, scheme-4 showed relatively stable and good performances compared with others. Also, it showed better performances than *benchmark-1* for all the different parts of the queries, and better performances than *benchmark-2* for desc+title or title-only queries.

  With relevance feedback and with $rew_{IfD}(t)(t)$, scheme-3 showed the best performances compared with others. Also, it showed significantly better performances than *benchmark-1*, *benchmark-2* and *benchmark-3* for all the different parts of the queries.

- Experimentally investigated the effectiveness of the score functions.

  $score^*_I(t)$ and $score^*_J(t)$ exhibited similar performances. These two functions and $score^*_K(t)$ were consistently most effective when used for desc+title or title-only queries.

  With pseudo-relevance feedback and with $rew_{IfD}(t)$, $score^*_I(t)$ and $score^*_J(t)$ showed markedly better performances than *benchmark-1*, and better performances than *benchmark-2*; $score^*_K(t)$ showed better performances than *benchmark-1*, but poorer performances than *benchmark-2*.

  With relevance feedback and with $rew_{IfD}(t)$, these three score functions showed significantly better performances than *benchmark-1* and *benchmark-2*. $score^*_I(t)$ and $score^*_J(t)$ gained markedly further improved performances compared with *benchmark-3* at all the evaluation points; $score^*_K(t)$ gained markedly further improved performances compared with *benchmark-3* at most evaluation points.

The measure of precision at-5 was greatly increased by these three functions on pseudo-relevance feedback; the measures of precision at-5 and at-10 were greatly increased by these three functions on relevance feedback. These experimentally verified that our methods are precision (and also recall) devices, and are effective in improving retrieval performance.

- Proposed a reweighting function for expanded query terms.

  Our reweighting function, $rew_{IfD}(t)$, emphasized both importance of query terms and association of selected terms with the query context. The weights of query terms and the scores of selected terms were properly adjusted and linearly combined to produce appropriate weights for expanded query terms.

  Experimental results showed that weighting expanded query terms using $rew_{IfD}(t)$ performs better than using $rew_{Roc}(t)$ on pseudo-relevance feedback with scheme-4, and significantly better than using the formula on relevance feedback with scheme-3. The better performances indicated that the incorporation of the discrimination information of terms into weights of expanded query terms is beneficial.

- Experimentally showed relation between query length and feedback.

  For the expanded queries obtained from $score^*_I(t)$, $score^*_J(t)$ and $score^*_K(t)$ using relevance-feedback, title-only queries worked markedly better than desc+title queries, which in turn worked markedly better than the full-text queries. These contrast strongly with the corresponding expanded queries obtained from the three functions using pseudo-relevance feedback: the full-text queries worked much better than desc+title queries, which in turn worked much better than title-only queries.

  Experimental results demonstrated that the query expansion obtained from the three functions suits shorter queries on relevance feedback, whereas it is more effective for longer queries on pseudo-relevance feedback.

9.2 Further Work

In IR, information of terms for discrimination is a perennial subject, and query representation (i.e., formulation/reformulation) is one of the central issues. Based on the discussions given in this thesis, we suggest some directions for further study.

- Choice of the discrimination measures.

  There may be some debate on the choice of a discrimination measure for comparing term probability distributions. In fact, a fairly large class of discrimination measures may lead to rather similar retrieval results. A choice of an appropriate measure for a particular application is needed. This is done by testing a range of measures against a set, whose complete relevance information is known. The measure that gives results most consistent with the available information may be chosen. Such a procedure has been followed in choosing one of the possible discrimination measures in our studies. In IR, it is of some importance that the discrimination measure should reflect, in an easily interpretable way, the divergence/difference of term probability distributions. Further

theoretical and experimental investigations are necessary to obtain more insight into these problems.

- Estimation of probability distributions.

  The estimation of probability distributions is crucial in determining retrieval performance. The best way of achieving an effective estimation of either the term probability distributions, or the term state distributions, remains an open problem, and is a significant subject for further study. We have established a theoretical framework for designing effective estimations. However, more estimation schemes/methods need to be considered.

- Investigation of effectiveness of the score functions.

  Discrimination on mutual information of terms was formally discussed in Section 7.2. Three specific estimation methods of the term state distributions were presented. Further experimental investigation into performance and comparison of the three methods need to be developed. Also, two score functions were proposed in Section 7.6, which apply the concept of query-based association. Further experiments need to be designed to test the two score functions, and find what retrieval performances can be obtained.

- Segmentation of long documents.

  The effectiveness of estimation of the discrimination measures usually lies in how documents are indexed by terms. Usually, long (full-text) documents cover multiple topics. Breaking documents into short subdocuments (such as, passages or paragraphs) may avoid problems arising from using terms from unrelated parts of documents for query expansion. Consequently, when many long documents are involved, the discrimination measures may be improved by using on information entity with appropriate units. It would be worth attempting to automatically segment long documents into subdocuments such that each of them covers only one topic. A further experimental investigation needs to be designed for estimating the discrimination measures and selecting good terms over the sample set of subdocuments.

- Multiple representations of the query.

  The user information can be represented by multiple queries. Some experimental studies, [201] for instance, have shown that multiple query representations can produce better retrieval performance than the use of a single query representation.

  Hence, a further work is to attempt multiple query representations for the information need in the $\mathcal{I}f\mathcal{D}$ framework. This can be achieved by using different score functions. Expanded queries obtained from different score functions can combinatorially be used to form new expanded queries for more effectively expressing the information need. An experimental study of such a combination method needs to be made. We want to test how these methods can work, and what levels of performance can be obtained.

- Consideration of nouns and noun phrases.

  Past studies, [94] for instance, have shown that nouns and noun phrases may be more informative than other types of terms and phrases. Thus, some further experiments need to be designed, which select only nouns and noun phrases as expansion terms for query expansion.

- Apply our methods to interactive environments.

  An alternative way to enhance queries is to interact with users. Our methods facilitate interactive query expansion for the user who is not a professional in retrieval. Once the user submits his query, the system can perform an initial retrieval, and then output a list of potential good terms that are identified by using our methods. Based on the user's judgement, some of them are selected and others are discarded. Some further experiments need to be designed for testing the interactive query expansion with real users. We want to know how our methods can work in an interactive retrieval environment, and what level of performance can be obtained.

- Test larger collection of TREC.

  We anticipate that the $\mathcal{I}f\mathcal{D}$ model can produce good results in a large-scale operational information retrieval environment. Having achieved good retrieval results in the experiments, we are encouraged to implement an IR system for accessing commercial-scale collections containing millions of documents.

9.3 Conclusions

In this thesis, we have made a thorough study of the fundamental issue of measuring the power of discrimination of terms, and interpreted the meaning of the amount of information of terms rationally and explicitly. We have established the $\mathcal{I}f\mathcal{D}$ model, and shown its ability to express a variety of term discrimination information. We have put forward the necessary criteria for the divergence measures, and analysed the properties of the discrimination measures in the context of IR. We have defined the concepts of association of terms with the context of the query, and introduced the Generalized Association Hypothesis. We have proposed a series of methods for judging good terms, and elaborated on how they can be utilized in practical feedback procedures. We have implemented a basic environment for AQE, tested retrieval performance of our methods, and compared the different query expansion methods.

We believe that the $\mathcal{I}f\mathcal{D}$ model is potentially very powerful. As we have emphasized repeatedly, $\mathcal{I}f\mathcal{D}$ is at a formal level. This makes it a complementary method to many methods in the literature, particularly, for those methods which attempt to treat terms as discriminators, and to regard the power of discrimination of terms to be essential. This also makes it possible to use $\mathcal{I}f\mathcal{D}$ as a common base, and incorporate other aspects of many different methods.

In closing, we would like to say that the advantage and promise of $\mathcal{I}f\mathcal{D}$ lie in it being an indispensable part of a textual retrieval system. Many studies of IR have shown strong interest in the issue of the power of discrimination of terms. It might be possible to have some better methodology which is capable of measuring the amount of information in terms, or it might turn out that the methodology is only a myth. But whatever the methodology is, it is most important that the IR community investigate (1) the *representation* of documents, (2) the *representation* of queries and, (3) the relevance decision over the *representations*.

> The *representations* should be formulated and reformulated based on the amount of **information** in terms **for discrimination (ifd)** on relevance.

Chapter 10

Some Mathematical Details

In some of the earlier chapters, we promised to deal with certain mathematical details which we put off then so as not to disrupt the discourse. This chapter is divided into various independent sections, each of which deals with one of the topics mentioned earlier in the thesis.

10.1 Proof of the First Inequality

Let us now prove the first inequality for both Methods I and II of modifying the discrimination measure $\mathbf{ifd}'_J(t)$ discussed in Section 4.6. That is, we need to prove that $P'_{\Xi+}(t)$ and $P'_D(t)$ can satisfy condition (C3), i.e., prove that $\mathbf{ifd}'_J(t)$ can satisfy the first inequality given in Eq.(4.3).

10.1.1 Method I

As mentioned in Section 4.6, for Method I, we need only to prove the first inequality given in Eq.(4.3). For such a purpose, we need to establish the following theorem.

Theorem 4.6.1 Let term t_0 be the argument minimum of $\mathbf{ifd}_J(t)$ over $t \in V^{\Xi+}$. If $P_{\Xi+}(t_0) \geq P_D(t_0)$, then

$$\mathbf{ifd}_J(t_0) \geq \mathbf{ifd}'_J(t_0),$$

for all terms $t \in V^{\Xi+}$.

Proof. From $P_{\Xi+}(t_0) \geq P_D(t_0)$, we have $P_{\Xi+}(t_0) - P_D(t_0) \geq 0$, and $\log \frac{P_{\Xi+}(t_0)}{P_D(t_0)} \geq 0$ (since $\frac{P_{\Xi+}(t_0)}{P_D(t_0)} \geq 1$). Then, the following inequality is easily seen:

$$\begin{aligned}\mathbf{ifd}_J(t_0) &= \big(P_{\Xi+}(t_0) - P_D(t_0)\big) \log \frac{P_{\Xi+}(t_0)}{P_D(t_0)} \geq P_{\Xi+}(t_0)\big(P_{\Xi+}(t_0) - P_D(t_0)\big) \log \frac{P_{\Xi+}(t_0)}{P_D(t_0)} \\ &= \big(P^2_{\Xi+}(t_0) - P_{\Xi+}(t_0)P_D(t_0)\big) \log \frac{P_{\Xi+}(t_0)}{P_D(t_0)} = \mathbf{ifd}'_J(t_0)\end{aligned}$$

since $0 < P_{\Xi+}(t_0) < 1$. The proof is complete.

In a real retrieval environment, it should not be a problem for satisfying *one constraint* that $P_{\Xi+}(t_0) \geq P_D(t_0)$ for $t_0 \in V^{\Xi+}$ in Theorem 4.6.1 (we have experimentally verified this

viewpoint is correct). Thus, for practical applications, we can make the assumption that, for all terms $t \in V^{\Xi+}$, we have $P_{\Xi+}(t) \geq P_D(t)$.

10.1.2 Method II

Recall that in the proof of the first inequality for Method I, we did not give any condition for factor μ since it was fixed to $\mu = P_{\Xi+}(t_0)$. However, in the proof of two inequalities for Method II, we have to give some extra conditions for factor μ. In fact, the issue of proving two inequalities for Method II is that of finding conditions that μ should satisfy. We here prove only the first inequality by Theorem 4.6.2, whereas we give the proof of the second inequality elsewhere [21].

Theorem 4.6.2 Let term t_0 be the argument minimum of $\mathbf{ifd}_J(t)$ over $t \in V^{\Xi+}$. If $P_{\Xi+}(t_0) > P_D(t_0)$ then there exists a real number μ satisfying $1 > \mu > 0$ and $\mu P_{\Xi+}(t_0) \geq P_D(t_0)$, such that

$$\mathbf{ifd}_J(t_0) \geq \mathbf{ifd}'_J(t_0),$$

for all terms $t \in V^{\Xi+}$.

Proof. If $P_{\Xi+}(t_0) > P_D(t_0)$, then $1 > \frac{P_D(t_0)}{P_{\Xi+}(t_0)}$. According to the theory of real numbers, there are a large amount of real numbers (non-numerally infinite) μ in the real number interval $[\frac{P_D(t_0)}{P^+_{\Xi}(t_0)}, 1)$, i.e., $0 < \frac{P_D(t_0)}{P^+_{\Xi}(t_0)} \leq \mu < 1$. Each of these non-numerally infinite μ certainly satisfy $\mu P_{\Xi+}(t_0) \geq P_D(t_0)$ and $1 > \mu > 0$.

Then, from $\mu P_{\Xi+}(t_0) \geq P_D(t_0)$, we have

$$\frac{\mu P_{\Xi+}(t_0)}{P_D(t_0)} \geq 1, \quad \text{i.e.,} \quad \log \frac{\mu P_{\Xi+}(t_0)}{P_D(t_0)} \geq 0.$$

Also, from $0 < \mu < 1$, we have

$$\frac{P_{\Xi+}(t_0)}{P_D(t_0)} > \frac{\mu P_{\Xi+}(t_0)}{P_D(t_0)}, \quad \text{i.e.,} \quad \log \frac{P_{\Xi+}(t_0)}{P_D(t_0)} > \log \frac{\mu P_{\Xi+}(t_0)}{P_D(t_0)} \geq 0.$$

On the other hand, we have

$$P_{\Xi+}(t_0) > \mu P_{\Xi+}(t_0) \geq P_D(t_0),$$

i.e.,

$$P_{\Xi+}(t_0) - P_D(t_0) > \mu P_{\Xi+}(t_0) - P_D(t_0) \geq 0.$$

Thus, we find that

$$\big(P_{\Xi+}(t_0) - P_D(t_0)\big) \log \frac{P_{\Xi+}(t_0)}{P_D(t_0)} > \big(\mu P_{\Xi+}(t_0) - P_D(t_0)\big) \log \frac{\mu P_{\Xi+}(t_0)}{P_D(t_0)}.$$

The proof is complete.

Notice that Theorem 4.6.2 requires $P_{\Xi+}(t_0) > P_D(t_0)$, rather than $P_{\Xi+}(t_0) \geq P_D(t_0)$ as required in Theorem 4.6.1 in the first modification method. This is because $P_{\Xi+}(t_0) = P_D(t_0)$ would lead to contradiction $1 > \mu \geq 1$ from $1 > \mu \geq \frac{P_D(t_0)}{P^+_{\Xi}(t_0)} = 1$.

10.2 Discussion on Symmetric Discrimination Measure

Let us now return to Section 5.3 and consider the symmetric discrimination measure again. Notice that, specializing the divergence measure $K(\lambda_1, \lambda_2; P_R, P_{\bar{R}})$ to $\lambda_1 = \lambda_2 = \frac{1}{2}$, we can obtain

$$\begin{aligned} K(P_R, P_{\bar{R}}) &= K(\tfrac{1}{2}, \tfrac{1}{2}; P_R, P_{\bar{R}}) \\ &= \sum_{t \in V} \Big(\frac{1}{2} P_R(t) \log \frac{P_R(t)}{\frac{1}{2}P_R(t) + \frac{1}{2}P_{\bar{R}}(t)} + \frac{1}{2} P_{\bar{R}}(t) \log \frac{P_{\bar{R}}(t)}{\frac{1}{2}P_R(t) + \frac{1}{2}P_{\bar{R}}(t)} \Big) \\ &= \frac{1}{2} \sum_{t \in V} \Big(P_R(t) \log 2 + P_{\bar{R}}(t) \log 2 + \phi(t) \Big) \\ &= \frac{1}{2} \sum_{t \in V} \Big(P_R(t) + P_{\bar{R}}(t) + \phi(t) \Big), \end{aligned}$$

which is symmetric with respect to $P_R(t)$ and $P_{\bar{R}}(t)$, where

$$\phi(t) = P_R(t) \log \frac{P_R(t)}{P_R(t) + P_{\bar{R}}(t)} + P_{\bar{R}}(t) \log \frac{P_{\bar{R}}(t)}{P_R(t) + P_{\bar{R}}(t)}.$$

It is worth mentioning that divergence $K(P_R, P_{\bar{R}})$ cannot be reduced to

$$\Phi(t) = \sum_{t \in V} \phi(t) = \sum_{t \in V} \Big(\phi_1(t) + \phi_2(t) \Big)$$

by considering

$$\begin{aligned} K(P_R, P_{\bar{R}}) &= \frac{1}{2} \sum_{t \in V} \Big(P_R(t) + P_{\bar{R}}(t) + \phi(t) \Big) = \frac{1}{2} \Big(\sum_{t \in V} P_R(t) + \sum_{t \in V} P_{\bar{R}}(t) + \sum_{t \in V} \phi(t) \Big) \\ &= 1 + \frac{1}{2} \sum_{t \in V} \phi(t) = 1 + \frac{1}{2} \Phi(t), \end{aligned}$$

and eliminating coefficients 1 and $\frac{1}{2}$ in the last expression. In fact, summation $\Phi(t)$ cannot service as a divergence measure. First, summation $\Phi(t)$ is non-positive. This is because, for each term $t \in V$, two sub-items $\phi_1(t)$ and $\phi_2(t)$ in each item are both non-positive since $P_R(t), P_{\bar{R}}(t) \leq P_R(t) + P_{\bar{R}}(t)$, and so is the item $\phi(t)$ itself. Thus, the summation over individual items receives a non-positive value. We can also easily verify $\Phi(t) \leq 0$ from $0 \leq K(P_R, P_{\bar{R}}) = 1 + \frac{1}{2}\Phi(t) \leq 1$ (we have shown $0 \leq K(\lambda_1, \lambda_2; P_R, P_{\bar{R}}) \leq 1$ for a general case). Second, summation $\Phi(t)$ is dependent of the introduction or elimination of terms unrelated to the relevance classification. That is, its items

$$\begin{aligned} \phi(t) &= P_R(t) \log \frac{P_R(t)}{P_R(t) + P_{\bar{R}}(t)} + P_{\bar{R}}(t) \log \frac{P_{\bar{R}}(t)}{P_R(t) + P_{\bar{R}}(t)} \\ &= P_R(t) \log \frac{1}{2} + P_{\bar{R}}(t) \log \frac{1}{2} = -P_R(t) - P_{\bar{R}}(t) \neq 0 \end{aligned}$$

do not vanish when $P_R(t) = P_{\bar{R}}(t) \neq 0$. Thus, $\Phi(t)$ does not possess Criterion 2.

10.3 Jensen Difference

In Section 6.2, we discussed the definition of Jensen difference. We now derive the corresponding Jensen difference for three entropy functions given in Section 6.1.

10.3.1 Entropy Function H_{Sh}

For Shannon's entropy, we have

$$\begin{aligned}
H_{Sh}\Big(\sum_{k=1}^{r}\lambda_k P_{D_k}\Big) &= -\sum_{t\in V}\Big(\big(\sum_{k=1}^{r}\lambda_k P_{D_k}(t)\big)\log\big(\sum_{k=1}^{r}\lambda_k P_{D_k}(t)\big)\Big)\\
&= \sum_{t\in V}\Big(\sum_{k=1}^{r}\big(\lambda_k P_{D_k}(t)\log\frac{1}{\sum_{k=1}^{r}\lambda_k P_{D_k}(t)} + \lambda_k P_{D_k}(t)\log P_{D_k}(t) - \lambda_k P_{D_k}(t)\log P_{D_k}(t)\big)\Big)\\
&= \sum_{t\in V}\Big(\sum_{k=1}^{r}\lambda_k P_{D_k}(t)\log\frac{P_{D_k}(t)}{\sum_{k=1}^{r}\lambda_k P_{D_k}(t)} - \sum_{k=1}^{r}\lambda_k P_{D_k}(t)\log P_{D_k}(t)\Big)\\
&= \sum_{t\in V}\sum_{k=1}^{r}\lambda_k P_{D_k}(t)\log\frac{P_{D_k}(t)}{\sum_{k=1}^{r}\lambda_k P_{D_k}(t)} + \sum_{k=1}^{r}\lambda_k H_{Sh}(P_{D_k}).
\end{aligned}$$

Then, the Jensen difference with respect to Shannon's entropy can be written as

$$\begin{aligned}
J_{H_{Sh}}(\{\lambda_k\};\{P_{D_k}\}) &= H_{Sh}\Big(\sum_{k=1}^{r}\lambda_k P_{D_k}\Big) - \sum_{k=1}^{r}\lambda_k H_{Sh}(P_{D_k})\\
&= \sum_{t\in V}\sum_{k=1}^{r}\lambda_k P_{D_k}(t)\log\frac{P_{D_k}(t)}{\sum_{k=1}^{r}\lambda_k P_{D_k}(t)},
\end{aligned}$$

which is the *information radius* defined by Sibson [177].

10.3.2 Entropy Function H_{Re}

For Rényi's entropy, when $0 < \alpha < 1$, the Jensen difference can be written as

$$\begin{aligned}
J_{H_{Re}}(\{\lambda_k\};\{P_{D_k}\}) &= H_{Re}\Big(\sum_{k=1}^{r}\lambda_k P_{D_k}\Big) - \sum_{k=1}^{r}\lambda_k H_{Re}(P_{D_k})\\
&= \frac{1}{1-\alpha}\log\Big(\sum_{t\in V}\big(\sum_{k=1}^{r}\lambda_k P_{D_k}(t)\big)^{\alpha}\Big) - \sum_{k=1}^{r}\Big(\frac{\lambda_k}{1-\alpha}\log\big(\sum_{t\in V}P_{D_k}^{\alpha}(t)\big)\Big)\\
&= \frac{1}{1-\alpha}\log\Big(\sum_{t\in V}\big(\sum_{k=1}^{r}\lambda_k P_{D_k}(t)\big)^{\alpha}\Big) - \frac{1}{1-\alpha}\log\Big(\prod_{k=1}^{r}\big(\sum_{t\in V}P_{D_k}^{\alpha}(t)\big)^{\lambda_k}\Big)\\
&= \frac{1}{1-\alpha}\log\frac{\sum_{t\in V}\big(\sum_{k=1}^{r}\lambda_k P_{D_k}(t)\big)^{\alpha}}{\prod_{k=1}^{r}\big(\sum_{t\in V}P_{D_k}^{\alpha}(t)\big)^{\lambda_k}}.
\end{aligned}$$

Notice that, when $\alpha > 1$, $J_{H_{Re}}(\{\lambda_k\};\{P_{D_k}\})$ may be negative since Rényi's entropy is not in general concave for $\alpha > 1$. This appears to be a disadvantage of entropy H_{Re} as a measure of diversity.

10.3.3 Entropy Function H_{HC}

For the entropy of Havrda and Charvát, when $\alpha > 0$ ($\alpha \neq 1$), we have

$$H_{HC}\Big(\sum_{k=1}^{r}\lambda_k P_{D_k}\Big) = \frac{1}{1-2^{1-\alpha}}\Big[1-\sum_{t\in V}\big(\sum_{k=1}^{r}\lambda_k P_{D_k}(t)\big)^{\alpha}\Big]$$

$$= \frac{1}{1-2^{1-\alpha}}\Big[1-\sum_{t\in V}\big(\sum_{k=1}^{r}\lambda_k P^{\alpha}_{D_k}(t)\big)+\sum_{t\in V}\big(\sum_{k=1}^{r}\lambda_k P^{\alpha}_{D_k}(t)\big)-\sum_{t\in V}\big(\sum_{k=1}^{r}\lambda_k P_{D_k}(t)\big)^{\alpha}\Big]$$

$$= \frac{1}{1-2^{1-\alpha}}\Big[\big(\sum_{k=1}^{r}\lambda_k\big)-\sum_{k=1}^{r}\big(\sum_{t\in V}\lambda_k P^{\alpha}_{D_k}(t)\big)+\sum_{t\in V}\Big(\big(\sum_{k=1}^{r}\lambda_k P^{\alpha}_{D_k}(t)\big)-\big(\sum_{k=1}^{r}\lambda_k P_{D_k}(t)\big)^{\alpha}\Big)\Big]$$

$$= \frac{1}{1-2^{1-\alpha}}\Big[\sum_{k=1}^{r}\lambda_k\big(1-\sum_{t\in V}P^{\alpha}_{D_k}(t)\big)+\sum_{t\in V}\Big(\big(\sum_{k=1}^{r}\lambda_k P^{\alpha}_{D_k}(t)\big)-\big(\sum_{k=1}^{r}\lambda_k P_{D_k}(t)\big)^{\alpha}\Big)\Big]$$

$$= \sum_{k=1}^{r}\lambda_k H_{HC}(P_{D_k})+\frac{1}{1-2^{1-\alpha}}\sum_{t\in V}\Big(\big(\sum_{k=1}^{r}\lambda_k P^{\alpha}_{D_k}(t)\big)-\big(\sum_{k=1}^{r}\lambda_k P_{D_k}(t)\big)^{\alpha}\Big).$$

Then, the Jensen difference with respect to the entropy of Havrda and Charvát can be written as follows.

$$J_{H_{HC}}(\{\lambda_k\};\{P_{D_k}\}) = H_{HC}\Big(\sum_{k=1}^{r}\lambda_k P_{D_k}\Big)-\sum_{k=1}^{r}\lambda_k H_{HC}(P_{D_k})$$

$$= \frac{1}{1-2^{1-\alpha}}\sum_{t\in V}\Big(\big(\sum_{k=1}^{r}\lambda_k P^{\alpha}_{D_k}(t)\big)-\big(\sum_{k=1}^{r}\lambda_k P_{D_k}(t)\big)^{\alpha}\Big).$$

10.4 Proofs of Some Theorems

This section proves a series of useful theorems which were put forward in Chapter 7.

10.4.1 Proof of Theorem 7.2.1

We now prove that the estimation, in Method A, of $P_d(\delta_i,\delta_j)$ given in Eq.(7.5) constitutes a probability distribution over $\Omega\times\Omega$ by establishing the following theorem.

Theorem 7.2.1 For arbitrary terms $t_j, t_j \in V^d$, and for the expression given in Eq.(7.5), we have:

(1) $p_d(t_i) \geq \gamma_A(t_i,t_j)$ if and only if

$$\sum_{i'<j';\,t_{i'},t_{j'}\in V^d-\{t_j\}} f_d(t_{i'})f_d(t_{j'}) \geq f_d(t_j)f_d(t_j);$$

(2) $p_d(t_j) \geq \gamma_A(t_i,t_j)$ if and only if

$$\sum_{i'<j';\,t_{i'},t_{j'}\in V^d-\{t_i\}} f_d(t_{i'})f_d(t_{j'}) \geq f_d(t_i)f_d(t_i).$$

Proof. We only prove (1). The proof of (2) is similar to (1).

For document d, let $V^d = \{t_{i_1}, t_{i_2}, ..., t_{i_s}\} \subseteq \{t_1, t_2, ..., t_n\} = V$, where $1 \leq i_1 < i_2 < ... < i_s \leq n$, and $|V^d| = s \geq 2$. Without losing generality, let us suppose $t_j = t_{i_1}$ (Otherwise, let $t_j = t_{i_h}$. Notice the order of the elements in the set is unnecessary, so we can rewrite $V^d = \{t_{i_1}, t_{i_2}, ..., t_{i_{h-1}}, t_{i_h}, t_{i_{h+1}}, ..., t_{i_s}\}$ by $V^d = \{t_{i_h}, t_{i_1}, t_{i_2}, ..., t_{i_{h-1}}, t_{i_{h+1}}, ..., t_{i_s}\}$. we thus have $V^d - \{t_j\} = \{t_{i_1}, t_{i_2}, ..., t_{i_{h-1}}, t_{i_{h+1}}, ..., t_{i_s}\}$ with $1 \leq i_1 < i_2 < ... < i_{h-1} < i_{h-2} < ... < i_s \leq n$. So the discussion below still holds.)

Denote the denominator of probability $P_d(\delta_i = 1, \delta_j = 1)$ by

$$\varpi = \sum_{i'<j'; t_{i'}, t_{j'} \in V^d} f_d(t_{i'}) f_d(t_{j'}),$$

which is the sum of products $f_d(t_{i'}) f_d(t_{j'})$ for $i' < j'$; $i', j' \in \{i_1, i_2, ..., i_s\}$. Then,

$$\begin{aligned}
\varpi &= f_d(t_{i_1})\big[f_d(t_{i_2}) + f_d(t_{i_3}) + ... + f_d(t_{i_s})\big] + f_d(t_{i_2})\big[f_d(t_{i_3}) + f_d(t_{i_4}) + ... + f_d(t_{i_s})\big] \\
&\quad + ... + f_d(t_{i_{s-2}})\big[f_d(t_{i_{s-1}}) + f_d(t_{i_s})\big] + f_d(t_{i_{s-1}})\big[f_d(t_{i_s})\big] \\
&= f_d(t_j)\big[||d|| - f_d(t_j)\big] + f_d(t_{i_2}) \sum_{j'=i_3,...,i_s} f_d(t_{j'}) + ... + f_d(t_{i_{s-1}}) \sum_{j'=i_s} f_d(t_{j'}) \\
&= ||d|| f_d(t_j) - f_d(t_j) f_d(t_j) + \sum_{i'<j'; t_{i'}, t_{j'} \in V^d - \{t_j\}} f_d(t_{i'}) f_d(t_{j'}).
\end{aligned}$$

Therefore, we have

$$\varpi_j = \sum_{i'<j'; t_{i'}, t_{j'} \in V^d - \{t_j\}} f_d(t_{i'}) f_d(t_{j'}) \geq f_d(t_j) f_d(t_j)$$

if and only if

$$-f_d(t_j) f_d(t_j) + \varpi_j \geq 0,$$

i.e.,

$$\varpi = ||d|| f_d(t_j) - f_d(t_j) f_d(t_j) + \varpi_j \geq ||d|| f_d(t_j),$$

i.e.,

$$p_d(t_i) = \frac{f_d(t_i)}{||d||} \geq \frac{f_d(t_i) f_d(t_j)}{\varpi} = \gamma_A(t_i, t_j).$$

The proof is complete.

10.4.2 Proof of Theorem 7.2.3

We now analyse the absolute continuity of the estimation, in Method C, of distribution $P_{\Xi^+}(\delta_i, \delta_j)$ with respect to the product, $P_{\Xi^+}(\delta_i) \cdot P_{\Xi^+}(\delta_j)$, of the marginal distributions by proving the following theorem.

Theorem 7.2.3 For arbitrary terms $t_i, t_j \in V^{\Xi^+}$, $P_{\Xi^+}(\delta_i, \delta_j) \ll P_{\Xi^+}(\delta_i) \cdot P_{\Xi^+}(\delta_j)$ for $\delta_i, \delta_j = 1, 0$.

Proof. According to whether $\phi_{\Xi^+}(t_i) = 1$ and/or $\phi_{\Xi^+}(t_j) = 1$, there are four cases to be considered, that is,

(c1) $0 < \phi_{\Xi^+}(t_i) < 1$ and $0 < \phi_{\Xi^+}(t_j) < 1$,

(c2) $\phi_{\Xi^+}(t_i) = 1$ but $0 < \phi_{\Xi^+}(t_j) < 1$,

(c3) $0 < \phi_{\Xi^+}(t_i) < 1$ but $\phi_{\Xi^+}(t_j) = 1$,

(c4) $\phi_{\Xi^+}(t_i) = 1$ and $\phi_{\Xi^+}(t_j) = 1$.

We here prove only (c2). For (c1), see the discussion for the unified expression given in Section 7.2. For (c3) and (c4), the discussion is similar to (c2).

Suppose that we are given terms $t_i, t_j \in V^{\Xi^+}$ satisfying $F_{\Xi^+}(t_i) = |\Xi^+|$ and $F_{\Xi^+}(t_j) < |\Xi^+|$ (namely t_i occurs in all relevant sample documents, but t_j does not). In this case, it has $\phi_{\Xi^+}(t_i) = 1$ and $0 < \phi_{\Xi^+}(t_j) < 1$, and $F_{\Xi^+}(t_j) = F_{\Xi^+}(t_i, t_j)$. Therefore:

(a) We have $P_{\Xi^+}(\delta_i = 1) \cdot P_{\Xi^+}(\delta_j = 1) > 0$ since $P_{\Xi^+}(\delta_i = 1) = 1$, and $0 < P_{\Xi^+}(\delta_j = 1) < 1$. Thus, $P_{\Xi^+}(\delta_i = 1, \delta_j = 1) \ll P_{\Xi^+}(\delta_i = 1) \cdot P_{\Xi^+}(\delta_j = 1)$ for $(\delta_i, \delta_j) = (1, 1)$.

(b) We have $P_{\Xi^+}(\delta_i = 1) \cdot P_{\Xi^+}(\delta_j = 0) > 0$ since $P_{\Xi^+}(\delta_i = 1) = 1$ and $0 < P_{\Xi^+}(\delta_j = 0) < 1$. Thus, $P_{\Xi^+}(\delta_i = 1, \delta_j = 0) \ll P_{\Xi^+}(\delta_i = 1) \cdot P_{\Xi^+}(\delta_j = 0)$ for $(\delta_i, \delta_j) = (1, 0)$.

(c) We have $P_{\Xi^+}(\delta_i = 0) \cdot P_{\Xi^+}(\delta_j = 1) = 0$ since $P_{\Xi^+}(\delta_i = 0) = 0$ and $0 < P_{\Xi^+}(\delta_j = 1) < 1$. Also, we have $P_{\Xi^+}(\delta_i = 0, \delta_j = 1) = \frac{1}{|\Xi^+|}\left[F_{\Xi^+}(t_j) - F_{\Xi^+}(t_i, t_j)\right] = 0$. Thus, $P_{\Xi^+}(\delta_i = 0, \delta_j = 1) \ll P_{\Xi^+}(\delta_i = 0) \cdot P_{\Xi^+}(\delta_j = 1)$ for $(\delta_i, \delta_j) = (0, 1)$. (

(d) We have $P_{\Xi^+}(\delta_i = 0) \cdot P_{\Xi^+}(\delta_j = 0) = 0$ since $P_{\Xi^+}(\delta_i = 0) = 0$ and $0 < P_{\Xi^+}(\delta_j = 0) < 1$. Also, we have $P_{\Xi^+}(\delta_i = 0, \delta_j = 0) = \frac{1}{|\Xi^+|}\left[|\Xi^+| - F_{\Xi^+}(t_i) - F_{\Xi^+}(t_j) + F_{\Xi^+}(t_i, t_j)\right] = \frac{1}{|\Xi^+|}\left[(|\Xi^+| - F_{\Xi^+}(t_i)) - (F_{\Xi^+}(t_j) - F_{\Xi^+}(t_i, t_j))\right] = 0$. Thus, $P_{\Xi^+}(\delta_i = 0, \delta_j = 0) \ll P_{\Xi^+}(\delta_i = 0) \cdot P_{\Xi^+}(\delta_j = 0)$ for $(\delta_i, \delta_j) = (0, 0)$.

Consequently, $P_{\Xi^+}(\delta_i, \delta_j) \ll P_{\Xi^+}(\delta_i) \cdot P_{\Xi^+}(\delta_j)$ for $\delta_i, \delta_j = 1, 0$. The proof is complete.

10.4.3 Proof of Theorem 7.2.4

Let us look at an interesting property of measure $emim_{\Xi^+}(\delta_i, \delta_j)$, discussed in Method C, by considering the following theorem.

Theorem 7.2.4 For arbitrary terms $t_i, t_j \in V^{\Xi^+}$, $emim_{\Xi^+}(\delta_i, \delta_j) \leq 0$.

Proof. Let us prove that the individual items of $emim_{\Xi^+}(\delta_i, \delta_j)$ are non-positive. For this purpose, we need prove (1) $\frac{n_{11}}{n_{1\cdot}n_{\cdot 1}} \leq 1$, (2) $\frac{n_{10}}{n_{1\cdot}n_{\cdot 0}} \leq 1$, (3) $\frac{n_{01}}{n_{0\cdot}n_{\cdot 1}} \leq 1$, and (4) $\frac{n_{00}}{n_{0\cdot}n_{\cdot 0}} \leq 1$.

(1) From $F_{\Xi^+}(t_i, t_j) \leq F_{\Xi^+}(t_i)$ and $F_{\Xi^+}(t_i, t_j) \leq F_{\Xi^+}(t_j)$, we obtain immediately

$$\frac{n_{11}}{n_{1\cdot}n_{\cdot 1}} = \frac{F_{\Xi^+}(t_i, t_j)}{F_{\Xi^+}(t_i)F_{\Xi^+}(t_j)} \leq 1.$$

(2) From $F_{\Xi^+}(t_i) - F_{\Xi^+}(t_i, t_j) \leq F_{\Xi^+}(t_i)$ and

$$F_{\Xi^+}(t_i) - F_{\Xi^+}(t_i, t_j) \leq |\Xi^+| - F_{\Xi^+}(t_i, t_j) \leq |\Xi^+| - F_{\Xi^+}(t_j),$$

it follows

$$\frac{n_{10}}{n_{1\cdot}n_{\cdot 0}} = \frac{F_{\Xi^+}(t_i) - F_{\Xi^+}(t_i, t_j)}{F_{\Xi^+}(t_i)\left(|\Xi^+| - F_{\Xi^+}(t_j)\right)} \leq 1.$$

(3) The proof is similar to (2).

(4) Notice that

$$\begin{aligned}&|\Xi^+| - F_{\Xi+}(t_i) - F_{\Xi+}(t_j) + F_{\Xi+}(t_i, t_j)\\ &= |\Xi^+| - F_{\Xi+}(t_i) - \left[F_{\Xi+}(t_j) - F_{\Xi+}(t_i, t_j)\right] \leq |\Xi^+| - F_{\Xi+}(t_i),\end{aligned}$$

$$\begin{aligned}&|\Xi^+| - F_{\Xi+}(t_i) - F_{\Xi+}(t_j) + F_{\Xi+}(t_i, t_j)\\ &= |\Xi^+| - F_{\Xi+}(t_j) - \left[F_{\Xi+}(t_i) - F_{\Xi+}(t_i, t_j)\right] \leq |\Xi^+| - F_{\Xi+}(t_j).\end{aligned}$$

From which we have

$$\frac{n_{00}}{n_{0\cdot}n_{\cdot 0}} = \frac{|\Xi^+| - F_{\Xi+}(t_i) - F_{\Xi+}(t_j) + F_{\Xi+}(t_i, t_j)}{\left(|\Xi^+| - F_{\Xi+}(t_i)\right)\left(|\Xi^+| - F_{\Xi+}(t_j)\right)} \leq 1.$$

The proof is complete.

10.4.4 Proof of Theorem 7.2.5

Let us prove Theorem 7.2.5, which reveals a property of measure $I_{\Xi+}(\delta_i, \delta_j)$ that, when it is estimated with Method C and one of the terms t_i and t_j occurs in all sample documents, terms t_i and t_j would not provide any information on relevance.

Theorem 7.2.5 For arbitrary terms $t_i, t_j \in V^{\Xi+}$, if $F_{\Xi+}(t_i) = |\Xi^+|$ then $I_{\Xi+}(\delta_i, \delta_j) = 0$.

Proof. Let us prove that the individual items of $I_{\Xi+}(\delta_i, \delta_j)$ are zero. Notice that, from $F_{\Xi+}(t_i) = |\Xi^+|$, we have $F_{\Xi+}(t_j) = F_{\Xi+}(t_i, t_j)$, Thus,

(1) for $(\delta_i, \delta_j) = (1, 1)$, we have

$$\frac{F_{\Xi+}(t_i, t_j)}{|\Xi^+|} \log \frac{\frac{F_{\Xi+}(t_i, t_j)}{|\Xi^+|}}{\frac{F_{\Xi+}(t_i)}{|\Xi^+|}\frac{F_{\Xi+}(t_j)}{|\Xi^+|}} = \frac{F_{\Xi+}(t_i, t_j)}{|\Xi^+|} \log \frac{F_{\Xi+}(t_i, t_j)}{1 \times F_{\Xi+}(t_j)} = \frac{F_{\Xi+}(t_i, t_j)}{|\Xi^+|} \log 1 = 0;$$

(2) for $(\delta_i, \delta_j) = (1, 0)$, we have

$$\begin{aligned}&\frac{F_{\Xi+}(t_i) - F_{\Xi+}(t_i, t_j)}{|\Xi^+|} \log \frac{\frac{F_{\Xi+}(t_i) - F_{\Xi+}(t_i, t_j)}{|\Xi^+|}}{\frac{F_{\Xi+}(t_i)}{|\Xi^+|}\left(1 - \frac{F_{\Xi+}(t_j)}{|\Xi^+|}\right)}\\ &= \frac{F_{\Xi+}(t_i) - F_{\Xi+}(t_i, t_j)}{|\Xi^+|} \log \frac{F_{\Xi+}(t_i) - F_{\Xi+}(t_i, t_j)}{1 \times \left(|\Xi^+| - F_{\Xi+}(t_j)\right)}\\ &= \frac{F_{\Xi+}(t_i) - F_{\Xi+}(t_i, t_j)}{|\Xi^+|} \log \frac{|\Xi^+| - F_{\Xi+}(t_j)}{|\Xi^+| - F_{\Xi+}(t_j)}\\ &= \frac{F_{\Xi+}(t_i) - F_{\Xi+}(t_i, t_j)}{|\Xi^+|} \log 1 = 0;\end{aligned}$$

(3) for $(\delta_i, \delta_j) = (0, 1)$, we have

$$\begin{aligned}&\frac{F_{\Xi+}(t_j) - F_{\Xi+}(t_i, t_j)}{|\Xi^+|} \log \frac{\frac{F_{\Xi+}(t_j) - F_{\Xi+}(t_i, t_j)}{|\Xi^+|}}{\left(1 - \frac{F_{\Xi+}(t_i)}{|\Xi^+|}\right)\frac{F_{\Xi+}(t_j)}{|\Xi^+|}}\\ &= \frac{F_{\Xi+}(t_j) - F_{\Xi+}(t_i, t_j)}{|\Xi^+|} \log \frac{F_{\Xi+}(t_j) - F_{\Xi+}(t_i, t_j)}{(1 - 1) \times F_{\Xi+}(t_j)}\end{aligned}$$

$$= \frac{0}{|\Xi^+|}\log\frac{0}{0 \times F_{\Xi^+}(t_j)} = 0\log\frac{0}{0} = 0;$$

(4) for $(\delta_i, \delta_j) = (0,0)$, we have

$$\begin{aligned}
&\frac{|\Xi^+| - F_{\Xi^+}(t_i) - F_{\Xi^+}(t_j) + F_{\Xi^+}(t_i,t_j)}{|\Xi^+|}\log\frac{\frac{|\Xi^+|-F_{\Xi^+}(t_i)-F_{\Xi^+}(t_j)+F_{\Xi^+}(t_i,t_j)}{|\Xi^+|}}{\big(1-\frac{F_{\Xi^+}(t_i)}{|\Xi^+|}\big)\big(1-\frac{F_{\Xi^+}(t_j)}{|\Xi^+|}\big)} \\
&= \frac{|\Xi^+| - |\Xi^+| - F_{\Xi^+}(t_j) + F_{\Xi^+}(t_j)}{|\Xi^+|}\log\frac{|\Xi^+| - |\Xi^+| - F_{\Xi^+}(t_j) + F_{\Xi^+}(t_j)}{\big(1-1\big)\big(|\Xi^+| - F_{\Xi^+}(t_j)\big)} \\
&= \frac{0-0}{|\Xi^+|}\log\frac{0-0}{0 \times \big(|\Xi^+| - F_{\Xi^+}(t_j)\big)} = 0\log\frac{0}{0} = 0.
\end{aligned}$$

The proof is complete.

10.4.5 Proof of Theorem 7.3.1

In order to investigate the properties of the discrimination measures $\mathbf{ifd}_M^E(t_i^{\delta_i}, t_j^{\delta_j})$ posed in Section 7.3, let us prove Theorem 7.3.1. Again, the signs of the inequalities we are about to prove in the theorem should be carefully noticed.

Theorem 7.3.1 For arbitrary $t_i, t_j \in V^E$, suppose that $I_E(\delta_i, \delta_j)$ is estimated by using the unified expressions given in Eq.(7.15) and Eq.(7.16), we have

(1) if $\gamma_E(t_i,t_j) = \psi_E(t_i)\psi_E(t_j)$, then $\mathbf{ifd}_M^E(t_i,t_j) = 0$, $\mathbf{ifd}_M^E(\bar{t}_i,\bar{t}_j) = 0$, $\mathbf{ifd}_M^E(t_i,\bar{t}_j) = 0$ and $\mathbf{ifd}_M^E(\bar{t}_i,t_j) = 0$;

(2) if $\gamma_E(t_i,t_j) > \psi_E(t_i)\psi_E(t_j)$, then $\mathbf{ifd}_M^E(t_i,t_j) > 0$, $\mathbf{ifd}_M^E(\bar{t}_i,\bar{t}_j) > 0$, $\mathbf{ifd}_M^E(t_i,\bar{t}_j) \leq 0$ and $\mathbf{ifd}_M^E(\bar{t}_i,t_j) \leq 0$;

(3) if $\gamma_E(t_i,t_j) < \psi_E(t_i)\psi_E(t_j)$, then $\mathbf{ifd}_M^E(t_i,t_j) \leq 0$, $\mathbf{ifd}_M^E(\bar{t}_i,\bar{t}_j) \leq 0$, $\mathbf{ifd}_M^E(t_i,\bar{t}_j) \geq 0$ and $\mathbf{ifd}_M^E(\bar{t}_i,t_j) \geq 0$.

Proof. The proof of (1) is obvious.

(2) From $\gamma_E(t_i,t_j) > \psi_E(t_i)\psi_E(t_j)$, we obtain the following inequalities:

$$\begin{aligned}
&\gamma_E(t_i,t_j) > \psi_E(t_i)\psi_E(t_j), \\
&\psi_E(t_i) - \gamma_E(t_i,t_j) < \psi_E(t_i) - \psi_E(t_i)\psi_E(t_j) = \psi_E(t_i)\big(1-\psi_E(t_j)\big), \\
&\psi_E(t_j) - \gamma_E(t_i,t_j) < \psi_E(t_j) - \psi_E(t_i)\psi_E(t_j) = \psi_E(t_j)\big(1-\psi_E(t_i)\big), \\
&1 - \psi_E(t_i) - \psi_E(t_j) + \gamma_E(t_i,t_j) > 1 - \psi_E(t_i) - \psi_E(t_j) + \psi_E(t_i)\psi_E(t_j), \\
&\qquad\qquad = \big(1-\psi_E(t_i)\big)\big(1-\psi_E(t_j)\big),
\end{aligned}$$

which correspond respectively to

$$\begin{aligned}
&\frac{\gamma_E(t_i,t_j)}{\psi_E(t_i)\psi_E(t_j)} > 1, && \frac{\psi_E(t_i) - \gamma_E(t_i,t_j)}{\psi_E(t_i)\big(1-\psi_E(t_j)\big)} < 1, \\
&\frac{\psi_E(t_i) - \gamma_E(t_i,t_j)}{\psi_E(t_i)\big(1-\psi_E(t_j)\big)} < 1, && \frac{1 - \psi_E(t_i) - \psi_E(t_j) + \gamma_E(t_i,t_j)}{\big(1-\psi_E(t_i)\big)\big(1-\psi_E(t_j)\big)} > 1.
\end{aligned}$$

On the other hand, notice that $P_E(\delta_i, \delta_j)$ is a probability distribution, and that $\psi_E(t) > 0$ and $1 - \psi_E(t) > 0$ for $t \in V^E$. Thus, from Eq.(7.17), we have

$$\begin{aligned}
&\gamma_E(t_i, t_j) > \psi_E(t_i)\psi_E(t_j) > 0,\\
&\psi_E(t_i) - \gamma_E(t_i, t_j) \geq 0,\\
&\psi_E(t_j) - \gamma_E(t_i, t_j) \geq 0,\\
&1 - \psi_E(t_i) - \psi_E(t_j) + \gamma_E(t_i, t_j) > \big(1 - \psi_E(t_i)\big)\big(1 - \psi_E(t_j)\big) > 0.
\end{aligned}$$

Hence, by Eq.(7.18), we can see that the four inequalities in (2) hold.

(3) From $\gamma(t_i, t_j) < \psi_E(t_i)\psi_E(t_j)$, we obtain the following inequalities:

$$\begin{aligned}
&\gamma_E(t_i, t_j) < \psi_E(t_i)\psi_E(t_j),\\
&\psi_E(t_i) - \gamma_E(t_i, t_j) > \psi_E(t_i) - \psi_E(t_i)\psi_E(t_j),\\
&\psi_E(t_j) - \gamma_E(t_i, t_j) > \psi_E(t_j) - \psi_E(t_i)\psi_E(t_j),\\
&1 - \psi_E(t_i) - \psi_E(t_j) + \gamma_E(t_i, t_j) < 1 - \psi_E(t_i) - \psi_E(t_j) + \psi_E(t_i)\psi_E(t_j).
\end{aligned}$$

Hence, from Eq.(7.17) and Eq.(7.18), we can see that the four inequalities in (3) hold. The proof is complete.

10.4.6 Proof of Theorem 7.5.1

For the concept of the term-based association introduced in Section 7.5, we have the following theorem.

Theorem 7.5.1 Given term $t_i \in V^{\Xi^+}$, we have

$$ats_M(t_i^{\delta_i}, \Xi^+) = \frac{1}{|\Xi^+|} \sum_{t_j \in V^d - \{t_i\}; d \in \Xi^+} \mathbf{ifd}_M^d(t_i^{\delta_i}, t_j^{\delta_i}).$$

Proof. Notice that, by Definition 7.3.2, $\mathbf{ifd}_M^d(t_i^{\delta_i}, t_j^{\delta_i}) = 0$ if $t_j \notin V^d$. Thus, from Definitions 7.4.1 and 7.4.2, we have immediately

$$\begin{aligned}
ats_M(t_i^{\delta_i}, \Xi^+) &= \sum_{t_j \in V^{\Xi^+} - \{t_i\}} \Big[\frac{1}{|\Xi^+|} \sum_{d \in \Xi^+} \mathbf{ifd}_M^d(t_i^{\delta_i}, t_j^{\delta_i}) \Big]\\
&= \frac{1}{|\Xi^+|} \sum_{t_j \in V^{\Xi^+} - \{t_i\}; d \in \Xi^+} \mathbf{ifd}_M^d(t_i^{\delta_i}, t_j^{\delta_i})\\
&= \frac{1}{|\Xi^+|} \sum_{t_j \in V^{\Xi^+} - \{t_i\}; t_j \in V^d; d \in \Xi^+} \mathbf{ifd}_M^d(t_i^{\delta_i}, t_j^{\delta_i})\\
&= \frac{1}{|\Xi^+|} \sum_{t_j \in V^d - \{t_i\}; d \in \Xi^+} \mathbf{ifd}_M^d(t_i^{\delta_i}, t_j^{\delta_i}).
\end{aligned}$$

The proof is complete.

10.4.7 Proof of Theorem 7.6.1

In order to understand the relationship between the two score functions discussed in Section 7.6, we need to show that order $score_{M_2}(t_1) \leq score_{M_2}(t_2)$ may not guarantee the same order $score_{M_1}(t_1) \leq score_{M_1}(t_2)$, that is, we need to prove the following theorem.

Theorem 7.6.1 Given two terms $t_1, t_2 \in V^{\Xi+} - V^q$, if $score_{M_2}(t_1) \leq score_{M_2}(t_2)$, then there exists a function $\nu(t_1, t_2)$, such that

$$score_{M_1}(t_1) \leq score_{M_1}(t_2) + \nu(t_1, t_2),$$

where $\nu(t_1, t_2)$ may not be always equal to zero.

Proof. To show $\nu(t_1, t_2) = 0$ is not always true, consider the definitions of the score functions. It follows that $score_{M_2}(t_1) \leq score_{M_2}(t_2)$ if and only if

$$\begin{aligned}
&\sum_{(t_1,t_j)_d \in \mho_{t_1}^{\Xi+}} f_q(t_j)\big[\mathbf{ifd}_M^d(t_1, t_j) + \mathbf{ifd}_M^d(\bar{t}_1, \bar{t}_j)\big] \\
&\leq \sum_{(t_2,t_j)_d \in \mho_{t_2}^{\Xi+}} f_q(t_j)\big[\mathbf{ifd}_M^d(t_2, t_j) + \mathbf{ifd}_M^d(\bar{t}_2, \bar{t}_j)\big],
\end{aligned}$$

i.e.,

$$\begin{aligned}
&score_{M_1}(t_1) + \sum_{(t_1,t_j)_d \in \mho_{t_1}^{\Xi+}} f_q(t_j)\mathbf{ifd}_M^d(\bar{t}_1, \bar{t}_j) \\
&\leq score_{M_1}(t_2) + \sum_{(t_2,t_j)_d \in \mho_{t_2}^{\Xi+}} f_q(t_j)\mathbf{ifd}_M^d(\bar{t}_2, \bar{t}_j),
\end{aligned}$$

i.e.,

$$score_{M_1}(t_1) \leq score_{M_1}(t_2) + \nu(t_1, t_2),$$

in which,

$$\begin{aligned}
\nu(t_1, t_2) &= \sum_{(t_2,t_j)_d \in \mho_{t_2}^{\Xi+}} f_q(t_j)\mathbf{ifd}_M^d(\bar{t}_2, \bar{t}_j) - \sum_{(t_1,t_j)_d \in \mho_{t_1}^{\Xi+}} f_q(t_j)\mathbf{ifd}_M^d(\bar{t}_1, \bar{t}_j) \\
&= \sum_{(t_2,t_j)_d \in \mho_{t_2}^{\Xi+}} f_q(t_j)\big[\big(1 - p_d(t_2) - p_d(t_j) + \gamma_d(t_2, t_j)\big) \times \\
&\qquad \times \log \frac{1 - p_d(t_2) - p_d(t_j) + \gamma_d(t_2, t_j)}{\big(1 - p_d(t_2)\big)\big(1 - p_d(t_j)\big)}\big] \\
&\quad - \sum_{(t_1,t_j)_d \in \mho_{t_1}^{\Xi+}} f_q(t_j)\big[\big(1 - p_d(t_1) - p_d(t_j) + \gamma_d(t_1, t_j)\big) \times \\
&\qquad \times \log \frac{1 - p_d(t_1) - p_d(t_j) + \gamma_d(t_1, t_j)}{\big(1 - p_d(t_1)\big)\big(1 - p_d(t_j)\big)}\big].
\end{aligned}$$

It is clear that $\nu(t_1, t_2)$ is a function of terms t_1 and t_2, and would not always be equal to zero. In fact, $\nu(t_1, t_2) = 0$ for all $t_1, t_2 \in V^{\Xi^+} - V^q$ if all relations

$$\mho_{t_2}^{\Xi^+} = \mho_{t_1}^{\Xi^+}, \qquad p_d(t_2) = p_d(t_1), \qquad \gamma_d(t_2, t_j) = \gamma_d(t_1, t_j),$$

hold for all terms $t_1, t_2 \in V^{\Xi^+} - V^q$, which is unlikely. The proof is complete.

10.5 A General Situation of Domain

In Section 7.2, we concentrated on the study of the estimates of the term state distributions, which were determined by two functions $\gamma_E(t_i, t_j)$ and $\psi_E(t)$. Recall that there the variables of $\gamma_E(t_i, t_j)$ were restricted to lie in a range of $V^E \times V^E \subseteq V \times V$ and the variable of $\psi_E(t)$ was restricted to lie in a range of $V^E \subseteq V$, where E was an given entity. Let us now consider the more general situation where these two functions are defined on $V \times V$ and V, respectively. It can be seen that the restrictions are immaterial, and that the extensions are very easy, but they can make some things simpler to explain.

10.5.1 Extension of Domain

Given an information entity E, based on the statistical data within E, introduce a non-negative function ψ_E: $V \to [0, 1)$ and

$$\psi_E(t) \begin{cases} > 0 & \text{when } t \in V^E \\ = 0 & \text{when } t \in V - V^E, \end{cases}$$

which may be or may not be a term probability distribution.

Based on function $\psi_E(t)$, for each term $t \in V$, define

$$P_E(\delta = 1) = \psi_E(t) \quad \text{and} \quad P_E(\delta = 0) = 1 - \psi_E(t),$$

which is a probability distribution over $\Omega = \{1, 0\}$. Thus, $0 < P_E(\delta) < 1$ for $\delta = 1, 0$ when $t \in V^E$; $P_E(\delta = 1) = 0$ and $P_E(\delta = 0) = 1$ when $t \in V - V^E$.

Also, introduce a non-negative function γ_E: $V \times V \to [0, 1]$ and

$$\gamma_E(t_i, t_j) \begin{cases} \geq 0 & \text{when } (t_i, t_j) \in V^E \times V^E \\ = 0 & \text{when } (t_i, t_j) \in (V \times V) - (V^E \times V^E), \end{cases}$$

satisfying $\gamma_E(t_i, t_j) \leq \psi_E(t_i)$, $\gamma_E(t_i, t_j) \leq \psi_E(t_j)$, and $\gamma_E(t_i, t_j) \geq \psi_E(t_i) + \psi_E(t_j) - 1$.

Based on function $\gamma_E(t_i, t_j)$, for an arbitrary term pair $(t_i, t_j) \in V \times V$ $(i \neq j)$, define

$$\begin{aligned} P_E(\delta_i = 1, \delta_j = 1) &= \gamma_E(t_i, t_j), \\ P_E(\delta_i = 1, \delta_j = 0) &= \psi_E(t_i) - \gamma_E(t_i, t_j), \\ P_E(\delta_i = 0, \delta_j = 1) &= \psi_E(t_j) - \gamma_E(t_i, t_j), \\ P_E(\delta_i = 0, \delta_j = 0) &= 1 - \psi_E(t_i) - \psi_E(t_j) + \gamma_E(t_i, t_j), \end{aligned}$$

which is a probability distribution over $\Omega \times \Omega$.

Obviously, when $(t_i, t_j) \in V^E \times V^E$ and $t \in V^E$, functions $\gamma_E(t_i, t_j)$ and $\psi_E(t)$ are identical with the ones introduced in the unified expressions in Section 7.2.4. Thus, the state distributions $P_E(\delta_i, \delta_j)$ and $P_E(\delta)$ defined here are completely the same as the ones defined in Eq.(7.16) and Eq.(7.15), respectively. The difference between the distributions given here and there are for those terms which take values outside V^E.

For instance, for the case where $t_i \in V^E$ but $t_j \notin V^E$ (then $(t_i, t_j) \notin V^E \times V^E$), we have $\gamma_E(t_i, t_j) = 0$ and $\psi_E(t_j) = 0$, and thus

$$\begin{aligned} P_E(\delta_i = 1, \delta_j = 1) &= 0, \\ P_E(\delta_i = 1, \delta_j = 0) &= \psi_E(t_i), \\ P_E(\delta_i = 0, \delta_j = 1) &= 0, \\ P_E(\delta_i = 0, \delta_j = 0) &= 1 - \psi_E(t_i). \end{aligned}$$

Whereas for such a term pair, the joint distribution $P_E(\delta_i, \delta_j)$ given in Eq.(7.16) was not defined. It does not make sense to ask what happens to $P_E(\delta_i, \delta_j)$ when $(t_i, t_j) \notin V^E \times V^E$ for the estimation given Eq.(7.16).

Because $\psi_E(t) = 0$ when $t \in V - V^E$, we need thus analyse the absolute continuity of the joint state distribution with respect to the product of the marginal distributions. The following theorem is for this purpose.

Theorem 7.2.6 For arbitrary terms $t_i, t_j \in V$, $P_E(\delta_i, \delta_j) \ll P_E(\delta_i) \cdot P_E(\delta_j)$ for $\delta_i, \delta_j = 1, 0$.

Proof. According to whether $t_i \in V^E$ and/or $t_j \in V^E$, there are four cases that should be considered:

(c1) $t_i, t_j \in V^E$,

(c2) $t_i \in V^E$ but $t_j \notin V^E$,

(c3) $t_i \notin V^E$ but $t_j \in V^E$,

(c4) $t_i, t_j \notin V^E$.

And for each case, we need verify four distinct state values, respectively.

For (c1): we have $P_E(\delta_i) \cdot P_E(\delta_j) \neq 0$ since $0 < P_E(\delta_i), P_E(\delta_j) < 1$ for $\delta_i, \delta_j = 0, 1$. Thus, $P_E(\delta_i, \delta_j) \ll P_E(\delta_i) \cdot P_E(\delta_j)$, for $\delta_i, \delta_j = 0, 1$. For (c2):

(a) We have $P_E(\delta_i = 1) \cdot P_E(\delta_j = 1) = 0$ since $0 < P_E(\delta_i = 1) < 1$ and $P_E(\delta_j = 1) = 0$. Also, we have $P_E(\delta_i = 1, \delta_j = 1) = \gamma_E(t_i, t_j) = 0$. Thus, $P_E(\delta_i = 1, \delta_j = 1) \ll P_E(\delta_i = 1) \cdot P_E(\delta_j = 1)$ for $(\delta_i, \delta_j) = (1, 1)$.

(b) We have $P_E(\delta_i = 1) \cdot P_E(\delta_j = 0) \neq 0$ since $0 < P_E(\delta_i = 1) < 1$ and $P_E(\delta_j = 0) = 1$. Thus, $P_E(\delta_i = 1, \delta_j = 0) \ll P_E(\delta_i = 1) \cdot P_E(\delta_j = 0)$ for $(\delta_i, \delta_j) = (1, 0)$.

(c) We have $P_E(\delta_i = 0) \cdot P_E(\delta_j = 1) = 0$ since $0 < P_E(\delta_i = 0) < 1$ and $P_E(\delta_j = 1) = 0$. Also, we have $P_E(\delta_i = 0, \delta_j = 1) = \psi_E(t_j) - \gamma_E(t_i, t_j) = 0$. Thus, $P_E(\delta_i = 0, \delta_j = 1) \ll P_E(\delta_i = 0) \cdot P_E(\delta_j = 1)$ for $(\delta_i, \delta_j) = (0, 1)$.

(d) We have $P_E(\delta_i = 0) \cdot P_E(\delta_j = 0) \neq 0$ since $0 < P_E(\delta_i = 0) < 1$ and $P_E(\delta_j = 0) = 1$. Thus, $P_E(\delta_i = 0, \delta_j = 0) \ll P_E(\delta_i = 0) \cdot P_E(\delta_j = 0)$ for $(\delta_i, \delta_j) = (0, 0)$.

For (c3) and (c4), we have a similar discussion to (c2). The proof is complete.

10.5.2 An Alternative Way to View $P_d(\delta_i = 1, \delta_j = 1)$ in Eq.(7.5)

After extending the domains on which functions $\gamma_E(t_i, t_j)$ and $\psi_E(t)$ are defined, the state distributions $P_d(\delta_i, \delta_j)$ and $P_d(\delta)$ can be given for arbitrary terms $t_i, t_j \in V \supseteq V^E$. Thus, it would be now more convenient for us to view what is intuitively meant by $P_d(\delta_i = 1, \delta_j = 1)$. in Method A given in Section 7.2, through an $n \times n$ matrix.

More specifically, suppose that we are given a document d which is represented by $M_d = [w_d(t)]_{1\times n} = [f_d(t)]_{1\times n}$. Thus, from matrix M_d, it follows that

$$
\begin{aligned}
M_d' \times M_d &= \begin{bmatrix} f_d(t_1) \\ f_d(t_2) \\ \cdots \\ f_d(t_n) \end{bmatrix} \times [\, f_d(t_1)\ f_d(t_2)\ \ldots\ f_d(t_n)\,] \\
&= \begin{bmatrix} f_d(t_1)f_d(t_1) & f_d(t_1)f_d(t_2) & \ldots & f_d(t_1)f_d(t_n) \\ f_d(t_2)f_d(t_1) & f_d(t_2)f_d(t_2) & \ldots & f_d(t_2)f_d(t_n) \\ \cdots & \cdots & \cdots & \cdots \\ f_d(t_n)f_d(t_1) & f_d(t_n)f_d(t_2) & \ldots & f_d(t_n)f_d(t_n) \end{bmatrix} \\
&= \Big[f_d(t_i)f_d(t_j)\Big]_{n\times n} = X\Big[\frac{1}{X} f_d(t_i)f_d(t_j)\Big]_{n\times n} \\
&= X \begin{bmatrix} P_d(\delta_1 = 1, \delta_1 = 1) & P_d(\delta_1 = 1, \delta_2 = 1) & \ldots & P_d(\delta_1 = 1, \delta_n = 1) \\ P_d(\delta_2 = 1, \delta_1 = 1) & P_d(\delta_2 = 1, \delta_2 = 1) & \ldots & P_d(\delta_2 = 1, \delta_n = 1) \\ \cdots & \cdots & \cdots & \cdots \\ P_d(\delta_n = 1, \delta_1 = 1) & P_d(\delta_n = 1, \delta_2 = 1) & \ldots & P_d(\delta_n = 1, \delta_n = 1) \end{bmatrix} \\
&= X\Big[P_d(\delta_i = 1, \delta_j = 1)\Big]_{n\times n}.
\end{aligned}
$$

Thus, it shows that probability $P_d(\delta_i = 1, \delta_j = 1)$, for $i, j = 1, ..., n$, can be represented by an $n \times n$ matrix $\big[P_d(\delta_i = 1, \delta_j = 1)\big]_{n\times n}$ with a scale factor ϖ. Its numerator $f_d(t_i)f_d(t_j)$ characterizes the co-occurrence frequencies of t_i and t_j in document d. Whereas its denominator ϖ, the sum of all possible numerators $f_d(t_i)f_d(t_j)$ for $i < j; i, j = 1, 2, ..., n$, is a normalization factor for the probability.

Matrix $\big[f_d(t_i)f_d(t_j)\big]_{n\times n}$, which is symmetric, is called the *co-occurrence frequency matrix* of terms concerning d. Thus, from relation

$$\Big[P_d(\delta_i = 1, \delta_j = 1)\Big]_{n\times n} = \frac{1}{X}\Big[f_d(t_i)f_d(t_j)\Big]_{n\times n},$$

we can see that matrix $\big[P_d(\delta_i = 1, \delta_j = 1)\big]_{n\times n}$ is in fact the *normalized* co-occurrence frequency matrix of terms concerning d.

Notice that assumption $|V^d| \geq 2$ ensures that at least one term pair can be drawn from document d, and thus there exists at least one non-zero element in matrix $\big[P_d(\delta_i = 1, \delta_j = 1)\big]_{n\times n}$, i.e., $\big[P_d(\delta_i = 1, \delta_j = 1)\big]_{n\times n} \neq [0]_{n\times n}$. Notice also that, subject to the condition that no two components of (t_i, t_j) can be the same, the cases where $i = j$, corresponding to elements $P_d(\delta_i = 1, \delta_i = 1)$ for $i = 1, ..., n$, are not considered in our context. However, it is only for mathematical convenience in notation to include these cases.

10.5.3 Appropriateness of Definition 7.3.2

Given an entity E, for the case where $t_i, t_j \in V^E$, we have the general expressions of the discrimination measures given Eq.(7.18) as shown in Section 7.3. Now further, for the case

where $t_i \in V^E$ but $t_j \notin V^E$, we have $\gamma_E(t_i, t_j) = 0$ and $\psi_E(t_j) = 0$. Thus, from Eq.(7.18) we obtain immediately

$$\mathbf{ifd}_M^E(t_i, t_j) = 0 \log \frac{0}{\psi_E(t_i)0} = 0,$$

$$\mathbf{ifd}_M^E(t_i, \bar{t}_j) = \big(\psi_E(t_i) - 0\big) \log \frac{\psi_E(t_i) - 0}{\psi_E(t_i)\big(1 - 0\big)} = 0,$$

$$\mathbf{ifd}_M^E(\bar{t}_i, t_j) = \big(0 - 0\big) \log \frac{0 - 0}{\big(1 - \psi_E(t_i)\big)0} = 0,$$

$$\mathbf{ifd}_M^E(\bar{t}_i, \bar{t}_j) = \big(1 - \psi_E(t_i) - 0 + 0\big) \log \frac{1 - \psi_E(t_i) - 0 + 0}{\big(1 - \psi_E(t_i)\big)\big(1 - 0\big)} = 0.$$

That is, the contributions made by the individual state values to the expected mutual information are zero, and hence we have $I_E(\delta_i, \delta_j) = 0$. Similar results can be obtained for the case where $t_i \notin V^E$ but $t_j \in V^E$, and the case where $t_i, t_j \notin V^E$. Such mathematical results are consistent with our intuitive understanding: when one term appears in document d but another does not, these two terms should not be regarded as containing statistically mutual information concerning the entity, similarly for two terms neither of which appear in document E. Therefore, we believe that two terms contain mutual information, whether more or less, only when they at least co-occur in some entity.

10.6 Examples

Let us now give some examples to illustrate the computation involved in the method proposed in Chapter 7. Suppose that $\Xi^+ = \{d_1, d_2, d_3\} \subseteq D$ is the relevant sample set with respect to query q, and that their corresponding statistical data is shown below.

t_i	t_1	t_2	t_3	t_4	t_5	t_6	t_7	t_8	...	t_n
$f_q(t_i)$		1			2					
$f_{d_1}(t_i)$	1	2		1	1	1		2		
$f_{d_2}(t_i)$	2		1	1	3	1	2			
$f_{d_3}(t_i)$		1		2		1				
$P_q(\delta_i = 1)$		$\frac{1}{3}$			$\frac{2}{3}$					
$P_{d_1}(\delta_i = 1)$	$\frac{1}{8}$	$\frac{2}{8}$		$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{8}$		$\frac{2}{8}$		
$P_{d_2}(\delta_i = 1)$	$\frac{2}{10}$		$\frac{1}{10}$	$\frac{1}{10}$	$\frac{3}{10}$	$\frac{1}{10}$	$\frac{2}{10}$			
$P_{d_2}(\delta_i = 1)$		$\frac{1}{4}$		$\frac{2}{4}$		$\frac{1}{4}$				

In the following three examples, we will calculate the scores of term t_1 using the score functions given in Sections 7.6 and 7.7 for the different estimations Methods. Example A corresponds to Method A, Example B corresponds to Method B and Example C corresponds to Method C.

10.6.1 Example A

For term $t_1 \in V^{\Xi^+}$, the association set of term t_1 with query q concerning set Ξ^+ can be written as

$$\mho_{t_1}^{\Xi^+} = \{(t_1, t_2)_{d_1}, (t_1, t_5)_{d_1}, (t_1, t_5)_{d_2}\}.$$

From Eq.(7.6) and $V^{d_1} = \{t_1, t_2, t_4, t_5, t_6, t_8\}$, it is easy to calculate ϖ by

$$\sum_{i'<j'; t_{i'}, t_{j'} \in V^{d_1}} f_{d_1}(t_{i'}) f_{d_1}(t_{j'}) = 1 \times (2+1+1+1+2) + 2 \times (1+1+1+2) + 1 \times (1+1+2) + 1 \times (1+2) + 1 \times (2) = 26.$$

Similarly, $V^{d_2} = \{t_1, t_3, t_4, t_5, t_6, t_7\}$, thus

$$\sum_{i'<j'; t_{i'}, t_{j'} \in V^{d_2}} f_{d_2}(t_{i'}) f_{d_2}(t_{j'}) = 2 \times (1+1+3+1+2) + 1 \times (1+3+1+2) + 1 \times (3+1+2) + 3 \times (1+2) + 1 \times (2) = 40.$$

Notice that we need not calculate ϖ for d_3 since $t_1 \notin V^{d_3}$, and thus t_1 will not co-occur with any terms in d_3.

Then, from Eq.(7.5), we can estimate the joint state distribution for the corresponding element in the following table (the joint state distributions $P_d(\delta_1, \delta_j)$ of the corresponding element $(t_1, t_j)_d \in \mho_{t_1}^{\Xi+}$ is listed in the column below the element).

$(t_1, t_j)_d$	$(t_1, t_2)_{d_1}$	$(t_1, t_5)_{d_1}$	$(t_1, t_5)_{d_2}$
$P_d(\delta_1 = 1, \delta_j = 1)$	$\frac{1\times2}{26} = \frac{8}{104}$	$\frac{1\times1}{26} = \frac{4}{104}$	$\frac{2\times3}{40} = \frac{3}{20}$
$P_d(\delta_1 = 1, \delta_j = 0)$	$\frac{1}{8} - \frac{2}{26} = \frac{5}{104}$	$\frac{1}{8} - \frac{1}{26} = \frac{9}{104}$	$\frac{2}{10} - \frac{3}{20} = \frac{1}{20}$
$P_d(\delta_1 = 0, \delta_j = 1)$	$\frac{2}{8} - \frac{2}{26} = \frac{18}{104}$	$\frac{1}{8} - \frac{1}{26} = \frac{9}{104}$	$\frac{3}{10} - \frac{3}{20} = \frac{3}{20}$
$P_d(\delta_1 = 0, \delta_j = 0)$	$1 - \frac{1}{8} - \frac{2}{8} + \frac{2}{26} = \frac{73}{104}$	$1 - \frac{1}{8} - \frac{1}{8} + \frac{1}{26} = \frac{82}{104}$	$1 - \frac{2}{10} - \frac{3}{10} + \frac{3}{20} = \frac{13}{20}$

Then, for terms t_1 and t_2, for instance, it follows immediately that

$$\begin{aligned} I_{d_1}(\delta_1, \delta_2) &= \frac{8}{104} \log \frac{\frac{8}{104}}{\frac{1}{8}\frac{2}{8}} + \frac{5}{104} \log \frac{\frac{5}{104}}{\frac{1}{8}(1-\frac{2}{8})} + \frac{18}{104} \log \frac{\frac{18}{104}}{(1-\frac{1}{8})\frac{2}{8}} + \frac{73}{104} \log \frac{\frac{73}{104}}{(1-\frac{1}{8})(1-\frac{2}{8})} \\ &= \frac{1}{13} \log \frac{32}{13} + \frac{5}{104} \log \frac{20}{39} + \frac{9}{52} \log \frac{72}{91} + \frac{73}{104} \log \frac{292}{273} \\ &\approx 0.0769 \log 2.4615 + 0.0481 \log 0.5128 + 0.1731 \log 0.7912 + 0.7019 \log 1.0696 \\ &\approx 0.0769 \times 0.9008 + 0.0481 \times (-0.6679) + 0.1731 \times (-0.2342) + 0.7019 \times 0.0673 \\ &\approx 0.0693 - 0.0321 - 0.0405 + 0.0472 = 0.0439. \end{aligned}$$

As mentioned before, $I_{d_1}(\delta_1, \delta_2)$ is expected over the state value space, and hence each of its items offset one another. Thus, we have, for instance, $I_{d_1}(\delta_1, \delta_2) \approx 0.0439 < 0.1165 \approx \mathbf{ifd}_M^{d_1}(t_1, t_2) + \mathbf{ifd}_M^{d}(\bar{t}_i, \bar{t}_j)$.

For term $t_1 \in V^{\Xi+}$, we have

$$\begin{aligned} score_{M_1}(t_1) &= \sum_{(t_1, t_j)_d \in \mho_{t_1}^{\Xi+}} f_q(t_j) \mathbf{ifd}_M^{d}(t_1, t_j) \\ &= [f_q(t_2) \mathbf{ifd}_M^{d_1}(t_1, t_2) + f_q(t_5) \mathbf{ifd}_M^{d_1}(t_1, t_5)] + [f_q(t_5) \mathbf{ifd}_M^{d_2}(t_1, t_5)] \\ &= 1 \times \frac{2}{26} \log \frac{\frac{2}{26}}{\frac{1}{8}\frac{2}{8}} + 2 \times \frac{1}{26} \log \frac{\frac{1}{26}}{\frac{1}{8}\frac{1}{8}} + 2 \times \frac{3}{20} \log \frac{\frac{3}{20}}{\frac{2}{10}\frac{3}{10}} \\ &= \frac{1}{13} \log \frac{32}{13} + \frac{1}{13} \log \frac{32}{13} + \frac{3}{10} \log \frac{5}{2} \end{aligned}$$

$$\approx 0.0769 \log 2.4615 + 0.0769 \log 2.4615 + 0.3000 \log 2.5000$$
$$\approx 0.0769 \times 0.9008 + 0.0769 \times 0.9008 + 0.3000 \times 0.9163$$
$$\approx 0.0693 + 0.0693 + 0.2749 = 0.4135 > 0.0000.$$

$$score_{M_2}(t_1) = \sum_{(t_1,t_j)_d \in \mho_{t_1}^{\Xi^+}} f_q(t_j)\left[\mathbf{ifd}_M^d(t_1,t_j) + \mathbf{ifd}_M^d(\bar{t}_1,\bar{t}_j)\right]$$
$$\approx 0.4135 + \left[f_q(t_2)\mathbf{ifd}_M^{d_1}(\bar{t}_1,\bar{t}_2) + f_q(t_5)\mathbf{ifd}_M^{d_1}(\bar{t}_1,\bar{t}_5)\right] + \left[f_q(t_5)\mathbf{ifd}_M^{d_2}(\bar{t}_1,\bar{t}_5)\right]$$
$$= 0.4135 + 1 \times \frac{73}{104} \log \frac{\frac{73}{104}}{(1-\frac{1}{8})(1-\frac{2}{8})} + 2 \times \frac{82}{104} \log \frac{\frac{82}{104}}{(1-\frac{1}{8})(1-\frac{1}{8})}$$
$$+ 2 \times \frac{13}{20} \log \frac{\frac{13}{20}}{(1-\frac{2}{10})(1-\frac{3}{10})}$$
$$= 0.4135 + \frac{73}{104} \log \frac{292}{273} + \frac{41}{26} \log \frac{656}{637} + \frac{13}{10} \log \frac{65}{56}$$
$$\approx 0.4135 + 0.7019 \log 1.0696 + 1.5769 \log 1.0298 + 1.3000 \log 1.1607$$
$$\approx 0.4135 + 0.7019 \times 0.0673 + 1.5769 \times 0.0294 + 1.3000 \times 0.1490$$
$$\approx 0.4135 + 0.0472 + 0.0464 + 0.1937 = 0.7008 > 0.0000.$$

Comparing two score functions, it is easily seen that the statistical information of the none-occurrence of terms is equally as important as that of the co-occurrence of terms. ♠

10.6.2 Example B

Similar to Example A, we have the same association set of term t_1 with query q concerning set Ξ^+. Also, from Eq.(7.9), we can estimate the joint state distribution for the corresponding element in the table as follows.

$(t_1,t_j)_d$	$(t_1,t_2)_{d_1}$	$(t_1,t_5)_{d_1}$	$(t_1,t_5)_{d_2}$
$P_d(\delta_1=1,\delta_j=1)$	$\frac{1}{8} \times \frac{2}{8-1} = \frac{2}{56}$	$\frac{1}{8} \times \frac{1}{8-1} = \frac{1}{56}$	$\frac{2}{10} \times \frac{3}{10-1} = \frac{6}{90}$
$P_d(\delta_1=1,\delta_j=0)$	$\frac{1}{8} \times (1-\frac{2}{8-1}) = \frac{5}{56}$	$\frac{1}{8} \times (1-\frac{1}{8-1}) = \frac{6}{56}$	$\frac{2}{10} \times (1-\frac{3}{10-1}) = \frac{12}{90}$
$P_d(\delta_1=0,\delta_j=1)$	$\frac{2}{8} \times (1-\frac{1}{8-1}) = \frac{12}{56}$	$\frac{1}{8} \times (1-\frac{1}{8-1}) = \frac{6}{56}$	$\frac{3}{10} \times (1-\frac{2}{10-1}) = \frac{21}{90}$
$P_d(\delta_1=0,\delta_j=0)$	$1-\frac{2}{56}-\frac{5}{56}-\frac{12}{56} = \frac{37}{56}$	$1-\frac{1}{56}-\frac{6}{56}-\frac{6}{56} = \frac{43}{56}$	$1-\frac{6}{90}-\frac{12}{90}-\frac{21}{90} = \frac{51}{90}$

Then, for terms t_1 and t_2, for instance, it follows immediately that

$$I_{d_1}(\delta_1,\delta_2) = \frac{2}{56} \log \frac{\frac{2}{56}}{\frac{1}{8}\frac{2}{8}} + \frac{12}{56} \log \frac{\frac{12}{56}}{\frac{2}{8}(1-\frac{1}{8})} + \frac{5}{56} \log \frac{\frac{5}{56}}{(1-\frac{2}{8})\frac{1}{8}} + \frac{37}{56} \log \frac{\frac{37}{56}}{(1-\frac{1}{8})(1-\frac{2}{8})}$$
$$= \frac{1}{28} \log \frac{8}{7} + \frac{3}{14} \log \frac{48}{49} + \frac{5}{56} \log \frac{20}{21} + \frac{37}{56} \log \frac{148}{147}$$
$$\approx 0.0357 \log 1.1429 + 0.2143 \log 0.9796 + 0.0893 \log 0.9524 + 0.6607 \log 1.0068$$
$$\approx 0.0357 \times 0.1336 + 0.2143 \times (-0.0206) + 0.0893 \times (-0.0488) + 0.6607 \times 0.0068$$
$$\approx 0.0048 - 0.0044 - 0.0044 + 0.0045 = 0.0005 > 0.0000.$$

It also has $I_{d_1}(\delta_1,\delta_2) \approx 0.0005 < 0.0093 \approx \mathbf{ifd}_M^{d_1}(t_1,t_2) + \mathbf{ifd}_M^{d_1}(\bar{t}_i,\bar{t}_j)$ in this example.

For term $t_1 \in V^{\Xi+}$, we have

$$
\begin{aligned}
score_{M_1}(t_1) &= \sum_{(t_1,t_j)_d \in \mho_{t_1}^{\Xi+}} f_q(t_j)\mathbf{ifd}_M^d(t_1,t_j) \\
&= \left[f_q(t_2)\mathbf{ifd}_M^{d_1}(t_1,t_2) + f_q(t_5)\mathbf{ifd}_M^{d_1}(t_1,t_5)\right] + \left[f_q(t_5)\mathbf{ifd}_M^{d_2}(t_1,t_5)\right] \\
&= 1 \times \frac{2}{56}\log\frac{\frac{2}{56}}{\frac{1}{8}\frac{2}{8}} + 2 \times \frac{1}{56}\log\frac{\frac{1}{56}}{\frac{1}{8}\frac{1}{8}} + 2 \times \frac{6}{90}\log\frac{\frac{6}{90}}{\frac{2}{10}\frac{3}{10}} \\
&= \frac{1}{28}\log\frac{8}{7} + \frac{1}{28}\log\frac{8}{7} + \frac{2}{15}\log\frac{10}{9} \\
&\approx 0.0357\log 1.1429 + 0.0357\log 1.1429 + 0.1333\log 1.1111 \\
&\approx 0.0357 \times 0.1336 + 0.0357 \times 0.1336 + 0.1333 \times 0.1054 \\
&\approx 0.0048 + 0.0048 + 0.0140 = 0.0236 > 0.0000. \\
score_{M_2}(t_1) &= \sum_{(t_1,t_j)_d \in \mho_{t_1}^{\Xi+}} f_q(t_j)\left[\mathbf{ifd}_M^d(t_1,t_j) + \mathbf{ifd}_M^d(\bar{t}_1,\bar{t}_j)\right] \\
&\approx 0.0236 + \left[f_q(t_2)\mathbf{ifd}_M^{d_1}(\bar{t}_1,\bar{t}_2) + f_q(t_5)\mathbf{ifd}_M^{d_1}(\bar{t}_1,\bar{t}_5)\right] + \left[f_q(t_5)\mathbf{ifd}_M^{d_2}(\bar{t}_1,\bar{t}_5)\right] \\
&= 0.0236 + 1 \times \frac{37}{56}\log\frac{\frac{37}{56}}{(1-\frac{1}{8})(1-\frac{2}{8})} + 2 \times \frac{43}{56}\log\frac{\frac{43}{56}}{(1-\frac{1}{8})(1-\frac{1}{8})} \\
&\qquad + 2 \times \frac{51}{90}\log\frac{\frac{51}{90}}{(1-\frac{2}{10})(1-\frac{3}{10})} \\
&= 0.0236 + \frac{37}{56}\log\frac{148}{147} + \frac{43}{28}\log\frac{344}{343} + \frac{17}{15}\log\frac{85}{84} \\
&\approx 0.0236 + 0.6607\log 1.0068 + 1.5357\log 1.0029 + 1.1333\log 1.0119 \\
&\approx 0.0236 + 0.6607 \times 0.0068 + 1.5357 \times 0.0029 + 1.1333 \times 0.0118 \\
&\approx 0.0236 + 0.0045 + 0.0045 + 0.0134 = 0.0460 > 0.0000. \quad \spadesuit
\end{aligned}
$$

10.6.3 Example C

For term t_1 and t_2, we have $F_{\Xi+}(t_1) = 2$, $F_{\Xi+}(t_2) = 2$, $F_{\Xi+}(t_1,t_2) = 1$. Thus,

$$
\begin{aligned}
I_{\Xi+}(\delta_1,\delta_2) &= \frac{1}{3}\log\frac{\frac{1}{3}}{\frac{2}{3}\frac{2}{3}} + \frac{2-1}{3}\log\frac{\frac{2-1}{3}}{\frac{2}{3}(1-\frac{2}{3})} + \frac{2-1}{3}\log\frac{\frac{2-1}{3}}{(1-\frac{2}{3})\frac{2}{3}} \\
&\quad + \frac{3-2-2+1}{3}\log\frac{\frac{3-2-2+1}{3}}{(1-\frac{2}{3})(1-\frac{2}{3})} \\
&= \frac{1}{3}\log\frac{3}{4} + \frac{1}{3}\log\frac{3}{2} + \frac{1}{3}\log\frac{3}{2} + 0\log 0 \\
&\approx -0.0959 + 0.1352 + 0.1352 - 0.0000 = 0.1745. \\
emim_{\Xi+}(\delta_1,\delta_2) &= 1 \times \log\frac{1}{2\times 2} + (2-1)\log\frac{2-1}{2\times(3-2)} + (2-1)\log\frac{2-1}{(3-2)\times 2} \\
&\quad + (3-2-2+1)\log\frac{3-2-2+1}{(3-2)\times(3-2)} \\
&= \log\frac{1}{4} + \log\frac{1}{2} + \log\frac{1}{2} + 0\log\frac{0}{1}
\end{aligned}
$$

$$\approx -1.3863 - 0.6931 - 0.6931 - 0.0000 = -2.7725.$$

For term t_1 and t_5, we have $F_{\Xi+}(t_1) = 2$, $F_{\Xi+}(t_5) = 2$, $F_{\Xi+}(t_1, t_5) = 2$. Thus,

$$\begin{aligned}
I_{\Xi+}(\delta_1, \delta_5) &= \frac{2}{3}\log\frac{\frac{2}{3}}{\frac{2}{3}\frac{2}{3}} + \frac{2-2}{3}\log\frac{\frac{2-2}{3}}{\frac{2}{3}(1-\frac{2}{3})} + \frac{2-2}{3}\log\frac{\frac{2-2}{3}}{(1-\frac{2}{3})\frac{2}{3}} \\
&\quad + \frac{3-2-2+2}{3}\log\frac{\frac{3-2-2+2}{3}}{(1-\frac{2}{3})(1-\frac{2}{3})} \\
&= \frac{2}{3}\log\frac{3}{2} + 0\log 0 + 0\log 0 + \frac{1}{3}\log 3 \\
&\approx 0.2703 - 0.0000 - 0.0000 + 0.3662 = 0.6365. \\
emim_{\Xi+}(\delta_1, \delta_5) &= 2 \times \log\frac{2}{2\times 2} + (2-2)\log\frac{2-2}{2\times(3-2)} + (2-2)\log\frac{2-2}{(3-2)\times 2} \\
&\quad + (3-2-2+2)\log\frac{3-2-2+2}{(3-2)\times(3-2)} \\
&= 2\log\frac{1}{2} + 0\log 0 + 0\log 0 + \log 1 \\
&\approx -1.3863 - 0.0000 - 0.0000 - 0.0000 = -1.3863.
\end{aligned}$$

Next, for calculating the scores of term t_1, we need write out the association set of term t_1 with query q:

$$\mho_{t_1} = \{(t_1, t_2)_{\Xi+}, (t_1, t_5)_{\Xi+}\}.$$

Thus, from the score function given in Section 7.6, we have

$$\begin{aligned}
score_{M_1}(t_1) &= \sum_{(t_i,t_j)_{\Xi+} \in \mho_{t_i}} f_q(t_j)\mathbf{ifd}_M^{\Xi+}(t_i, t_j) \\
&= f_q(t_2)\mathbf{ifd}_M^{\Xi+}(t_1, t_2) + f_q(t_5)\mathbf{ifd}_M^{\Xi+}(t_1, t_5) \\
&= 1 \times \frac{1}{3}\log\frac{\frac{1}{3}}{\frac{2}{3}\frac{2}{3}} + 2 \times \frac{2}{3}\log\frac{\frac{2}{3}}{\frac{2}{3}\frac{2}{3}} \\
&\approx 1 \times (-0.0959) + 2 \times 0.2703 \\
&\approx -0.0959 + 0.5406 = 0.4447 > 0.0000. \\
score_{M_2}(t_1) &= \sum_{(t_i,t_j)_{\Xi+} \in \mho_{t_i}} f_q(t_j)\left[\mathbf{ifd}_M^{\Xi+}(t_i, t_j) + \mathbf{ifd}_M^{\Xi+}(\bar{t}_i, \bar{t}_j)\right] \\
&\approx 0.4447 + f_q(t_2)\mathbf{ifd}_M^{\Xi+}(\bar{t}_1, \bar{t}_2) + f_q(t_5)\mathbf{ifd}_M^{\Xi+}(\bar{t}_1, \bar{t}_5) \\
&= 0.4447 + 1 \times \frac{3-2-2+1}{3}\log\frac{\frac{3-2-2+1}{3}}{(1-\frac{2}{3})(1-\frac{2}{3})} + \\
&\qquad + 2 \times \frac{3-2-2+2}{3}\log\frac{\frac{3-2-2+2}{3}}{(1-\frac{2}{3})(1-\frac{2}{3})} \\
&\approx 0.4447 + 1 \times 0.0000 + 2 \times 0.3662 \\
&\approx 0.4447 + 0.0000 + 0.7324 = 1.1771 > 0.0000. \quad \spadesuit
\end{aligned}$$

Bibliography

[1] J. Aczel and Z. Daroczy. *On Measures of Information and Their Characterizations*. Academic Press, New York, 1975.

[2] J. Allan. Relevance feedback with too much data. In *Proceedings of the 18th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*, pages 337–343, 1995.

[3] J. Allan. Incremental relevance feedback for information filtering. In *Proceedings of the 19th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*, pages 270–278, 1996.

[4] J. Allan, L. Ballesteros, J. P. Callan, W. B. Croft, and Z. Lu. Current experiments with INQUERY. In *The 4th Text REtrieval Conference (TREC-4)*, pages 49–64. NIST Special Publication, 1996.

[5] G. Amati and C. J. Van Rijsbergen. Probabilistic models of information retrieval based on measuring the divergence from randomness. *ACM Transactions on Information Systems*, 20(4):357–389, 2002.

[6] R. Attar and A. S. Fraenkel. Local feedback in full-text retrieval systems. *Journal of the Association for Computing Machinery*, 24(3):397–417, 1977.

[7] R. Attar and A. S. Fraenkel. Experiments in local metrical feedback in full-text retrieval systems. *Information Processing & Management*, 17(3):115–126, 1981.

[8] N. J. Belkin and W. B. Croft. Retrieval techniques. *Annual Review of Information Science and Technology*, 22:109–145, 1987.

[9] A. Berger and J. Lafferty. Information retrieval as statistical translation. In *Proceedings of the 22nd Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*, pages 222–229, 1999.

[10] T. Biru, A. El-Hamdouchi, R. S. Rees, and P. Willett. Inclusion of relevance information in the term discrimination model. *Journal of Documentation*, 45(2):85–109, 1989.

[11] D. Blocks, C. Binding, D. Cunliffe, and D. Tudhope. Qualitative evaluation of thesaurus-based retrieval. In *Proceedings of the 6th European Conference on Research and Advanced Technology for Digital Libraries*, volume 2458 of *Lecture Notes in Computer Science*, pages 346–361, 2002.

[12] A. Bookstein. Relevance. *Journal of the American Society for Information Science*, 30:269–273, 1979.

[13] A. Bookstein and D. R. Swanson. Probabilistic models for automatic indexing. *Journal of the American Society for Information Science*, 25(5):312–318, 1974.

[14] A. Bookstein and D. R. Swanson. A decision theoretic foundation for indexing. *Journal of the American Society for Information Science*, 26(1):45–50, 1975.

[15] C. Buckley, J. Allan, and G. Salton. Automatic routing and ad-hoc retrieval using SMART: TREC-2. In *The 2nd Text REtrieval Conference (TREC-2)*, pages 45–56. NIST Special Publication, 1994.

[16] C. Buckley and G. Salton. Optimisation of relevance feedback weights. In *Proceedings of the 18th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*, pages 351–357, 1995.

[17] C. Buckley, G. Salton, and J. Allan. Automatic retrieval with locality information using SMART. In *The 1st Text REtrieval Conference (TREC-1)*, pages 59–72. NIST Special Publication, 1993.

[18] C. Buckley, G. Salton, and J. Allan. The effect of adding relevance information in a relevance feedback environment. In *Proceedings of the 17th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*, pages 292–300, 1994.

[19] C. Buckley, G. Salton, J. Allan, and A. Singhal. Automatic query expansion using SMART: TREC-3. In *The 3rd Text REtrieval Conference (TREC-3)*, pages 69–79. NIST Special Publication, 1995.

[20] C. Buckley, A. Singhal, M. Mitra, and G. Salton. New retrieval approaches using SMART: TREC-4. In *The 4th Text REtrieval Conference (TREC-4)*, pages 25–48. NIST Special Publication, 1996.

[21] D. Cai and C. J. Van Rijsbergen. A case study for automatic query expansion based on divergence. Technical Report TR-2004-172, Department of Computing Science, University of Glasgow, 2004.

[22] D. Cai and C. J. Van Rijsbergen. Relevance weight revisited. Technical Report TR-2004-173, Department of Computing Science, University of Glasgow, 2004.

[23] D. Cai, C. J. Van Rijsbergen, and J. M. Jose. Automatic query expansion based on divergence. In *The Proceedings of the 10th International Conference on Information and Knowledge Management (ACM-CIKM)*, pages 419–426, 2001.

[24] J. P. Callan. Passage-level evidence in document retrieval. In *Proceedings of the 17th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*, pages 302–310, 1994.

[25] J. P. Callan, W. B. Croft, and J. Broglio. TREC and TIPSTER experiments with INQUERY. *Information Processing & Management*, 31(3):327–343, 1995.

[26] F. Can and E. A. Ozkarahan. Computation of term/document discrimination values by use of the cover coefficient concept. *Journal of the American Society for Information Science*, 38(3):171–183, 1987.

[27] D. Carmel, E. Farchi, Y. Petruschka, and A. Soffer. Automatic query refinement using lexical affinities with maximal information gain. In *Proceedings of the 25th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*, pages 283–290, 2002.

[28] C. Carpineto, R. Mori, and G. Romano. Informative term selection for automatic query expansion. In *The 7th Text REtrieval Conference (TREC-7)*, pages 363–369. NIST Special Publication, 1998.

[29] C. Carpineto, R. D. Mori, G. Romano, and B. Bigi. An information-theoretic approach to automatic query expansion. *ACM Transactions on Information Systems*, 19(1):1–27, 2001.

[30] C. Carpineto and G. Romano. TREC-8 automatic ad-hoc experiments at Fondazione Ugo Bordoni. In *The 8th Text REtrieval Conference*, pages 377–380, 1999.

[31] B. Cho, C. Lee, and G. G. Lee. Exploring term dependences in probabilistic information retrieval model. *Information Processing & Management*, 39:505–519, 2003.

[32] C. K. Chow and C. N. Liu. Approximating discrete probability distributions with dependence trees. *IEEE Transactions on Information Theory*, IT-14(3):462–467, 1968.

[33] Y. M. Chung and J. Y. Lee. A corpus-based approach to comparative evaluation of statistical term association measures. *Journal of the American Society for Information Science*, 52(4):283–296, 2001.

[34] K. W. Church and P. Hanks. Word association norms, mutual information, and lexicography. *Journal of the American Society for Information Science*, 16(1):22–29, 1990.

[35] W. S. Cooper. A definition of relevance for information retrieval. *Information Storage and Retrieval*, 7:19–37, 1971.

[36] W. S. Cooper. The suboptimality of retrieval rankings based on probability of usefulness. Technical report, School of Library and Information Studies, University of California, Berkeley, California, 1976.

[37] W. S. Cooper and P. Huizinga. The maximum entropy principle and its application to the design of probabilistic retrieval system. *Information Technology: Research and Development*, 1:99–112, 1982.

[38] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. Wiley series in telecommunications. Wiley-Interscience, New York, NY, 1991.

[39] R. G. Crawford. The computation of discrimination values. *Information Processing & Management*, 11:249–253, 1975.

[40] W. B. Croft. Experiments with representation in a document retrieval system. *Information Technology: Research and Development*, 2(1):1–21, 1983.

[41] W. B. Croft. Boolean queries and term dependencies in probabilistic models. *Journal of the American Society for Information Science*, 37(2):71–77, 1986.

[42] W. B. Croft and D. J. Harper. Using probabilistic models of document retrieval without relevance information. *Journal of Documentation*, 35(4):285–295, 1979.

[43] W. B. Croft and D. D. Lewis. An approach to natural language processing for document retrieval. In *Proceedings of the 10th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*, pages 26–32, 1987.

[44] S. Cronen-Townsend, Y. Zhou, and W. B. Croft. Predicting query performance. In *Proceedings of the 25th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*, pages 299–306, 2002.

[45] C. J. Crouch. An analysis of approximate versus exact discrimination values. *Information Processing & Management*, 24(1):5–16, 1988.

[46] C. J. Crouch. An approach to the automatic construction of global thesauri. *Information Processing & Management*, 26(5):629–640, 1990.

[47] C. J. Crouch and B. Yang. Experiments in automatic statistical thesaurus construction. In *Proceedings of the 15th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*, pages 77–88, 1992.

[48] I. Dagan, L. Lee, and F. Pereira. Similarity-based models of cooccurrence probabilities. *Machine Learning*, 34(1–3):43–69, 1999. Special issue on natural language learning.

[49] N. Dillon and J. Desper. Automatic relevance feedback in Boolean retrieval systems. *Journal of Documentation*, 36:197–208, 1980.

[50] T. E. Doszkocs. AID, an associative interactive dictionary for online searching. *Online Review*, 2(2):163–172, 1978.

[51] R. O. Duda and P. E. Hart. *Pattern classification and scene analysis.* Wiley, New York, 1973.

[52] E. N. Efthimiadis. *Interactive query expansion and relevance feedback for document retrieval systems.* PhD thesis, The City University, London, 1992.

[53] E. N. Efthimiadis. Query expansion. *Annual Review of Information Systems and Technology*, 31:121–187, 1996.

[54] E. N. Efthimiadis and P. V. Biron. UCLA-Okapi at TREC-2: query expansion experiments. In *The 2nd Text REtrieval Conference (TREC-2)*, pages 279–289. NIST Special Publication, 1994.

[55] A. El-Hamdoichi and P. Willett. An improved algorithm for the calculation of exact term discrimination values. *Information Processing & Management*, 24(1):17–22, 1988.

[56] P. C. Fishburn. *Utility Theory for Decision Making.* Wiley, New York, 1970.

[57] C. Fox. A stop list for general text. *ACM SIGIR Forum*, 24(1-2):19–35, 1990.

[58] C. Fox. Lexical analysis and stoplists. In *Information Retrieval: Data Structures & Algorithms*, pages 102–130. 1992.

[59] W. Francis and H. Kucera. *Frequency Analysis of English Usage.* Houghton Mifflin, New York, 1982.

[60] N. Fuhr. Optimum polynomial retrieval functions based on the probability ranking principle. *ACM Transactions on Information Systems*, 7(3):183–204, 1989.

[61] N. Fuhr. Probabilistic models in information retrieval. *The Computer Journal*, 35(3):243–255, 1992.

[62] N. Fuhr and C. Buckley. A probabilistic learning approach for document indexing. *ACM Transactions on Information Systems*, 9(3):223–248, 1991.

[63] G. W. Furnas, T. K. Landauer, L. M. Gomez, and S. T. Dumais. The vocabulary problem in human-system communication. *Communications of the ACM*, 30(11):964–971, 1987.

[64] S. Gauch and J. Wang. A corpus analysis approach for automatic query expansion. In *Proceedings of the 6th International Conference on Information and Knowledge Management (ACM-CIKM)*, pages 278–284, 1997.

[65] S. Gauch, J. Wang, and S. M. Rachakonda. A corpus analysis approach for automatic query expansion and its extension to multiple databases. *ACM Transactions on Information Systems*, 17(3):250–269, 1999.

[66] S. Goldman. *Information Theory.* Prentice-Hall, New York, 1953.

[67] I. J. Good. *Probability and the Weighing of Evidence.* Charles Griffin, London, 1950.

[68] S. Guiasu. *Information Theory with Applications.* McGraw-Hill, New York, 1977.

[69] M. Hancock-Beaulieu, M. Fieldhouse, and T. Do. An evaluation of interactive query expansion in an online library catalogue with a graphical user interface. *Journal of Documentation*, 51(3):225–243, 1995.

[70] D. Harman. An experimental study of factors important in document ranking. In *Proceedings of the 9th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*, pages 186–193, 1986.

[71] D. Harman. A failure analysis on the limitations of suffixing in an on-line environment. In *Proceedings of the 10th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*, pages 102–108, 1987.

[72] D. Harman. Towards interactive query expansion. In *Proceedings of the 11th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*, pages 321–331, 1988.

[73] D. Harman. How effective is suffixing? *Journal of the American Society for Information Science*, 42(1):7–15, 1991.

[74] D. Harman. Relevance feedback revisited. In *Proceedings of the 15th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*, pages 1–10, 1992.

[75] D. Harman. Overview of the second Text REtrieval Conference (TREC-2). In *The 2nd Text REtrieval Conference (TREC-2)*, pages 1–20. NIST Special Publication, 1994.

[76] D. J. Harper. *Relevance feedback in document retrieval system: an evaluation of probabilistic strategies.* PhD thesis, Computer Laboratory, University of Cambridge, UK, 1980.

[77] D. J. Harper and C. J. Van Rijsbergen. An evaluation of feedback in document retrieval using co-occurrence data. *Journal of Documentation*, 34(3):189–216, 1978.

[78] S. P. Harter. A probabilistic approach to automatic keyword indexing (parts I and II). *Journal of the American Society for Information Science*, 26(4,5):197–206,280–289, 1975.

[79] S. P. Harter. Psychological relevance and information science. *Journal of the American Society for Information Science*, 43(9):602–651, 1992.

[80] J. Havrda and F. Charvát. Quantification method of classification processes: concept of structural α-entropy. *Kybernetika*, 3:30–35, 1967.

[81] D. Hawking. Overview of the TREC-9 web track. In *The 9th Text REtrieval Conference (TREC-9)*, pages 87–102. NIST Special Publication, 2001.

[82] M. A. Hearst. Improving full-text precision on short queries using simple constraints. In *Proceedings of the 5th Annual Symposium on Document Analysis and Information Retrieval*, 1996.

[83] M. A. Hearst and C. Plaunt. Subtopic structuring for full-length document access. In *Proceedings of the 16th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*, pages 59–68, 1993.

[84] D. Hindle. Noun classification from predicate-argument structures. In *Proceedings of the 28th Annual Meeting of the Association for Computational Linguistics*, pages 268–275, 1990.

[85] P. G. Hoel. *Introduction to Mathematical Statistics.* Wiley, New York, 4th edition, 1971.

[86] D. A. Hull. Stemming algorithms: A case study for detailed evaluation. *Journal of the American Society for Information Science*, 47(1):70–84, 1996.

[87] E. Ide. New experiments in relevance feedback. In *The SMART retrieval system — experiments in automatic document processing*, pages 337–354. 1971.

[88] P. Ingwersen. A cognitive view of three selected online search facilities. *Online Review*, 8(5):465–492, 1984.

[89] P. Ingwersen. *Information Retrieval Interaction.* Taylor Graham, London, 1992.

[90] E. L. Ivin. *Search procedure based on measures of relatedness between documents.* PhD thesis, M.I.T., 1966. Report MAC-TR-29.

[91] B. J. Jansen, A. Spink, and T. Saracevic. Real life, real users, and real needs: A study and analysis of users on the web. *Information Processing & Management*, 36(2):207–227, 2000.

[92] N. Jardine and R. Sibson. *Mathematical Taxonomy.* John Wiley & Sons Ltd, 1971.

[93] N. Jardine and C. J. Van Rijsbergen. The use of hierarchic clustering in information retrieval. *Information Storage and Retrieval*, 7:217–240, 1971.

[94] Y. Jing and W. Croft. An association thesaurus for information retrieval. In *Proceedings of the RIAO'94 Conference*, pages 146–160, 1994.

[95] K. Kageura. Bigram statistics revisited: A comparative examination of some statistical measures in morphological analysis of Japanese kanji sequences. Internal memo, Natural Language Research Group, Computer Science Department, University of Sheffield, 1997.

[96] H. Kang and K. Choi. Two-level document ranking using mutual information in natural language information retrieval. *Information Processing & Management*, 33(3):289–306, 1997.

[97] P. B. Kantor. Maximum entropy and the optimal design of automated information retrieval systems. *Information Technology: Research and Development*, 3(2):88–94, 1984.

[98] P. B. Kantor and J. J. Lee. The maximum entropy principle in information retrieval. In *Proceedings of the 9th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*, pages 269–274, 1986.

[99] P. B. Kantor and J. J. Lee. Testing the maximum entropy principle for information retrieval. *Journal of the American Society for Information Science*, 49(6):557–566, 1998.

[100] E. M. Keen. Presenting results of experimental retrieval comparisons. *Information Processing & Management*, 28(4):491–502, 1992.

[101] J. Kekäläinen. *The effects of query complexity, expansion and structure on retrieval performance in probabilistic text retrieval.* PhD thesis, University of Tampere, 1999.

[102] M. Kim and K. Choi. A comparison of collocation-based similarity measures in query expansion. *Information Processing & Management*, 35(1):19–30, 1999.

[103] S. Klink, A. Hust, M. Junker, and A. Dengel. Collaborative learning of term-based concepts for automatic query expansion. In *Proceedings of the 13th European Conference on Machine Learning (ECML)*, pages 195–206, 2002.

[104] J. Kristensen. Expanding end-users' query statements for free-text searching with a search-aid thesaurus. *Information Processing & Management*, 29(6):733–744, 1993.

[105] R. Krovetz. Viewing morphology as an inference process. In *Proceedings of the 16th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*, pages 191–202, 1993.

[106] S. Kullback. *Information Theory and Statistics.* Wiley, New York, 1959.

[107] S. Kullback and R. A. Leibler. On information and sufficiency. *Annual Mathematical Statistics*, 22:79–86, 1951.

[108] O. Kwon, M. Kim, and K. Choi. Query expansion using domain-adapted, weighted thesaurus in an extended Boolean model. In *Proceedings of the 3rd International Conference on Information and Knowledge Management (ACM-CIKM)*, pages 140–146, 1994.

[109] J. Lafferty and C. Zhai. Document language models, query models, and risk minimization for information retrieval. In *Proceedings of the 24th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*, pages 111–119, 2001.

[110] M. Lalmas. Dempster-Shafer's theory of evidence applied to structured documents: modelling uncertainty. In *Proceedings of the 20th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*, pages 110–118, 1997.

[111] A. M. Lam-Adesina and G. J. F. Jones. Applying summarization techniques for term selection in relevance feedback. In *Proceedings of the 24th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*, pages 1–9, 2001.

[112] V. Lavrenko and W. B. Croft. Relevance-based language models. In *Proceedings of the 24th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*, pages 120–127, 2001.

[113] M. E. Lesk. Word-word associations in document retrieval systems. *American Documentation*, 20(1):8–36, 1969.

[114] D. D. Lewis and W. B. Croft. Term clustering of syntactic phrases. In *Proceedings of the 13th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*, pages 185–404, 1990.

[115] R. M. Losee, Jr. *The Science of Information Measurement and Applications.* Academic Press, Inc., 1990.

[116] R. M. Losee, Jr. Term dependence: Truncating the Bahadur Lazarsfeld expansion. *Information Processing & Management*, 30(2):293–303, 1994.

[117] R. M. Losee, Jr. and A. Bookstein. Integrating Boolean queries in conjunctive normal form with probabilistic retrieval models. *Information Processing & Management*, 24(3):315–321, 1988.

[118] B. J. Lovins. Development of a stemming algorithm. *Mechanical Translation and Computational Linguistics*, 11:22–31, 1968.

[119] H. P. Luhn. A statistical approach to mechanized encoding and searching of literary information. *IBM Journal of Research and Development*, 1(4):309–317, 1957.

[120] H. P. Luhn. The automatic creation of literature abstracts. *IBM Journal of Research and Development*, 2(3):150–165, 1958.

[121] Y. S. Maarek and F. A. Smadja. Full text indexing based on lexical relations — an application: software libraries. In *Proceedings of the 12th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*, pages 198–206, 1989.

[122] M. Magennis and C. J. Van Rijsbergen. The potential and actual effectiveness of interactive query expansion. In *Proceedings of the 20th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*, pages 324–332, 1997.

[123] R. Mandala, T. Tokunaga, and H. Tanaka. Query expansion using heterogeneous thesauri. *Information Processing & Management*, 36(3):361–378, 2000.

[124] M. E. Maron and J. L. Kuhns. On relevance, probabilistic indexing and information retrieval. *Journal of the Association for Computing Machinery*, 7(3):216–244, 1960.

[125] W. A. Martin. Helping the less experienced user. In *Proceedings of the 6th International Online Meeting*, pages 67–76, 1982.

[126] A. M. Mathai and P. N. Rathie. *Basic Concepts in Information Theory and Statistics.* Wiley, New York, 1975.

[127] D. R. H. Miller, T. Leek, and R. M. Schwartz. A hidden Markov model information retrieval system. In *Proceedings of the 22nd Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*, pages 214–221, 1999.

[128] G. Miller. WordNet: An on-line lexical database. *International Journal of Lexicography*, 3(4):235–244, 1990. Special issue.

[129] J. Minker, G. A. Wilson, and B. H. Zimmerman. An evaluation of query expansion by the addition of clustered terms for a document retrieval system. *Information Storage and Retrieval*, 8:329–348, 1972.

[130] M. Mitra, A. Singhal, and C. Buckley. Improving automatic query expansion. In *Proceedings of the 21st Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*, pages 206–214, 1998.

[131] T. K. Nayak. On diversity measures based on entropy functions. *Communications in Statistics. Theory and Method*, 14(1):203–215, 1985.

[132] N. J. Nilsson. *Learning Machines — Foundations of Trainable Pattern Classifying Systems.* McGraw-Hill, New York, 1965.

[133] Bollmann P. and V. V. Raghavan. On the necessity of term dependence in a query space for weighted retrieval. *Journal of the American Society for Information Science*, 49(13):1161–1168, 1998.

[134] Bollmann P. and S. K. M. Wong. Adaptive linear information retrieval models. In *Proceedings of the 10th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*, pages 157–163, 1987.

[135] H. J. Peat and P. Willett. The limitations of term co-occurrence data for query expansion in document retrieval systems. *Journal of the American Society for Information Science*, 42(5):378–383, 1991.

[136] J. M. Ponte and W. B. Croft. A language modeling approach to information retrieval. In *Proceedings of the 21st Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*, pages 275–281, 1998.

[137] M. F. Porter. An algorithm for suffix stripping. *Program*, 14(3):130–137, 1980.

[138] M. F. Porter. A proposal for writing a probabilistic IR system. *Information Technology: Research and Development*, 1:131–156, 1982.

[139] M. F. Porter and V. Galpin. Relevance feedback in a public access catalogue for a research library. *Program*, 22(1):1–20, 1988.

[140] Y. Qiu. *Automatic query expansion based on a similarity thesaurus.* PhD thesis, Swiss Federal Institute of Technology Zurich, 1995.

[141] Y. Qiu and H. P. Frei. Concept based on query expansion. In *Proceedings of the 16th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*, pages 160–169, 1993.

[142] V. V. Raghavan and S. K. M. Wong. A critical analysis of vector space model for information retrieval. *Journal of the American Society for Information Science*, 37(5):279–287, 1986.

[143] C. R. Rao. Diversity and dissimilarity coefficients: a unified approach. *Journal of Theoretical Population Biology*, 21:24–43, 1982.

[144] C. R. Rao. Diversity: Its measurement, decomposition, apportionment and analysis. *Sankhya: Indian Journal of Statistics*, 44:1–22, 1982.

[145] A. Rényi. On measures of entropy and information. In *Proceedings of the 4th Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 547–561, 1961.

[146] S. E. Robertson. The probabilistic character of relevance. *Information Processing & Management*, 13(13):247–251, 1977.

[147] S. E. Robertson. The probabilistic ranking principle in IR. *Journal of Documentation*, 33(4):294–304, 1977.

[148] S. E. Robertson. Term frequency and term value. *ACM SIGIR Forum*, pages 22–29, 1981.

[149] S. E. Robertson. On relevance weight estimation and query expansion. *Journal of Documentation*, 42(3):182–188, 1986.

[150] S. E. Robertson. On term selection for query expansion. *Journal of Documentation*, 46(4):359–364, 1990.

[151] S. E. Robertson and M. Beaulieu. Research and evaluation in information retrieval. *Journal of Documentation*, 53(1):51–57, 1997.

[152] S. E. Robertson and K. Sparck Jones. Relevance weighting of search terms. *Journal of the American Society for Information Science*, 27(3):129–146, 1976.

[153] S. E. Robertson, C. J. Van Rijsbergen, and Porter. M. F. Probabilistic models of indexing and searching. *Information Retrieval Research*, pages 35–56, 1981.

[154] S. E. Robertson and S. Walker. Some simple effective approximations to the 2-Poisson model for probabilistic weighted retrieval. In *Proceedings of the 17th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*, pages 232–241, 1994.

[155] S. E. Robertson and S. Walker. Okapi/Keenbow at TREC-8. In *The 8th Text REtrieval Conference (TREC-8)*, pages 151–161. NIST Special Publication, 1999.

[156] S. E. Robertson, S. Walker, and M. Beaulieu. Okapi at TREC-7: Automatic ad hoc, filtering, VLC, and interactive track. In *The 7th Text REtrieval Conference (TREC-7)*, pages 253–264. NIST Special Publication, 1999.

[157] S. E. Robertson, S. Walker, S. Jones, M. M. Hancock-Beaulieu, and M. Gatford. Okapi at TREC-3. In *The 3rd Text REtrieval Conference (TREC-3)*, pages 109–126. NIST Special Publication, 1995.

[158] J. J. Rocchio. Relevance feedback in information retrieval. In *The SMART retrieval system — experiments in automatic document processing*, pages 313–323. 1971.

[159] I. Ruthven. Re-examining the potential effectiveness of interactive query expansion. In *Proceedings of the 26th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*, pages 213–220, 2003.

[160] G. Salton. Experiments in automatic thesaurus construction for information retrieval. *Information Processing*, 71:115–123, 1971.

[161] G. Salton. Relevance feedback and the optimisation of retrieval effectiveness. In *The SMART Retrieval System — Experiments in Automatic Document Processing*, pages 324–336. 1971.

[162] G. Salton. *Automatic Text Processing*. Addison Wesley, Reading, 1989.

[163] G. Salton. Developments in automatic text retrieval. *Science*, 253:974–980, 1991.

[164] G. Salton and C. Buckley. Term weighting approaches in automatic text retrieval. *Information Processing & Management*, 24(5):513–523, 1988.

[165] G. Salton and C. Buckley. Improving retrieval performance by relevance feedback. *Journal of the American Society for Information Science*, 41(4):288–297, 1990.

[166] G. Salton, E. A. Fox, and E. Voorhees. Advanced feedback methods in information retrieval. *Journal of the American Society for Information Science*, 36(3):200–210, 1985.

[167] G. Salton and M. H. McGill. *Introduction to Modern Information Retrieval*. McGraw-Hill, New York, 1983.

[168] G. Salton, E. Voorhees, and E. A. Fox. A comparison of two methods for Boolean query relevance feedback. *Journal of the American Society for Information Science*, 36:200–210, 1985.

[169] G. Salton, A. Wong, and C. S. Yang. A vector space model for automatic indexing. *Communications of the ACM*, 18:613–620, 1975.

[170] G. Salton, A. Wong, and C. T. Yu. Automatic indexing using term discrimination and term precision measurement. *Information Processing & Management*, 12:43–51, 1976.

[171] G. Salton and C. S. Yang. On the specification of term values in automatic indexing. *Journal of Documentation*, 29(4):351–372, 1973.

[172] G. Salton, C. S. Yang, and C. T. Yu. A theory of term importance in automatic text analysis. *Journal of the American Society for Information Science*, 26(1):33–44, 1975.

[173] T. Saracevic. Relevance: A review of and framework for the thinking on the notion in information science. *Journal of the American Society for Information Science*, 26(6):321–343, 1975.

[174] T. Saracevic. Relevance reconsidered '96. In *Proceedings of the 2nd International Conference on Conceptions of Library and Information Science*, pages 201–218, 1996.

[175] H. Schütze and J. O. Pedersen. A cooccurrence-based thesaurus and two applications to information retrieval. *Information Processing & Management*, 33(3):307–318, 1997.

[176] C. E. Shannon. A mathematical theory of communication. *Bell System and Technical Journal*, 27:379–423,623–656, 1948.

[177] R. Sibson. Information radius. *Z. Wahrsch'theorie and verw. Geb*, 14:149–160, 1969.

[178] W. T. Silva and R. L. Milidiú. Belief function model for information retrieval. *Journal of the American Society for Information Science*, 44(1):10–18, 1993.

[179] A. Singhal, J. Choi, D. Hindle, D. Lewis, and F. Pereira. AT&T at TREC-7. In *The 7th Text REtrieval Conference (TREC-7)*, pages 239–252. NIST Special Publication, 1999.

[180] A. F. Smeaton and C. J. Van Rijsbergen. The nearest neighbor problem in information retrieval. In *Proceedings of the 4th Annual International Conference on Information Storage and Retrieval*, pages 83–87, 1981.

[181] A. F. Smeaton and C. J. Van Rijsbergen. The retrieval effects of query expansion on a feedback document retrieval system. *The Computer Journal*, 26(3):239–246, 1983.

[182] F. Song and W. B. Croft. A general language model for information retrieval. In *Proceedings of the 22nd Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*, pages 279–280, 1999.

[183] K. Sparck Jones. The use of automatically-obtained keyword classification for information retrieval. *Information Storage and Retrieval*, 5(3):175–201, 1970.

[184] K. Sparck Jones. *Automatic Keyword Classification for Information Retrieval*. Butterworth, 1971.

[185] K. Sparck Jones. A statistical interpretation of term specificity and its application to retrieval. *Journal of Documentation*, 28(1):11–21, 1972.

[186] K. Sparck Jones. Collection properties influencing automatic term classification performance. *Information Storage and Retrieval*, 9(9):499–513, 1973.

[187] K. Sparck Jones. Experiments in relevance weighting of search terms. *Information Processing & Management*, 15(3):133–144, 1979.

[188] K. Sparck Jones. Search term relevance weighting given little relevance information. *Journal of Documentation*, 35(1):30–48, 1979.

[189] K. Sparck Jones. Search term relevance weighting — some recent results. *Journal of Information Science*, 1(6):325–332, 1980.

[190] K. Sparck Jones. A look back and a look forward. In *Proceedings of the 11th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*, pages 13–29, 1988.

[191] K. Sparck Jones. Reflection on TREC. *Information Processing & Management*, 31(3):291–314, 1995.

[192] K. Sparck Jones and E. O. Barber. What makes an automatic keyword classification effective? *Journal of the American Society for Information Science*, 22(3):166–175, 1971.

[193] K. Sparck Jones and R. G. Bates. Research on automatic indexing 1974-1976. Technical report, Computer Laboratory, University of Cambridge, 1977.

[194] K. Sparck Jones and R. M. Needham. Automatic term classification and retrieval. *Information Storage & Retrieval*, 4(1):91–100, 1968.

[195] K. Sparck Jones and C. A. Webster. Research on relevance weighting 1976-1979. British Library Report 5553, Cambridge, UK, 1980.

[196] P. Srinivasan. Query expansion and MEDLINE. *Information Processing & Management*, 32(4):431–443, 1996.

[197] H. E. Stiles. The association factor in information retrieval. *Journal of the Association for Computing Machinery*, 8(2):271–279, 1961.

[198] D. R. Swanson. Historical note: information retrieval and the future of an illusion. *Journal of the American Society for Information Science*, 39(4):92–98, 1988.

[199] J. M. Tague. A Bayesian approach to interactive retrieval. *Information Storage and Retrieval*, 9(3):129–142, 1973.

[200] TREC Eval., 1992. "The trec eval" program is available via ftp from the SMART site at Cornell University. Available at: ftp://fp.cs.cornell.edu/pub/smart.

[201] H. Turtle and W. B. Croft. Inference networks for document retrieval. In *Proceedings of the 13th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*, pages 1–24, 1990.

[202] H. Turtle and W. B. Croft. A comparison of text retrieval models. *The Computer Journal*, 35(3):279–290, 1992.

[203] M. J. Usher. *Information Theory for Information Technologists.* Macmillan, London, 1984.

[204] C. J. Van Rijsbergen. *Automatic information structuring and retrieval.* PhD thesis, University of Cambridge, 1972.

[205] C. J. Van Rijsbergen. Further experiments with hierarchic clustering in document retrieval. *Information Storage & Retrieval*, 4(1):1–14, 1974.

[206] C. J. Van Rijsbergen. A theoretical basis for the use of co-occurrence data in information retrieval. *Journal of Documentation*, 33(2):106–119, 1977.

[207] C. J. Van Rijsbergen. *Information Retrieval.* Butterworths, London, 2nd edition, 1979.

[208] C. J. Van Rijsbergen. A non-classical logic for information retrieval. *The Computer Journal*, 39:481–485, 1986.

[209] C. J. Van Rijsbergen. Towards a new information logic. In *Proceedings of the 12th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*, pages 77–86, 1989.

[210] C. J. Van Rijsbergen, D. J. Harper, and M. F. Porter. The selection of good search terms. *Information Processing & Management*, 17:77–91, 1981.

[211] C. J. Van Rijsbergen and K. Sparck Jones. A test for the separation of relevant and non-relevant documents in experimental retrieval collections. *Journal of Documentation*, 29(3):251–257, 1973.

[212] E. M. Voorhees. The efficiency of inverted index and cluster searches. In *Proceedings of the 9th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*, pages 164–174, 1986.

[213] E. M. Voorhees. Using WordNet to disambiguate word senses for text retrieval. In *Proceedings of the 16th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*, pages 171–180, 1993.

[214] E. M. Voorhees. Query expansion using lexical-semantic relations. In *Proceedings of the 17th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*, pages 61–69, 1994.

[215] E. M. Voorhees and D. Harman. Overview of the seventh Text REtrieval Conference (TREC-7). In *The 7th Text REtrieval Conference (TREC-7)*, pages 1–23. NIST Special Publication, 1999.

[216] E. M. Voorhees and Y. W. Hou. Vector expansion in a large collection. In *The 1st Text REtrieval Conference (TREC-1)*, pages 343–351. NIST Special Publication, 1993.

[217] H. D. White. Toward automated search strategies. In *Proceedings of the 13th International Online Information Meeting*, pages 12–14, 1989.

[218] K. V. M. Whitney. Minimal spanning tree. *Communications of the Association for Computing Machinery*, 15(4):273–274, 1972.

[219] R. Wilkinson. Effective retrieval of structured documents. In *Proceedings of the 17th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*, pages 311–317, 1994.

[220] P. Willett. An algorithm for the calculation of exact term discrimination values. *Information Processing & Management*, 21(3):225–232, 1985.

[221] S. K. M. Wong and Y. Y. Yao. A probability distribution model for information retrieval. *Information Processing & Management*, 25(1):39–53, 1989.

[222] S. K. M. Wong and Y. Y. Yao. Query formulation in linear retrieval models. *Journal of the American Society for Information Science*, 41(5):334–341, 1990.

[223] S. K. M. Wong and Y. Y. Yao. A probabilistic inference model for information retrieval. *Information Systems*, 16(3):301–321, 1991.

[224] S. K. M. Wong and Y. Y. Yao. An information-theoretic measure of term specificity. *Journal of the American Society for Information Science*, 43(1):54–61, 1992.

[225] S. K. M. Wong, Y. Y. Yao, and P. Bollmann. Linear structure in information retrieval. In *Proceedings of the 11th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*, pages 219–232, 1988.

[226] S. K. M. Wong, Y. Y. Yao, G. Salton, and C. Buckley. Evaluation of an adaptive linear model. *Journal of the American Society for Information Science*, 42(10):723–730, 1991.

[227] S. K. M. Wong, W. Ziarko, V. V. Raghavan, and P. C. N. Wong. On extending the vector space model for Boolean query processing. In *Proceedings of the 9th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*, pages 175–185, 1986.

[228] S. K. M. Wong, W. Ziarko, and P. C. N. Wong. Generalised vector space model in information retrieval. In *Proceedings of the 8th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*, pages 18–25, 1985.

[229] H. Wu and G. Salton. A comparison of search term weighting: term relevance vs. inverse document frequency. *ACM SIGIR Forum*, pages 30–39, 1981.

[230] J. Xu. *Solving the word mismatch problem through automatic text analysis.* PhD thesis, University of Massachusetts at Amherst, 1997.

[231] J. Xu and W. B. Croft. Query expansion using local and global document analysis. In *Proceedings of the 19th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*, pages 4–11, 1996.

[232] J. Xu and W. B. Croft. Cluster-based language models for distributed retrieval. In *Proceedings of the 22nd Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*, pages 254–261, 1999.

[233] C. T. Yu, C. Buckley, H. Lam, and G. Salton. A generalized term dependence model in information retrieval. *Information Technology: Research and Development*, 2(4):129-154, 1983.

[234] C. T. Yu, W. Meng, and S. Park. A framework for effective retrieval. *ACM Transactions on Database Systems*, 14(2):147–167, 1998.

[235] C. T. Yu and G. Salton. Precision weighting — an effective automatic indexing method. *Journal of the Association for Computing Machinery*, 23(1):76–88, 1976.

[236] C. T. Yu and G. Salton. Effective information retrieval using term accuracy. *Journal of the Association for Computing Machinery*, 20(3):135–142, 1977.

[237] G. U. Yule. *The Statistical Study of Literary Vocabulary.* Cambridge University Press, Cambridge, 1944.

[238] C. Zhai and J. Lafferty. Model-based feedback in the language modeling approach to information retrieval. In *Proceedings of the 24th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*, pages 403–410, 2001.

[239] C. Zhai and J. Lafferty. A study of smoothing methods for language models applied to ad hoc information retrieval. In *Proceedings of the 24th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*, pages 334–342, 2001.

[240] G. K. Zipf. *Human behavior and the principle of least effect.* Addison-Wesley, Reading, MA, 1949.