# Probability Models for Information Retrieval based on Divergence from Randomness

Giambattista Amati

Thesis submitted for the degree of Doctor of Philosophy,

Department of Computing Science

Faculty of Information and Mathematical Sciences

University of Glasgow

June 2003

# Abstract

This thesis devises a novel methodology based on probability theory, suitable for the construction of term-weighting models of Information Retrieval. Our term–weighting functions are created within a general framework made up of three components. Each of the three components is built independently from the others. We obtain the term–weighting functions from the general model in a purely theoretic way instantiating each component with different probability distribution forms.

The underpinning idea on which we are able to systematically construct the term–weighting models is based on the notion of divergence from randomness. The leading theme of the divergence-from-randomness approach is that the informative content of a term can be measured by examining how much the term-frequency distribution departs from a "benchmark" distribution, that is the distribution described by a random process.

Following this idea, the first two components of the framework provide an explanation to the duality existing in Information Retrieval between the distributions of topic–terms in a small set of documents (the elite set of a topic) and in the rest of the collection. The third component deals with the term-frequency normalization and is able to compare term frequencies within documents of different lengths. As a consequence, different probability distributions can be used in the framework of the divergence-from-randomness approach. Our experiments utilise some of them to show that the framework is sound and robust and generates different but highly effective Information Retrieval models.

The thesis begins with investigating the nature of the statistical inference involved in Information Retrieval. We explore the estimation problem underlying the process of sampling. De Finetti's theorem is used to show how to convert the frequentist approach into Bayesian inference and we display and employ the derived estimation techniques in the context of Information Retrieval.

We initially pay a great attention to the construction of the basic sample spaces of Information Retrieval. The notion of single or multiple sampling from different populations in the context of Information Retrieval is extensively discussed and used throughout the thesis. The language modelling approach and the standard probabilistic model are studied under the same foundational view and are experimentally compared to the divergence-from-randomness approach.

In revisiting the main information retrieval models in the literature, we show that even language modelling approach can be exploited to assign term-frequency normalization to the models of divergence from randomness.

We finally introduce a novel framework for the query expansion. This framework is based on the models of divergence-from-randomness and it can be applied to arbitrary models of IR, divergence-based, language modelling and probabilistic models included.

We have done a very large number of experiments and results show that the framework generates highly effective Information Retrieval models.

# Acknowledgments

I would like to express my gratitude to my supervisor, Prof. Cornelis Joost Van Rijsbergen, for his scientific input, knowledge, advice, understanding, feedback, generosity and for allowing me to freely explore new ideas.

I am grateful to Dr. Roderick Murray-Smith for his kindness and optimism.

I am especially indebted to Juliet Van Rijsbergen who greatly improved the readability of the manuscript by adding scientific comments, providing suggestions for improving its structure, and even going through the mathematics and pointing out where my thoughts and explanations were unclear.

I thank the Committee members, Prof. Norbert Führ, Dr. Ronald R. Poet and Dr. Lewis M. Mackenzie for having appreciated my work.

And finally my thanks to my wife, Carmen, who encouraged me to believe only the important things in my life.

# Legenda

Symbols and notations

| | |
|---|---|
| $D$ | a text collection |
| $t$ | a term |
| $q$ | a query |
| $d$ | a document |
| $w(q\|d)$ | the weight of the query $q$ given the document $d$ |
| $tf_q$ | the term–frequency of $t$ in the query $q$ |
| $tf$ | the term–frequency of $t$ in the document $d$ |
| $E$ | a sample of the collection |
| $E_q$ | *the elite set of the query*, the set of topmost documents satisfying the query $q$ according to the weight $w(q\|\_)$ |
| $E_t$ | *the elite set of the term*, the set of documents containing the term $t$ |
| $N$ | the number of documents in the collection $D$ |
| $avg\_l$ | the average length of a document in the collection |
| $l, l_d$ | the length of the document $d$ |
| $F, F_t, F_E$ | the total number of tokens of $t$ in the collection, in $E_t$, and in an arbitrary subset $E$ |
| $TotFr_D, TotFr_E$ | the total number of tokens in the collection $D$ and in a subset $E$ of $D$ |
| $p_D, p_D(t)$ | the relative frequency $\frac{F}{TotFr_D}$ of $t$ in the collection |
| $p_d, p_d(t)$ | the relative frequency $\frac{tf}{l_d}$ of $t$ in the document |
| $n, n_t$ | the document-frequency, the cardinality of $E_t$, $n = n_t =\mid E_t \mid$ |

Symbols and notations

| | |
|---|---|
| $B(F,k,p)$ | the binomial distribution of $F$ trials with probability $p$ of success and $k$ successes |
| $n_e$ | the number of documents containing a term according to the binomial distribution, $N \cdot (1 - B(F_t, 0, p))$ |
| $r$, $r_t$ | the number of relevant documents containing the term $t$ |
| $R$ | the number of relevant documents of a query |
| $\mu$ | the parameter of the Dirichlet priors |
| $\alpha$ | the parameter of the query expansion |
| $c$ | the parameter of the term frequency normalization $H2$ |
| $D$ | the Divergence of two distributions |
| $\chi$ | the $\chi$ divergence of two distributions |
| $KL$ | the Kullback-Leibler divergence of two distributions |
| $Inf(t\|E)$, $Inf_E(t)$ | the informative content of the term in $E$ |

Basic Divergence-based Models

| | |
|---|---|
| $D$ | the Divergence approximation of the binomial |
| $P$ | the Poisson approximation of the binomial |
| $B_E$ | the Bose-Einstein distribution |
| $G$ | the geometric approximation of the Bose-Einstein |
| $I(n)$ | the Inverse Document Frequency model |
| $I(F)$ | the Inverse Term Frequency model |
| $I(n_e)$ | the Inverse Expected Document Frequency model |

First Normalization Models

| | |
|---|---|
| $L$ | the Laplace normalization |
| $B$ | the Bernoulli ratio normalization |

Second Normalization: Term Frequency Normalization

| | |
|---|---|
| $H1$ | the uniform distribution of term frequencies |
| $H2$ | the logarithmic normalization |
| $H3$ | the Dirichlet normalization |
| $Z$ | the Zipfian normalization |

Normalized Models

| | |
|---|---|
| $DL1$ | the divergence basic model $D$, normalized by Laplace normalization $L$ and by term frequency normalization **H1** |
| ⋮ | ⋮ |
| $I(n_e)BZ$ | the Inverse Expected Document Frequency basic model, normalized by the Bernoulli ratio normalization $B$ and by the Zipfian term frequency normalization $Z$ |

# Contents

# List of Figures

# List of Tables

# Chapter 1

# Theoretical Information Retrieval

This dissertation devises a methodology based on probability theory, suitable for the construction of models of Information Retrieval (IR). The term *information retrieval* refers to a very practical problem. We imagine a user who wishes to retrieve all the most relevant documents from a text collection. A system of IR has the task of producing an ordered presentation of documents in decreasing weight of relevance in response to the user inquiry. The kernel of an IR system is thus the *model*, that is the theoretical component which leads to the determination of the document–ranking. In short, we can say that IR modelling finds solutions to the inductive problem of predicting the relevance of a document to a query. Although the notion of relevance remains a most difficult, subjective and controversial notion to be defined [21], we postpone this problem and start with the frequentist approach of the statistician. The statistical data for IR comes from observations on the distribution of word occurrences within documents and over the entire collection. In statistics observations from empirical data are explained by distributions for which an exact mathematical form exists. Once the hypothesis on the type of distribution is formulated and successfully tested, we are in the position to estimate the values of the inherent parameters of the distribution-form. Consequently, the first general problem which may be stated by a statistical approach to Information Retrieval is that of determining the distribution-form of the word–frequencies.

Early works on IR focused on this problem [30, 113, 14, 52, 53, 54, 13], and the form distribution of the word frequencies was indeed found to be the 2-Poisson model by Harter [52]. The 2-Poisson model is a mixture [114] of two Poisson distributions.

Its generalization, the mixture of $N$ Poisson distributions, has been also studied more recently [110, 76].

Although the solution to this statistics problem has been known since the 70's, the exact connection of this result to the fundamental inductive problem of modelling relevance remains an open problem of Information Retrieval.

The source of the difficulty is that words distribute over a collection according to the *meaning*, so that an occurrence of a word attracts the occurrence of the same or different words with a probability value which changes significantly over the document–collection. Following the terminology of statistics, a document is a sample of an unknown population and large variations of term-frequency from one document to another may witness a change of the population from which we are sampling. Two arbitrary documents should be in principle considered samples belonging to different populations, because they are, in general, dealing with different subjects and topics, so terms in them appear rarely or frequently depending on their principal content. Instead, the most difficult question is when two documents can be regarded to be *samples of the same population.*

In general we have to accept that the *a priori* probability of occurrence of a word in *an arbitrary* document is not the same as the probability of occurrence of a word in *a given* document, since every document reflects some unknown population and has always a specific topic treated at a greater level of detail. In principle, there are $2^{|V|}$ possible "topics" or "queries" $q \in 2^V$ generated by a vocabulary $V$ of cardinality $|V|$. Each topic $q$ possesses an "*elite set*" $E_q$, with $E_q \subset 2^D$ possibly empty, describing at a greater level of detail the content of that topic, and all documents of this elite set can be regarded by a first approximation as if we were sampling from the *same population.* It would be as we had a single large document instead of a set of documents. This is what happens when we look at the content of the documents returned by any search engine. In many cases, this elite set can be pooled forming almost an homogeneous and coherent piece of text. An elite set can be considered as a set of samples from the same population, that is the population relative to the submitted query.

The notion of eliteness was first introduced by Harter [52, pages 68-74] to explain the 2-Poisson model. According to Harter, the idea of eliteness is used to reflect the level of treatment of a word in a certain small set of documents compared with the rest of the

collection. The main characterization of the elite set of a topic, is that a word occurs to a relatively greater extent than in all other documents. Harter defines eliteness through a probabilistic estimate which is interpreted as the set of documents which a human indexer assesses to be elite with respect to a word.

Once the notion of eliteness is ostensively characterized by a set of documents that are sharing a specific topic as principal subject of their content, as if we were sampling from the same population, we might find it difficult to distinguish eliteness from *relevance*. Unlike eliteness, which is implicitly defined by word–frequency distributions, relevance is a primitive concept, similarly to "Truth" in Logic, and it defines a binary relationship between queries and documents. The relevance relationship is defined by the user, by common sense or by experts and it holds or not for each query-document pair independently from the informative content contained in the rest of the collection. Therefore, relevance should be conceived as an external notion to the IR model. The treatment of relevance as an external feature of the system [27] is not, in general, accepted. For example, the standard probabilistic model [90, 91, 86] or the $BM25$ formula [87], which is one of the most used model of IR, has a user's relevance feedback mechanism incorporated into the model (see Section 5.2.2 in Chapter 5).

Our position in this dissertation is that relevance mainly concerns the evaluation of effectiveness of IR systems and it is the user-based or user-perceived counterpart of eliteness.

Beside the application of relevance to the evaluation task, the feedback on relevance received by the users or provided by the test collections can be processed as further observations. With user's feedback the relevance may come into the estimation problem with a set of unknown parameters. The estimation of such parameters with relevance data defines a *parametric* approach to IR modelling.

In a parametric approach data are assumed to be *incomplete* since the knowledge on relevance is provided by only a small number of query samples. A frequently used estimation technique consists in determining the values of the unknown parameters, for example maximizing the measure of the retrieval performance with a set of test queries. This estimation methodology is called the *Best Match* parametric method.

A "parameter–free" IR model instead does not possess parameters which need to be

learned from observations on relevance provided by the test queries.

An example of a non–parametric model is the vector space model (see Section 5.1). An example of a parametric model is the $BM25$ formula. The vector space model was outperformed by the more recent model $BM25$, but language modelling is now an emerging proposal for IR modelling [82, 60, 70, 132, 133]. However, $BM25$ still remains the most popular model among the participants of TREC, the main Conference dedicated to the evaluation of IR systems [46, 49, 50, 122, 123, 121].

## 1.1 The intention of this Thesis

The main objective of this dissertation is the definition of a novel methodology suitable for theoretical derivation of models of Information Retrieval.

Beside this, we aim at deriving parameter–free models of IR, where the term "parameter–free" refers to the absence of parameters whose estimation depends on relevance but not from the text collection only.

Our second objective is to investigate the nature of the statistical inference involved in Information Retrieval. This investigation is not only of theoretical interest but of practical purpose. Our initial motivation to undertake a theoretical approach to IR was that only a well founded theory could have led to the construction of highly effective IR models. Our term–weighting functions are thus created within a general framework, that is a "super-model" or, borrowing the term from the terminology of Logic, a "second-order model" of IR. This framework is made up of three components. Each of the three components is built independently from the others. These term–weighting functions are thus derived in a purely theoretic way from the general model instantiating each component with different probability distribution forms. The first two components provide an explanation to the duality existing in IR between the distributions of terms in the elite and non-elite sets of documents with respect to given topics. The third component is the term-frequency normalization, that is a component which is able to compare frequencies within documents of different lengths. As a consequence of our approach, its second most important feature is that different probability distributions can be used in the framework. Our experiments utilise them to show that the framework is sound and robust and generates different but highly effective IR models. This theoretical framework

has been successfully tested in the TREC-10 (see Table 7.13 at page 159) Conference [3] producing the best performing run at the WEB track (see Table 1.2 at page 39). In Chapter 8 we show that our theory of query expansion based on the divergence from randomness has further improved results (see Table 8.4 at page 180).

In summary, our theory devices the construction of a class of effective probability based retrieval models rather than a single retrieval model.

The basic idea of divergence from randomness is also applied to generate a query expansion framework. The first component of basic models of IR is used to weight terms of a new formulation of the original query. This time, the other two components are not needed because query expansion is reduced to the problem of having a unique sample relatively to an unknown population.

The first three components of our framework are briefly introduced in the following sections.

## 1.2 The origins of the proposal

Our proposal is strongly influenced by works on *automatic indexing* by Damerau, Bookstein, Swanson and Harter [30, 14, 53, 54]. These early models for automatic indexing were based on the distinction of the words into two complementary classes. There is the class of the *function* words, which have only a syntactical or modal role in the text, and the class of *specialty* words, which are informative content words. The function word distribution is closely modelled by a Poisson process, whilst specialty word–frequencies deviate from a distribution of a Poisson form. The specialty words appear more densely in a few "elite" documents, whereas function words, which are included in a list of words called *a stop list*, are *randomly* distributed over the collection, as predicted by a Poisson distribution with a mean of $\lambda$.

According to these early linguistic models a testable hypothesis is that the informative content of a word can be measured by examining how much the word-frequency distribution departs from a "benchmark" distribution, that is a distribution of non–informative words, in, for example, a Poisson distribution. *This is the underpinning idea on which we are able to systematically construct the models of IR based on the divergence from randomness.*

To exemplify our position we take our point of view from Stone and Bookstein [13]

> [...] a content bearing word is taken to be one whose appearance in a document does serve to distinguish it from other documents and which thus occurs *nonrandomly.*

We assume that *the more the word distribution does not fit the probabilistic model predicting a random appearance of the word, the more informative is.*

Harter assumed that a specialty word follows a second Poisson distribution on the elite set, obviously with a mean frequency $\mu$ greater than the mean frequency $\lambda$ in the rest of the collection. His model was able to assign 'sensible' index terms and was tested using a very small data collection and a few randomly chosen specialty words. Srinivasan and Margulis [110, 76] corroborated this finding on the N-Poisson distributions using a more robust experimentation.

However, eliteness is a hidden variable since it cannot be known in advance, and the estimation of the mean $\mu$ is thus problematic.

Harter's work was designed for automatic indexing, which concerns the automatic assignment of keywords to documents, but his proposal was not so general as to be included in any effective retrieval function.

In our work, we start again from the same viewpoint as these early works and develop the foundation in order to provide a well-founded theory for constructing models of Information Retrieval. We do not try to start from scratch, for we agree that, it is quite intuitive to believe that a good automatic indexing function, like that of Harter, can be exploited as a good term–weighting function. Indeed, the potential effectiveness of Harter's model for a direct exploitation of eliteness in retrieval was explored by Robertson, Van Rijsbergen, Porter, Williams and Walker [91, 87] who plugged the Harter 2-Poisson model [53] into the standard probabilistic model of Robertson and Sparck Jones [90] (see Section 5.2).

Robertson, Van Rijsbergen and Porter used notions of both eliteness and relevance. The evolution of the 2-Poisson model as designed by Robertson, Van Rijsbergen and Porter has motivated the birth of a family of term-weighting forms called $BM$s ($BM$ for Best Match) [87] (see Section 5.2.2 of Chapter 5). The most successful formula of this family, the $BM25$, was introduced in 1994 [87].

Although the 2–Poisson distribution may have inspired the $BM25$, the $BM25$ formula cannot quite be considered the retrieval counterpart of the 2-Poisson model, or more generally the natural evolution of a theory of eliteness for retrieval.

First, the $BM25$ formula was not formally derived from the Poisson model but it was introduced as a limit and a simplified version of the Robertson, Van Rijsbergen and Porter model (see the discussion in Section 5.2.2).

Second, the $BM25$ contains many parameters which need to be tuned from data on relevance. The rationale for their introduction is empirical and therefore the nature of these parameters is unknown and finding well-founded estimates for the parameters remains a problem. The solution is presented in Section 7.1, where we derive the unexpanded version of the $BM25$ formula from one of our models, the $I(n)L2$ model (see Section 7.1). Since $I(n)L2$ is a parameter–free model, we have also formally derived the empirical values of the unknown parameters of the $BM25$. It is surprising but satisfying to discover that the derived values for these parameters are very close to the default values of the $BM25$, which have been acquired by means of empirical data on relevance within the Best Match method. This result is also an evidence which corroborates our foundational theory.

Third, our aim is to model the inference process involved in IR. From the theory of our framework, we can explain why the $BM25$ has been considered for many years one of the most effective matching functions for document retrieval. We however make clear that in a class of highly performing models, of which the $BM25$ is an instance, other models can very often perform better than the $BM25$.

We may even generate a more effective version of the $BM25$ by using the third component of our framework. We show that, enlarging the magnitude of the document-length on which we compare the term-frequencies inside the third component of the framework, we derive more effective models of IR. This means that $BM25$ is incomplete. An improved version of the $BM25$ may be derived by the $I(n)L2$ model but with this different document-length (see the discussion in Section 6.2.1).

## 1.3 The generating term–weighting formula

We show that the weight of a term occurring $tf$ times in a document is a function of two probabilities $Prob_1(tf|D)$ and $Prob_2(tf|E_t)$ which are related by the following relation:

$$w = (1 - Prob_2(tf|E_t)) \cdot (-\log_2 Prob_1(tf|D)) = -\log_2 Prob_1(tf|D)^{1 - Prob_2(tf|E_t)} \quad (1.1)$$

where $D$ is the document–collection of size $N$ and $E_t$ is the elite set of the term. The term–weight is thus a decreasing function of both probabilities $Prob_1$ and $Prob_2$. In Relation 1.1 we assume that the elite set $E_t$ of $t$ is more simply the set of all documents containing the term. *Therefore from now on, the subscript t in elite set $E_t$ denotes the elite set of a term in our sense and not in Harter's sense.*

**Definition 1** The elite set $E_t$ of a term $t$ is the set of documents containing $t$.

The term–weighting function $w$ derived with the first two components is a function of four random variables, that is

$$w = w(tf, F_t, n_t, N)$$

where $F_t$ is the number of occurrences of the term in a collection of $N$ documents, $n_t$ the cardinality of the elite set of the term and $tf$ is the within–document term–frequency.

## 1.4 The first component: informative content

The distribution $Prob_1$ is introduced with similar arguments to those used by Harter. We suppose that terms which convey little information, are randomly distributed *on the whole set of documents.* We provide different basic probabilistic models, with a probability distribution $Prob_1$, that defines the notion of *randomness in the context of Information Retrieval.* We propose to define as models of randomness all probabilistic processes which use random drawings from urn models or random placement of coloured balls in urns. Instead of *urns* we have *documents*, and instead of different *colours* we have different *terms*, where one term $t$ can occur with a multiplicity $F$ in these set of urns as anyone of a number of related words or phrases which are called *tokens* of that term. We thus offer different processes as basic models of randomness. Among these processes, we

study the binomial distribution and its approximations, and the Bose-Einstein statistics and its approximations, and also the inverse document-frequency model and some of its variants.

The component of the weight of Formula 1.1

$$Inf_1 = -\log_2 Prob_1 \tag{1.2}$$

is defined as the *informative content* $Inf_1$ of the term in the document. The definition of informative content as defined in Relation 1.2, that is $-\log_2 Prob$, has appeared in semantic information theory [18, 10, 11, 62, 63] but the idea was actually that of Popper [83] (see Chapter 4). Fano used the term *self-information* for it [36].

A function which is decreasing monotonic with respect to $Prob_1$ and additive with respect to independent events is unique and is $Inf_1$ up to a multiplicative factor [23, 124].

Since $Prob_1$ is the probability of having *by chance*, according to the chosen model of randomness, $tf$ occurrences of a term $t$ in a document $d$, the smaller this probability, the less its tokens are distributed in conformity with the model of randomness, and therefore the higher the informative content of the term. For example, any tautology is the certain event but it is trivially informative.

Determining the informative content of a term can be conceived as an inverse test of randomness, that is as a measure of the extent the term distribution in the document departs from the random one. A uniform distribution over the space of events defines the random distribution. Obviously the elementary events of the space may be defined differently leading to have different models of randomness for Information Retrieval. This explains why we pay a lot of attention at the beginning in Chapter 2 to giving a clear definition of what is meant by the space of events in IR, since once the event space is circumscribed and we have clarified what are the samples and the populations involved in our inductive problem half of our work will be almost done.

Popper [83] gave an alternative definition to the informative content:

$$Inf = 1 - Prob_1$$

With this in mind and looking at the fundamental Formula 1.1 we can see it as the product of two informative content functions, the first function $Inf_1$ being related to the

whole document collection $D$ and the second one $Inf_2$ to the elite set $E_t$ of the term:

$$w = \quad Inf_1 \cdot Inf_2 \tag{1.3}$$

The factor $1 - Prob_2$ of Formula 1.1 is called the *First Normalization of the informative content* $Inf_1$.

### 1.4.1 An exemplification of informative content: Bernoulli model of divergence from randomness

We can illustrate the informative content defining just one of the models of randomness, that is the *Bernoulli model of randomness*, also called the *Binomial model of randomness.* The other models, which are fully introduced in Chapter 4 can be obtained with the same construction but varying the underlying sample space.

We define $Prob_1$ in the weighting function 1.1 by an example. Suppose that a lift is serving a building of 1024 floors and that 10 people take the lift at the basement floor independently of each other. Suppose that these 10 people have not arrived together. We assume that there is a uniform prior probability that a person gets off at a particular floor. The probability that any 4 people out of the 10 leave at any floor is

$$B\left(1024, 10, 4\right) = \left( \begin{array}{c} 10 \\ 4 \end{array} \right) p^4 q^6 = 0.00000000019$$

where $B$ is the binomial law, and $p = \dfrac{1}{1024}$ and $q = \dfrac{1023}{1024}$.

The informative content function 1.2 is

$$-\log_2 B\left(1024, 10, 4\right) = 32.29$$

This toy problem is abstractly equivalent to the IR problem. It is sufficient to change the terminology by replacing "floor" with "document", "people" with "tokens of the same term","leave" with "occur". The *term independence assumption* corresponds to the fact that people in the lift have not arrived together, or equivalently there is not a common cause which has brought any group of these people at the same time to take that lift. If $F$ is the total number of tokens of an observed term $t$ in a collection $D$ of $N$ documents, then we make the assumption that the tokens of a non-informative terms should distribute over the $N$ documents according to the binomial law.

In the Bernoulli model of a document $d$ we regard each token of a term as a trial of an *experiment.* A successful outcome for the term $t$ is when $t$ occurs in the document $d$. The *a priori* probability of success is the probability of retrieving the document $d$, which, in absence of knowledge, is obtained from a uniform distribution $p = \frac{1}{N}$.

Therefore the probability of $tf$ occurrences in a document (successes) is given by

$$Prob_1(tf) = Prob_1 = B(N, F, tf) = \begin{pmatrix} F \\ tf \end{pmatrix} p^{tf} q^{F-tf}$$

where $p = \frac{1}{N}$ and $q = \frac{N-1}{N}$.

Hence, the terms in a document with the highest probability $Prob_1$ of occurrence as predicted by such models of randomness are "non-specialty" terms. Equivalently, the terms whose probability $Prob_1$ of occurrence conforms most to the expected probability given by the basic models of randomness are non content bearing terms. Conversely, terms with the smallest expected probability $Prob_1$ are those which provide the *informative content* of the document.

Figure 1.1 shows the informative content of the term "osteoporosis" within the documents of the collection WT2g of TREC-8. The term occurs in 85 documents. Notice that the term-weights decrease very rapidly when the within–document term–frequency $tf$ diminishes.

## 1.5 The second component: apparent aftereffect of sampling

Now, we introduce the role played by the probability denoted by $Prob_2$ in the fundamental Equation 1.1. The informative words are rare in the collections but, in compensation, when they occur their frequency is very high. More specifically, the 2-Poisson model captures such a duality law. There is a statistical phenomenon called by statisticians an apparent *aftereffect* of sampling. It may happen that a sudden repetition of success of a rare event increases our expectation of a further success to almost certainty. Laplace's law of succession is one of the possible estimates of such an expectation.

Similarly, the 2-Poisson model of IR can be explained by an aftereffect phenomenon. As already observed, all informative terms $t$ occur to a relatively greater extent in a set

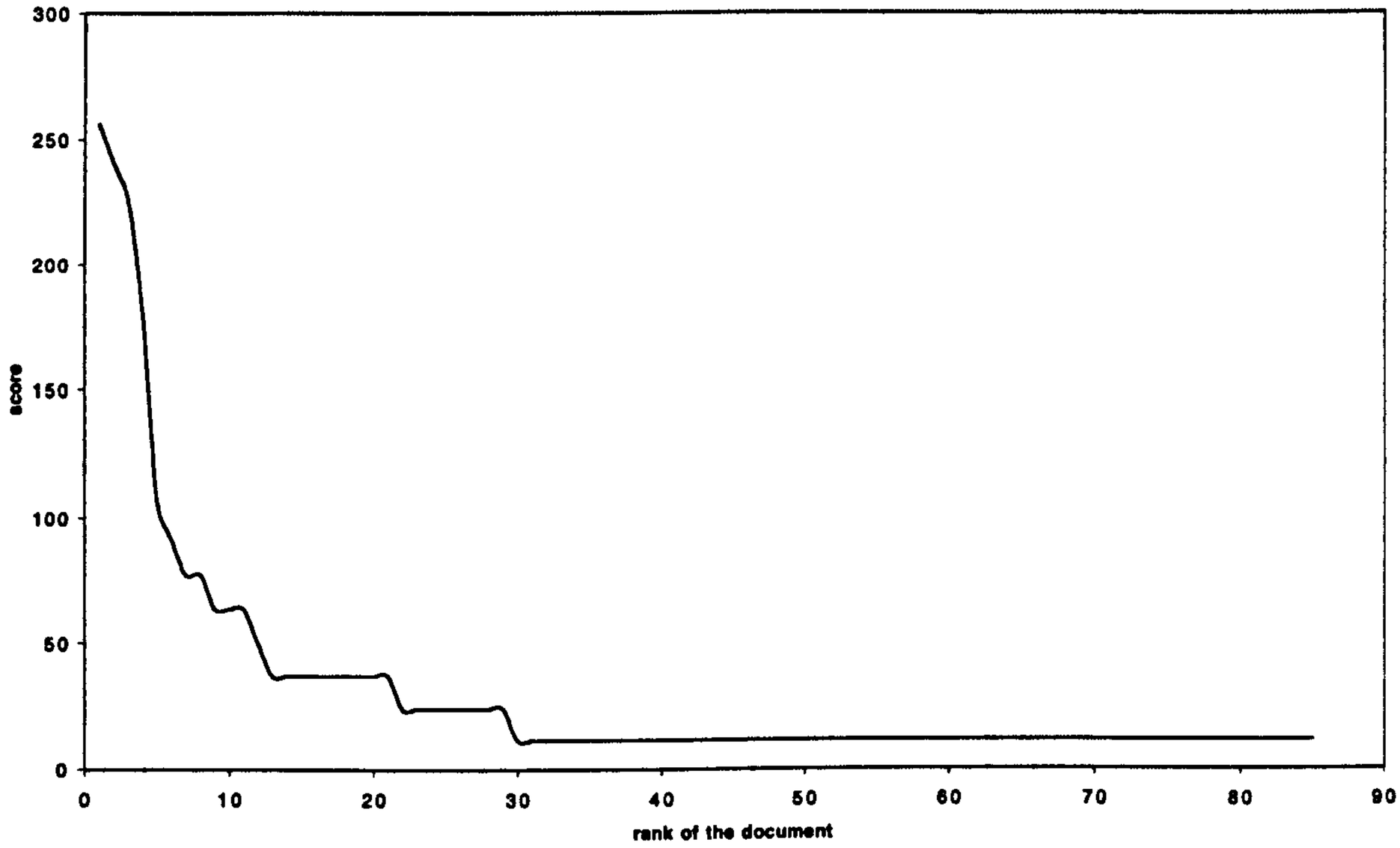


Figure 1.1: Informative content of the term "osteoporosis" with the Poisson approximation (model $P$) of the Bernoulli model over TREC-8 collection.

of a few "Elite" documents. If a very rare term becomes very frequent in a document then its informative content increases very rapidly as in Figure 1.1.

Let us suppose we observe in a document of the elite set of an informative word $t$ a term-frequency $tf$. We saw in the last section that observing the occurrence of the term $t$ within a given document can be abstractly studied as a success in a sequence of Bernoulli trials. Once a document $d$ is given we have seen in the example of last Section 1.4.1 that, under certain hypotheses, we get the binomial formula for the estimate of $Prob_1$. (see Section 2.1.3 and Section 2.2 of Chapter 2 for a formal treatment).

However, we know that if we had observed a different document then also the population would have possibly been different. Changing the population, the value of the term-frequency might well have had a different confidence interval for an accurate estimate. This happens independently from the size of the sample (the length of the document). Although we had sampled the whole document, we do not have further observations to decide whether we would have observed more occurrences of the term in the entire population of the elite set to which the document belongs. Obviously, if $tf$ is large in the document then we take a small risk in deciding that the term tokens appeared

non–randomly. Small risk corresponds to a high conditional probability $p(tf+1|tf)$ of having a further token of the term in the document if its length were longer than the actual one. A very small risk corresponds to a high probability, that is $p(tf+1|tf) \sim 1$.

In conclusion, the larger $tf$ is, the closer the conditional probability $p(tf+1|tf)$ is to certainty. We define $Prob_2(tf) = p(tf+1|tf)$ and $1 - Prob_2$, is the risk, of accepting the informative content as a weight of the term in the document.

When monetary values are involved in decisions we know that in a fair game the risk $1 - Prob_2$ of betting on an event is proportional to the gain.

> Assuming that the informative content $Inf_1(tf)$ of a term $t$ in a document $d$ is the monetary value involved in the decision of taking $t$ as the descriptor of the document, the weight of the term in the document $w$ turns out to be the part of the informative content $Inf_1(tf)$ gained with the decision of taking the term $t$ as a descriptor of the document.

In the next section we show one possible way to compute the apparent aftereffect of sampling.

### 1.5.1 An example of aftereffect model: Laplace's law of succession

Once $Inf_1$ has been computed by using a model of randomness, then the gain is computed with the conditional probability $Prob_2$. We will see that the so–called Laplace's law of succession provides one interpretation of the required conditional probability.

The law of succession (see Equation 3.20 at page 76)

$$Prob_2(tf) = \frac{tf + A}{tf + A + B}$$

can be derived with a Bayesian approach (see Sections 3.3.2 and 3.3.5 of Chapter 3). See Feller's book [37, page 123] for a frequentist derivation of the succession law obtained with an urn model of Type II. Urns model of Type II will be discussed in Section 3.3. The relationship between frequentist and Bayesian approach expressed by De Finetti's theorem, is presented in Sections 3, 3.1 and 3.2 of Chapter 3.

Observing that $1 - Prob_2 = \frac{B}{tf + A + B} \propto \frac{1}{tf + A + B}$ and setting $A = B = 0.5$ we

obtain that the gain computed by Eq. (1.1) is the model $PL$:

$$-\frac{1}{tf+1}log_2 \left( \begin{array}{c} F \\ tf \end{array} \right) p^{tf} q^{F-tf} \quad \text{[model } PL\text{]}$$

In the example of the lift the gain is only $\frac{B}{4+1} = \frac{0.5}{4+1} = \frac{1}{10}$ of the informative content. Considering the example of the query on osteoporosis, the term–weighting function based on the gain instead of the informative content attenuates the decrease rate of the weights as can be observed in Figure 1.2.

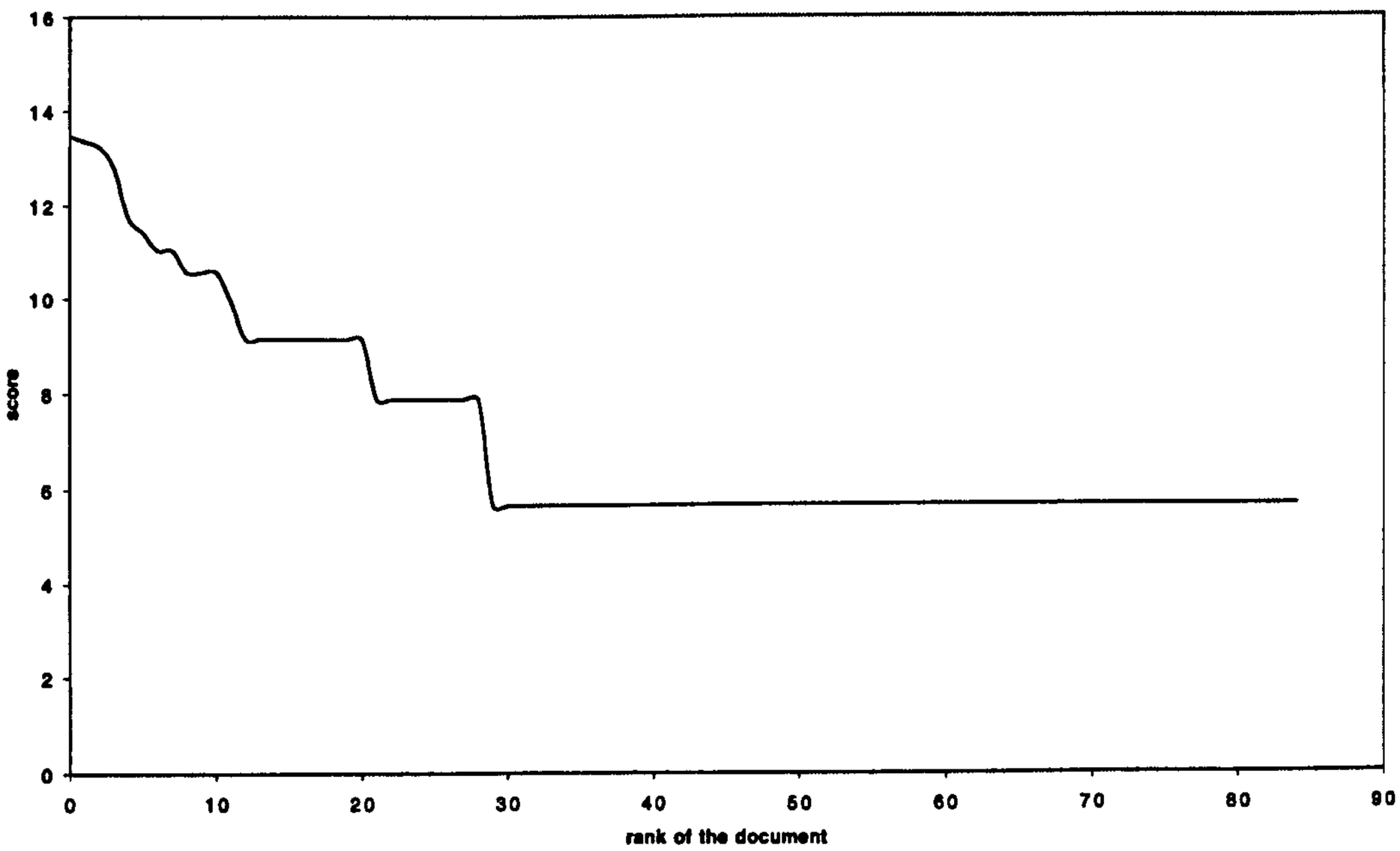


Figure 1.2: Informative content gain (model $PL$) of the term "osteoporosis" over TREC-8 collection.

## 1.6 The third component: term-frequency normalization

So far, we have introduced two probabilities: the probability $Prob_1$ of the term given by a model of randomness and the probability $Prob_2$ measuring the proneness of the term to appear frequently in the elite set, that is the aftereffect in sampling the term in the elite set. However, our intuition says that the magnitude of $tf$ also depends on the document length. On the contrary, all documents are equally likely to receive tokens

according to the urn models of randomness. Urns do not possess a predefined volume and all documents, whether long or short, are treated equally.

Briefly, the normalized term-frequency is the estimate of the expected term frequency when the document is compared with a given length (typically the average document length). We have a bivariate distribution of the number of tokens of a term and the length of documents. Once this distribution is obtained, the normalized term-frequency $tfn$ is used in Formula 1.1 instead of the non-normalized $tf$. We have called the process of substituting the normalized term-frequency for the actual term-frequency *the second normalization* of the informative content.

Despite our intuition about the dependence between frequency and length, Harter couldn't find any general relationship between the term-frequency and the document length. Harter asserts that [52, page 23]

> We assume that there is no relationship between the length of a document $d$ and the number of tokens of the term $t$ in $d$. In particular, we assume that there is no tendency for long documents to contain more tokens of $t$ than short documents. A reasonable alternative hypothesis suggests itself that the probability of a document's receiving a token to be taken to be proportional to its length.

We have run a similar experiment with a larger collection and we have arrived to a different conclusion. A positive correlation exists, and a detailed discussion about this dependence can be found in Chapter 6. We have also tested Harter's suggestion (see Hypothesis **H1** on page 127) comparing it with three other hypotheses using the Bayesian method against several test collections. The second hypothesis, that we have called **H2**, assumes that the relative term-frequency is not constant, as in **H1**, but decreasing with respect to the text length. **H1** approximates **H2** for large lengths (see Formula 6.7 on page 127). The third hypothesis, called **H3**, for term-frequency normalization uses the Dirichlet priors. This last hypothesis is a direct application of the language modelling. The probabilities of terms assigned by any language model can be applied to our framework as length normalization component (see Sections 5.5 and 6.4). In this dissertation we have used the most simple and effective language model, that is that based on Dirichlet's priors.

The fourth and final hypothesis, called **Z** for Zipf, comes from the Pareto-Feller-Zipf's law relating the frequency of a term in the collection, and its rank in decreasing order of magnitude of the frequency (see Sections 2.5 and 6.3).

The Pareto-Feller-Zipf's law establishes a relation between a given term-frequency and the length of the text. The problem is that the rank-frequency relation only holds when the size of the text is very large. We adopted some extra assumptions in order to apply Zipf's law to single documents.

For TREC-8 the comparison among three different models, that is the Bernoulli model only ($P$), the gain function of the Bernoulli model ($PL$) and the gain function of the Bernoulli model under the term-frequency normalization **H2** ($PL2$) is shown in Table 1.1.

Using the query "osteoporosis", the plot of $PL2$ against document rank of term–weighting is less steep than for the plots of $P$ and $PL$ of Figures 1.1 and 1.2, as shown in Figure 1.3.

A comparison with all other models, $BM25$ included, is shown in Table 7.10 on page 158.

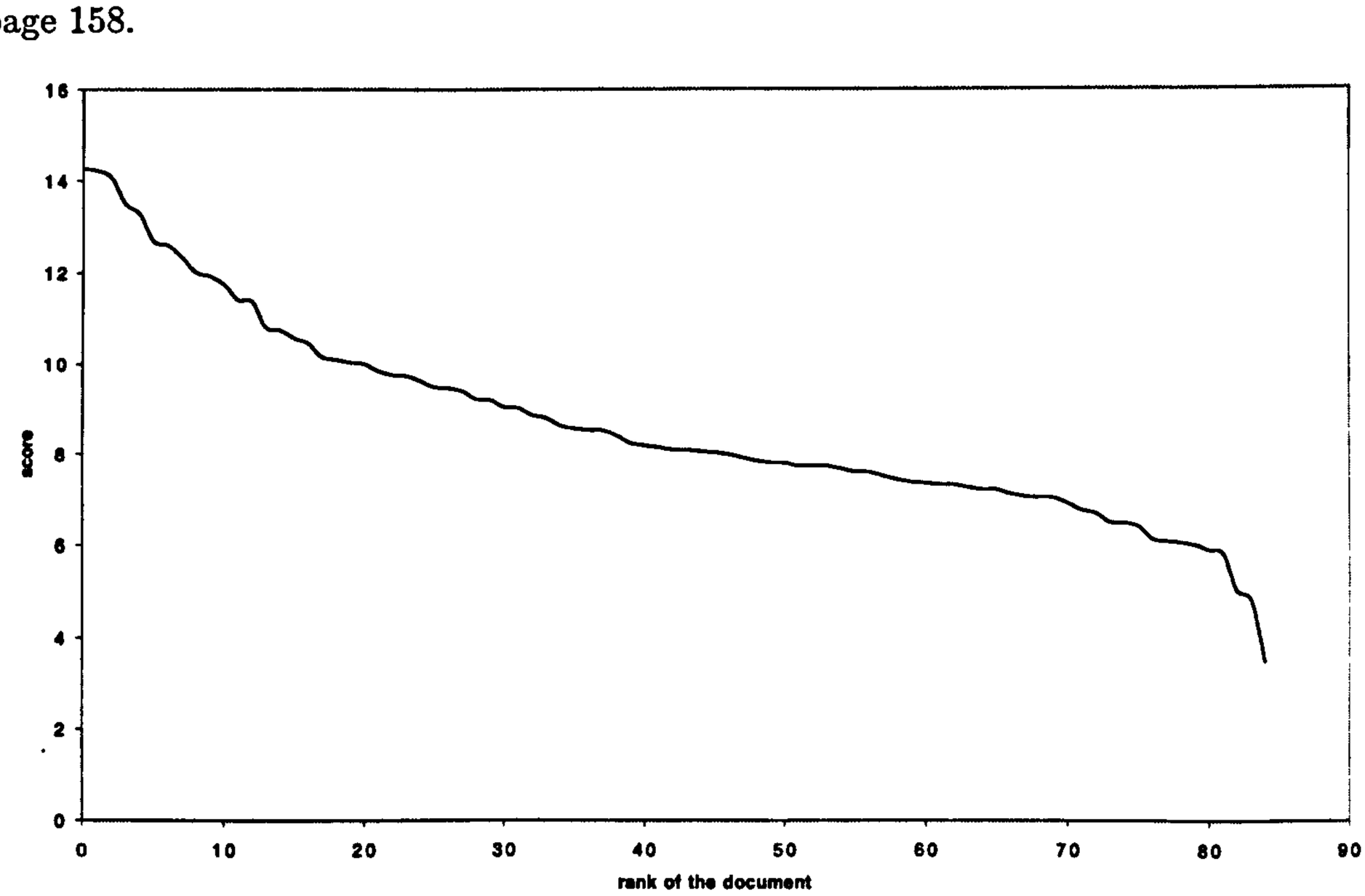


Figure 1.3: Score distribution using the term-frequency normalization component (model $PL2$) over the TREC-8 collection.

| Models | MAP | MAP@10 | Pr@5 | Pr@10 | Pr@20 | Pr@R | RelRet |
|---|---|---|---|---|---|---|---|
| *PL2* | 0.2477 | 0.3587 | 0.4880 | 0.4580 | 0.3970 | 0.2967 | 2866 |
| *PL* | 0.2037 | 0.2619 | 0.4160 | 0.3620 | 0.3260 | 0.2566 | 2645 |
| *P* | 0.0527 | 0.0537 | 0.1040 | 0.1100 | 0.0960 | 0.0863 | 1446 |

Table 1.1: Comparison among the basic model ($P$), the gain ($PL$), and the term-frequency normalization model ($PL2$) with the TREC-8 data. The evaluation measures are defined in Appendix B.1.

## 1.7 The probabilistic framework

Our probabilistic framework builds the weighting formulae in three sequential steps:

1. First, a probability $Prob_1$ is used to define a measure of informative content $Inf_1$ in Equation 1.2. We introduce five *basic models* which measure $Inf_1$. Two basic models are approximated by two formulae each, and thus we provide seven weighting formulae: $I(F)$ (for Inverse term Frequency), $I(n)$ (for Inverse document frequency where $n$ is the document-frequency), $I(n_e)$ (for Inverse expected document-frequency where $n_e$ is the document-frequency which is expected according to a Poisson), two approximations for the binomial distribution, $D$ (for divergence) and $P$ (for Poisson), and two approximations for the Bose-Einstein statistics, $G$ (for geometric) and $B_E$ (for Bose-Einstein).

2. Then, the first normalization computes *the information gain when accepting the term in the observed document as a good document descriptor.* We introduce two (first) normalization formulae: $L$ and $B$. The first formula derives from Laplace's law of succession and takes into account only the statistics of the observed document $d$. The second formula $B$ is obtained by a ratio of two Bernoulli processes and takes into account the elite set $E$ of a term.

3. Finally, we resize the term frequency in the light of the length of the document.

We test four hypotheses:

**H1** - Assuming we can represent the term-frequency within a document as a density function, we can take this to be a uniform distribution, that is the density function of the term-frequency is constant. The $H1$ hypothesis is a variant of the verbosity principle of Robertson [87].

**H2** - The density function of the term-frequency is inversely proportional to the length.

**H3** - Dirichlet's priors produce an expected probability for the relative term-frequency which is given by Equation 6.30 as introduced at page 137 in Section 6.4.

**Z** - Zipf's term-frequency normalization.

## 1.8 The naming of models

Models are represented by a sequence $\alpha\beta\gamma$ where $\alpha$ is one of the notations of the basic models, $\beta$ is one of the two first normalization factors, and $\gamma$ is either 1 , 2 3 or $Z$ according the second normalization **H1**, **H2**, **H3** or **Z**. For example, $PB1$ is the Poisson model $P$ with the normalization factor $B$ of 4.23 with the uniform substitution $tfn$ for $tf$ according to hypothesis **H1**, whilst $B_E L2$ is the Bose-Einstein model $B_E$ in 2.33 with the first normalization factor $L$ of 4.19 with the uniform substitution $tfn$ for $tf$ according to hypothesis **H2**. A summary showing all possible combinations is in Table 1.2.

## 1.9 The component of query expansion

In Chapter 8 we will use the same basic model of randomness used to define $Inf_1$ to expand also the original query. In principle, the query expansion problem is less difficult than the term–weighting problem.

Once a first ranking is produced by the query-document matching function, the top documents in the list are most probable members of the elite set of the query. The actual probability depends on the initial precision of the system. A set of the first few retrieved documents may be taken to be a sample of the elite set of the query. Thus we may pool the content of these document into a unique document–sample to be used in a second ranking. We do not need to normalize the frequencies in each document and the informative content $Inf_1$ can be used directly to extract a term-weight in the elite set. The terms with the highest score can be added to the original query with a query–weight proportional to the $Inf_1$ weight. From experiments we saw that for short queries like those submitted to the WEB search engines made up of at most three or four terms, only 3 documents and up to 10 new terms added to the query are sufficient to

| | BASIC MODELS | |
|---|---|---|
| $P$ | Poisson approximation of the binomial model | Formula 4.7 |
| $D$ | Approximation of the binomial model with the divergence | Formula 4.8 |
| $G$ | Geometric as limiting form of Bose-Einstein | Formula 4.10 |
| $B_E$ | Limiting form of Bose-Einstein | Formula 4.11 |
| $I(n_e)$ | Mixture of Poisson and inverse document-frequency | Formula 4.15 |
| $I(n)$ | Inverse document-frequency | Formula 4.13 |
| $I(F)$ | Approximation of $I(n_e)$ | Formula 4.16 |
| | FIRST NORMALIZATION | |
| $L$ | Laplace's law of succession | Formula 4.18 |
| $B$ | Ratio of two Bernoulli processes | Formula 4.22 |
| | SECOND (LENGTH) NORMALIZATION | |
| **H1** | Uniform distribution of the term-frequency | Formula 6.9 |
| **H2** | The term-frequency density is inversely related to the length | Formula 6.10 |
| **H3** | The term-frequency normalization is provided by Dirichlet's priors | Formula 6.31 |
| **Z** | The term-frequency normalization is provided by a Zipfian relation | Formula 6.29 |

Table 1.2: Models are made up of three components. For example $B_E B2$ uses the limiting form $B_E$ of Bose-Einstein Formula 2.33, normalized by the incremental rate $B$ of the Bernoulli process of Formula 4.22. The within–document term–frequency is normalized under hypothesis $H2$ of Formula 6.10

enhance significantly the performance in the second pass ranking. In the same chapter we will test 6 different basic formulae to perform query expansions, among these the Bose-Einstein statistics and the Bernoulli model. They perform similarly.

The Bernoulli model of a document is this time different from that presented in Section 1.4.1. Unlike the Bernoulli model of a document presented in Section 1.4.1, where the trials were only all tokens of a term occurring in the entire collection, we regard each term in a document as a trial of an *experiment.* Then the whole text becomes a sequence of trials. We then observe a specific term $t$. A successful outcome for the term $t$ is when $t$ occurs in the document. Unlike the Bernoulli model of a document presented in Section 1.4.1, where the *a priori* probability of a success is the probability of retrieving the document, the *a priori* probability of a success here is the relative frequency of the term in the collection.

## 1.10 Experimental work

Since the three components of a model are independent, we have $7 \times 2 \times 4 = 56$ basic models. We have also compared our models with the most effective models currently available in literature, that is the $BM25$ and the language models. If we also consider the query expansion component the number of combination are $56 \times 6 = 336$. Considering that we have now available several big text collections (TREC ones), we could have reported at least $4,000$ runs. However, empirical science is made of trials and errors and we have continuously trialled experiments testing different hypotheses with improving fortune and using ever more effective variants of our models which for the sake of space are not reported. Therefore, with the explosion of all possible combinations we do not claim to have tested our framework extensively and exhaustively, though we have done a massive number of experiments. Nevertheless, we offer a consistent number of tables together and a discussion about our achievement.

## 1.11 Outline of the Thesis

Chapter 2 begins with the construction of the sample spaces for IR. This introductory chapter describes the probabilistic distributions and their limiting forms which are used

in many parts of the dissertation.

Chapter 3 explores the estimation problem underlying the process of sampling. De Finetti's theorem is used to show how to convert the frequentist approach into Bayesian inference and the derived estimation techniques are explored in the context of IR.

Without a doubt, the problem of sampling from different populations is a central one in IR and therefore is treated extensively in this dissertation, for example in Sections 2.1 and 2.1.1, 2.1.2, 2.1.3 of Chapter 2, in many parts of Chapter 3, for example Sections 3, 3.1, 3.2, 3.3, 3.3.1, 3.3.2, 3.3.5 and 3.3.6. In the light of sampling from different populations we revisit a recent IR modelling approach, the language modelling (see Chapter 3 and Section 5.4).

Chapter 4 introduces the notions of informative content and information gain of a term in a document. These two notions are related and constitute the first two components of our models. Examples of models relatively to each component are displayed and these are direct applications of the distributions studied in Chapters 2 and 3.

Chapter 5 revisits the main IR models in the literature in the light of preceding chapters. We show that even language modelling approach can be exploited to assign term-frequency normalization to the models of divergence from randomness. For this reason this chapter precedes the term–frequency normalization fully developed in Chapter 6.

In Chapter 7 we merge the three components and introduce the full models of randomness.

Chapter 8 introduces a novel framework for the query expansion. This framework is based on the models of divergence from randomness and it can be applied to arbitrary models of IR, divergence-based, language modelling and probabilistic models included.

Experiments are diluted along with all chapters, but results are summarized in the final Chapter 9 together with a discussion about open problems and new research directions.

# Chapter 2

# Probability distributions for divergence based models of IR

This chapter introduces the appropriate statistical and mathematical tools to formalize several problems that we have encountered in Information Retrieval. They arise in connection with the following situations.

1. In the first chapter we have introduced the Bernoulli model of IR, in which a document was conceived as an experiment. A document is treated as a *sample* over a population. The outcome of an experiment is either the occurrence (*success*) or not (*failure*) of a specific term within the document. A document collection $D$ can be thus conceived as a collection of samples over different populations.

   What, on these data from multiple sampling, will be the probability that a given word occur in an arbitrary document? What is the probability that a given word frequency is observed in a given document?

2. Taking a different point of view of the same problem: a term distributes $F$ occurrences over a set of documents. What is the probability of having a particular within-document frequency configuration over the entire collection? What is the probability of observing a term-frequency $tf$ within an arbitrary document? The basic spaces formalizing these problems will be deployed to define the basic models of Information Retrieval.

Similarly, let $E$ be a subset of the collection $D$. What is the probability of observ-

ing a term-frequency $F_E$ within the subset $E$? The basic spaces formalizing this problem will be employed to define the models for query expansion in Information Retrieval.

3. We have a collection of documents of different lengths $l_d$. Our intuition says that the term–frequency is related to the length of the document. What is the correlation, if any, between the within-document term-frequency $tf$ and the length of a document? Any possible answer can be used to normalize the term–frequencies with respect to a standard document length within the basic models of Information Retrieval.

4. Let $X$ be the random variable counting the number of words which have a frequency $F$ in the collection. What is the distribution of $X$ varying $F$? How this number is related to the sample space size and the population of the experiment? How does the value $X = F$ affects the observation of a term-frequency $tf$ in a document as described in situation 3?

5. How can we model the fact that the terms which are rare events in the collection are the most informative and they become even more informative when they appear very densely in a few documents?

In order to find answers to all these questions we need to define the probability spaces that we will use (see Section 2.1). Then, we introduce the basic model of randomness for IR, that is the binomial model. We derive some computational forms, the *limiting forms*, necessary for an effective implementation in the Information Retrieval systems. These simplified versions of the binomial law will be central in many application contexts of Information Retrieval (see Sections 2.2, 2.2.1, 2.2.2, 2.2.3 and 2.2.4). The second part of question in 2 can be also answered using the hypergeometric distribution.

Using the terminology of statistics, when balls are drawn from one or more urns, while the binomial distribution assumes replacement in the hypergeometric model the extracted balls are not replaced into the urn. In practice, the hypergeometric distribution does not differ much from the binomial distribution when the sample size is very large. However, Information Retrieval deals with very small probabilities and thus the performance of the two models can be significantly different.

The hypergeometric distribution is useful to support the query expansion process and is the outcome of the compounding of the binomial with the Beta distribution (see Sections 2.3, 8.6 and 2.6.1). This specific compounding is introduced in connection with the language model and a particular term-frequency normalization of question 3.

Another alternative model to the binomial distribution is based on Bose-Einstein statistics. The Bose-Einstein model is introduced in Sections 2.4, 2.4.1 and 2.4.2. In Bose-Einstein statistics the balls of the same colour are indistinguishable, so that many possible arrangements become indistinguishable.

Finally, the fat–tailed distributions are introduced (see Section 2.5). These distributions occur when we try to classify the alternative outcomes of a sample space by their frequencies as stated by the question in 3. The most famous fat-tailed distribution in Information Retrieval is the Zipf distribution 2.5.1.

## 2.1 The probability space in Information Retrieval

Renyi, in his book on probability theory [85], recommends giving great attention to the construction of probability spaces, although in many applications of probability theory the probability space is implicitly assumed. In Information Retrieval we talk mainly of frequencies of terms in a collection of documents and therefore following Renyi's suggestion, our first intention is to provide a clear definition of these entities, that is *the terms* and *the documents*, within a probability space. Information Retrieval deals with discrete probability spaces. The first notion which is defined in a probability space is the *outcome* of an experiment. An outcome lies into a set of mutually exclusive results. We call this set the *basic space* or the *sample space* $\Omega$ of the probability space.

The *algebra of events* $\mathcal{A}$ of the probability space is made up of all subsets $E \subseteq \Omega$. The *probability distribution* $P$ is defined on the algebra of events $\mathcal{A}$. The algebra of events is interpreted as the set of all observable events.

### 2.1.1 The sample space $V$ of the terms

Cooper and Maron [22] developed a theory of indexing which they called Utility-Theoretic Indexing because it was based on utility theory. In their approach index terms are assigned to documents in such a way as to reflect the utility (or value) that the documents are expected to provide to users. Although Cooper and Maron use utilities and not probabilities, they observe that in both Probabilistic and Utility-Theoretic Indexing the fundamental conceptual construct is the event space $\Omega$. They say:

> Utility-Theoretic Indexing is related to (and if random draws are imagined to be made from it, can in fact be interpreted as) an "event space" in the statistician's sense.

We can define several basic spaces $\Omega$ for Information Retrieval. One basic space $\Omega$ of Information Retrieval is the set $V$ of of terms $t$. This set is called the *vocabulary* of the document collection. Since $\Omega = V$ is the set of all mutually exclusive events, $\Omega$ can also be the *certain* event with probability

$$P(V) = \sum_{t \in V} P(t) = 1$$

The probability distribution $P$ assigns thus probabilities to all sets of terms of the vocabulary.

We would try to use this sample space directly if all probabilities $P(t)$ were known in advance. Unfortunately, the basic problem of IR is to find an estimate for $P(t)$. Estimates are computed on the basis of *sampling* and the experimental text collection furnishes the samples needed for the estimation. The main question is how we formally treat two arbitrary but heterogenous pieces of texts, for example the text of this chapter as one and an article from a sport newspaper as the other. Can they be considered as two different samples over two different populations or according to the most reductive assumption as a single sample over the same population? Indeed, the full range of different perspectives can be useful and can be successfully exploited to define query–document matching functions. The next Sections and Chapter 3 develop this idea.

### 2.1.2 Sampling with a document

Another fundamental notion which needs be defined in Information Retrieval is how we conceive a document or a collection of documents in terms of our probability space. The relationship of the document with the experiments is made by the way in which the sample space is chosen.

The term *experiment*, or *trial*, is used here with a technical meaning rather than a general common sense. Thus, we can say that a document is an experiment and we mean that the document is a sequence of outcomes $t \in V$, or more simply a *sample* of a population. Similarly, we may talk of the event of observing a number $X_t = tf$ of occurrences of a given word $t$ in a sequence of experiments. In order to formally discuss this event space, however we should introduce the product of the probability spaces associated with the experiments of the sequence, but this formalism is unnecessarly pedantic. An easier way to introduce our sample space is to associate a point event with each possible configuration of the outcomes. The one-to-one correspondence defines the sample space as

$$\Omega = V^{l_d}$$

where $l_d$ is the number of trials of the experiment (in this case, the length of a document). We can suppose that each outcome does or does not depend on the outcomes of the previous experiments. If the experiments are designed so that an outcome is conditioning the next outcomes then the probability space is not invariant over the sequence of trials and thus the (projection of the) probability distribution on $V$ is different at each trial. To establish the simpler case when the probability space is invariant, in Information Retrieval, the *term independence assumption* is often made. Then, all possible configurations of $\Omega = V^{l_d}$ are considered equiprobable. In the case of equiprobable configurations, together with the assumption that the experiments are independent, we can consider each document a *a Bernoulli process.* The probability spaces of the product are invariant and the probability of a given sequence is the product of the probabilities at each trial. Therefore, if $p = P(t)$ is the *prior probability* that the outcome is $t$ and the

number of experiments is $l_d$ we obtain that the probability of $X_t = tf$ is equal to:

$$P(X_t = tf|p) = \begin{pmatrix} l_d \\ tf \end{pmatrix} p^{tf} q^{l_d - tf} \tag{2.1}$$

which is the sum of the probabilities of all possible configurations having $tf$ outcomes $t$ out of $l_d$. $P(X_t = tf|p)$ is a probability distribution because

$$\sum_{t \in V} P(X_t = tf|p) = (p+q)^{l_d} = 1$$

More generally, the probability distribution is the multinomial (see Equation B.11 in the Appendix)

$$P(\{tf_t\}_{t \in V}|\{p_t\}_{t \in V}) = \begin{pmatrix} l_d \\ tf_1 \ldots tf_V \end{pmatrix} p_1^{tf_1} \cdots p_V^{tf_V} \tag{2.2}$$

which holds in the case that we estimate the probability of an arbitrary configuration satisfying the condition

$$tf_1 + \ldots + tf_V = l_d \tag{2.3}$$

Turning back to the binary case, that is when only a term is observed, if we estimate the probability $p$ of occurrence, when $p$ is unknown, then the *Bayes' theorem* is used:

If $e_1, \ldots, e_n$ are mutually exclusive events of the basic space $\Omega$ and $Y \in \{e_1, \ldots, e_n\}$, then

$$P(Y|X) = \frac{P(Y)P(X|Y)}{\sum_{i=1}^{n} P(e_i)P(X|e_i)}$$

In our case:

$$P(p|X_t = tf) = \frac{P(p)P(X_t = tf|p)}{\sum_{i=1}^{n} P(e_i)P(X_t = tf|e_i)}$$

remembering $e_1, \ldots, e_n$ are all mutually exclusive events of the basic space;

$$P(p|X_t = tf) \propto P(p)P(X_t = tf|p)$$

The component $P(X|Y) = P(X_t = tf|p)$ is the *likelihood* of the *posterior probability* $P(Y|X)$. Details and discussions about the application of Bayes' theorem to the estimation problem in IR are all in Chapter 5. We will see that the likelihood is maximized when $p$ is the expected frequency $\frac{tf}{l_d}$. For this reason this expected frequency is also

called the *maximum likelihood.* We will see also that the *priors* $P(e_i)$ do not have much influence on the value of the posterior probability $P(p|X_t = tf)$ when the sample size, i.e. the number of trials $l_d$, is large. The estimation problem becomes critical when the sample size is small. In this situation, subjectivity arises in the possible choice of the prior form. The subjectivism in assigning priors is somehow paradoxical, because the term *prior* derives from the use in logic to denote the *a priori* statements which are independent of experience and thus they are logically true and objective. As it was observed by Jeffreys this term has been used in so many other senses that the only disambiguation would be to abandon it [66]. However we continue to use the term while understanding that its meaning is slippery.

With the choice of the prior distribution, the maximization of the likelihood may greatly diverge from the maximization of the *a posteriori* probability. De Finetti's Theorem is used to explain that the estimation problem with the Bayesian approach can be seen as an alteration of the "frequentist" underlying model (see Chapter 5). Instead of drawing balls from a single urn with independent trials and a constant probability of success one can use a model with several urns, one for each trial, and with an arbitrary (not uniform) probability distribution of success. The subjectivism of the Bayesian approach then consists in deciding the most suitable initial distribution for the set of outcomes.

### 2.1.3 Multiple sampling: placement of terms in a document collection

We here abandon the hypothesis of having a single sample, as an homogenous piece of text as was assumed in the last Section 2.1.2, and we are going to consider that we have several samples, for example a collection $D$ of documents. The situation of having a collection of $N$ documents is abstractly equivalent to the scheme of placing a certain number $TotFr_D$ of $V$ coloured types of balls in a collection of $N$ cells.

For each term $t \in V$ a possible configuration of ball placement satisfies the equation

$$tf_1 + \ldots + tf_N = F_t \tag{2.4}$$

and the condition

$$F_1 + \ldots + F_V = TotFr_D \tag{2.5}$$

where $F_t$ is the number of balls of the same colour $t$ to be distributed in the $N$ cells.

We have thus implicitly changed the basic space. The outcome of our experiment will be the documents $d$ in which the ball will be placed. Again we will have many possible configurations consistent with the number of coloured balls.

The number of solutions of Equation 2.4 is again the multinomial distribution (see Equation B.11 in the Appendix), under the hypotheses that all configurations are equiprobable and exchangeable:

$$M(F_t, \{tf_i\}_{I=1,\ldots,N}, \{p_i\}_{i=1,\ldots,N}) = \binom{F_t}{tf_1 \ldots tf_N} p_1^{tf_1} \cdots p_N^{tf_N} \tag{2.6}$$

where the priors $p_i$ is the prior probability that the document $d_i$ contains the given term $t$. In absence of further evidence, such as the document length $l$, we may assume the uniform distribution for $p_i$, that is $p = \frac{1}{N}$. Similarly, the number of solutions of Equation 2.5 is the multinomial distribution

$$M(TotFr_D, \{F_i\}_{I=1,\ldots,V}, \{p_i\}_{i=1,\ldots,N}) = \binom{TotFr_D}{F_1 \ldots F_V} p_1^{F_1} \cdots p_V^{F_V} \tag{2.7}$$

where the priors $p$ is the prior probability of having in a collection a term-frequency $F_i$. As we observe in Section 2.5 the priors may follow, for example, the Feller-Pareto law or a uniform distribution with $p_i = \frac{1}{V}$.

It is easy to reduce the multinomial case to the binary case for sake of simple implementation. It can be done by assuming that the only two outcomes are the success or failure of observing the term $t$ in a given document or in the collection. For the multinomial 2.6 the reduction is:

$$B(F_t, tf_i, p) = \binom{F_t}{tf_i} p^{tf_i} q^{F_t - tf_i} \tag{2.8}$$

We can further assume that all configurations which are equal under exchanges are also indistinguishable, that is all sequences which are equal under permutations must be counted as the same event. With that assumption we obtain the so called *Bose-Einstein* statistics (see Section 2.4). The basic sample spaces for Information Retrieval have now been introduced, and we can thus proceed to look at the main probability distributions which will be used to define the models for retrieval and query expansion.

## 2.2 Binomial distribution: limiting forms

We saw that the *term independence assumption* in Information Retrieval regards a document as a Bernoulli process. For example, we saw that in the binary sample case, a document is made up of a set of independent trials with a constant probability $p$ of success, which is the probability that we encounter a specific term in a given position of the text. Indeed, Bernoulli trials are repeated and independent trials with only two possible outcomes, having constant probabilities $p$ (*success*) and $q$ (*failure*). Similarly, in the multiple binary sampling we derive a binomial distribution. Thus, we encounter the binomial distribution for both single and multiple sampling. Let us now treat the binomial case independently from the specific type of sampling. We would like to find useful approximations of the binomial for practical reasons.

Since the outcomes of a Bernoulli process are exclusive events, the probabilities satisfy the condition:

$$p + q = 1$$

The probability of having $k$ successes out of $F$ Bernoulli trials is given by the combinatorial formula 2.8 that is

$$B(F, k, p) = \binom{F}{k} p^k q^{F-k}$$

$B(F, k, p)$ is a probability distribution because

$$1 = (p + q)^n = \sum_{k=0}^{F} B(F, k, p)$$

We use the binomial $B(F, k, p)$ extensively in the implementation of our retrieval models and query expansion models. Therefore a workable approximation of $B(F, k, p)$ is necessary. The next Sections display several limiting forms of the binomial distributions.

### 2.2.1 The Poisson distribution

Assuming that the probability $p$ decreases towards 0 when $F$ increases, but $\lambda = p \cdot F$ is constant, or moderate, an approximation of Equation 2.8 is the Poisson distribution

$$B(F, k, p) \sim \frac{e^{-\lambda} \lambda^k}{k!} \tag{2.9}$$

The value $\lambda$ is both the mean and the variance of the distribution. Further approximation may be obtained through the Stirling formula, which approximates the factorial number as follows [37]:

$$k! = \sqrt{2\pi} \cdot k^{k+0.5} e^{-tf} \tag{2.10}$$

A refinement of Equation 2.10 is

$$k! = \sqrt{2\pi} \cdot k^{k+0.5} e^{-tf} e^{(12 \cdot k+1)^{-1}} \tag{2.11}$$

For example, Feller [37] shows that the approximation error for 100! with equation 2.10 is "only" 0.08%.

$$B(F,k,p) \sim \frac{e^{-\lambda}\lambda^k}{\sqrt{2\pi} \cdot k^{k+0.5} e^{-k} e^{(12 \cdot k+1)^{-1}}} \tag{2.12}$$

The probability in expression 2.12 does not find a direct implementation, but it is used as argument of the logarithmic function, that is:

$$-\log_2 B(F,k,p) \sim k \cdot \log_2 \tfrac{k}{\lambda} + \left(\lambda + \tfrac{1}{12 \cdot k} - k\right) \cdot \log_2 e + 0.5 \cdot \log_2(2\pi \cdot k) \tag{2.13}$$

We will see that $-\log_2 B$ is conceived as the amount of information content related to the term $t$, when $F$ and $tf$ are interpreted as the frequencies of the term in the collection and in the document respectively. The notion of information content is introduced in Section 4.2.

### 2.2.2 The divergence $D$

The fundamental Formula 2.8 of the binomial distribution can be equivalently expressed using the information theoretic divergence $D$ [85], which is defined as:

$$D(\phi,p) = \phi \cdot \log_2 \frac{\phi}{p} + (1-\phi) \cdot \log_2 \frac{(1-\phi)}{(1-p)} \tag{2.14}$$

$D(\phi,p)$ is called the *divergence* of $\phi$ from $p$. With the divergence Formula 2.8 becomes

$$B(F,k,p) = \frac{2^{-F \cdot D(\phi,p)}}{(2\pi \cdot k(1-\phi))^{\frac{1}{2}}} \left(1 + O\left(\frac{1}{F}\right)\right) \tag{2.15}$$

where $k$ is the number of successes, out of $F$ Bernoulli trials, $p$ is the constant probability of success in each trial, $\phi = \frac{k}{F}$, $O\left(\frac{1}{F}\right)$ is the error of the approximation.

To obtain the new approximation of Formula 2.8 of the binomial distribution Renyi applied Stirling's formula. The version of Equation 2.15 with the logarithmic is:

$$-\log_2 B(F,k,p) \sim F \cdot [D(\phi,p) + 0.5\log_2(2\pi \cdot \phi \cdot (1-\phi))] \tag{2.16}$$

The error of the approximation with the logarithm is still $O\left(\frac{1}{F}\right)$, because

$$\log_2\left(1+O\left(\frac{1}{F}\right)\right) = O\left(\frac{1}{F}\right)$$

This equality is obtained using the MacLaurin series (Taylor series expanded about 0) of $\log_2(1+x)$.

### 2.2.3 Kullback-Leibler divergence

Let us assume the approximation of 2.15 of the binomial distribution with the information theoretic divergence of $\phi$ from $p$, where $\phi$ from $p$ are defined as in Section 2.2.2. Without loss of generality we may also assume that $p < \phi$ in Formula 2.16. Indeed, we can show that in the application of the binomial distribution to both term-weighting and query expansion, the definition of $p$ and $\phi$ will satisfy the relation $p < \phi$. Under the assumption of $p < \phi$, the contribution $(1-\phi)\log_2\left(\frac{1-\phi}{1-p}\right)$ in the divergence $D(\phi,p)$ is negative. Also, both $p$ an $\phi$ are very small and thus $\log_2\left(\frac{1-\phi}{1-p}\right)$, which can be easily shown to be approximately $p-\phi$, is also close to 0. Therefore, it is straightforward to derive a further approximation of the Bernoulli process by means of the so-called *asymmetric Kullback-Leibler divergence* $KL(\phi,p)$:

$$KL(\phi,p) = \phi \cdot \log_2 \frac{\phi}{p} \tag{2.17}$$

The approximation of the binomial with the logarithm is

$$-\log_2 B(F,k,p) \sim F \cdot [KL(\phi,p) + 0.5\log_2(2\pi \cdot \phi \cdot (1-\phi))] \tag{2.18}$$

The error of the divergence approximation is the same as in the previous section.

### 2.2.4 The $\mathcal{X}$ divergence

Now, we further approximate Formula ( 2.16). First, let us introduce the function

$$g(x) = x \cdot \ln \frac{x}{p}$$

with $0 \leq x, p \leq 1$.

Thus,

$$g(p) = 0, g'(x) = 1 + \ln \frac{x}{p}, g''(x) = \frac{1}{x}$$

The Taylor series of $g(x)$ is:

$$\begin{aligned} g(x) = & \quad g'(p)(x-p) + \tfrac{g''(p)}{2p}(x-p)^2 + O((x-p)^3) \\ = & \quad (x-p) + \tfrac{1}{2p}(x-p)^2 + O((x-p)^3) \end{aligned}$$

Using the notation $p_1 = p$, $p_2 = q$, $\phi_1 = \phi$ and $\phi_2 = 1 - \phi$, we can easily derive:

$$\sum_{i=1,2} g(\phi_i) = \quad \textstyle\sum_{i=1,2}(\phi_i - p_i) + \sum_{i=1,2} \frac{1}{2p_i}(\phi_i - p_i)^2 + O((\arg_{i=1,2} \max(\phi_i - p_i)^3)$$

Since

$$\sum_{i=1,2} (\phi_i - p_i) = \quad \textstyle\sum_{i=1,2} \phi_i - \sum_{i=1,2} p_i = 0$$

and

$$|\phi_1 - p_1|^3 = |\phi_2 - p_2|^3 = |\phi - p|^3$$

we derive:

$$\sum_{i=1,2} g(\phi_i) = \frac{1}{2} \sum_{i=1,2} \frac{(\phi_i - p_i)^2}{p_i} + O(|\phi_i - p_i|^3) \tag{2.19}$$

Therefore, the divergence $D$ can be approximated as:

$$\begin{aligned} D(\phi, p) = & \quad log_2 e \cdot \sum_{i=1,2} g(\phi_i) \\ = & \quad \frac{log_2 e}{2} \sum_{i=1,2} \frac{(\phi_i - p_i)^2}{p_i} + O(|\phi_i - p_i|^3) \\ = & \quad \frac{log_2 e}{2} \frac{(\phi_i - p_i)^2}{pq} + O(|\phi_i - p_i|^3) \end{aligned} \tag{2.20}$$

The function $\chi^2(\phi, p) = \sum_{i=1,2} \frac{(\phi_i - p_i)^2}{p_i}$ is called *the $\chi^2$ divergence of $\phi$ and $p$*. The approximation of the binomial is easily derived from Equation 2.16 by substituting $D(\phi, p)$ for the right hand side of Equation 2.20:

$$-\log_2 B(F, k, p) \quad \sim F \cdot \left[ \frac{log_2 e}{2} \chi(\phi, p) + 0.5 \log_2 (2\pi \cdot \phi \cdot (1 - \phi)) \right] \tag{2.21}$$

## 2.3 The hypergeometric distribution

The hypergeometric distribution plays an important role in sampling. One application of the hypergeometric distribution is shown in Section 8.6 and concerns the definition of a new model for query expansion. The hypergeometric distribution can be applied to the following problem of sampling. There is a population $D$ of $TotFr_D$ tokens and a number $F$ of tokens are of the same term $t$. A sample $E$ of $D$ is chosen at random. In the query expansion process $E$ will be instead a set of relevant or pseudo–relevant documents, that is a set of documents retrieved after a first retrieval pass. In the chosen sample $E$ we then observe a number $F_E$ of tokens of the same term $t$. The hypergeometric distribution defines the probability $P(F_E|D,E)$ of observing exactly $F_E$ tokens in the sample. The number of ways we can choose the tokens of the term $t$ in the sample is

$$\begin{pmatrix} F \\ F_E \end{pmatrix}$$

The number of possible ways of combining the remaining tokens in the sample is instead:

$$\begin{pmatrix} TotFr_D - F \\ TotFr_E - F_E \end{pmatrix}$$

The total number of possible ways of combining all tokens is:

$$\begin{pmatrix} TotFr_D \\ TotFr_E \end{pmatrix}$$

Then the probability of having the sample $E$ is thus the ratio:

$$P(F_E|D,E) = \frac{\begin{pmatrix} F \\ F_E \end{pmatrix}\begin{pmatrix} TotFr_D - F \\ TotFr_E - F_E \end{pmatrix}}{\begin{pmatrix} TotFr_D \\ TotFr_E \end{pmatrix}} \tag{2.22}$$

The last relation can be rewritten by swapping $TotFr_E$ and $F$:

$$P(F_E|D,E) = \frac{\begin{pmatrix} TotFr_E \\ F_E \end{pmatrix}\begin{pmatrix} TotFr_D - TotFr_E \\ F - F_E \end{pmatrix}}{\begin{pmatrix} TotFr_D \\ F \end{pmatrix}}$$

A limit theorem for the hypergeometric distribution (see [37, page 59]) is:

$$\binom{TotFr_E}{F_E}\left(p_D-\frac{F_E}{TotFr_D}\right)^{F_E}\left(q_D-\frac{TotFr_E-F_E}{TotFr_D}\right)^{TotFr_E-F_E}<$$

$$(2.23)<P(F_E|D,E)<\binom{TotFr_E}{F_E}p_D^{F_E}q_D^{TotFr_E-F_E}\left(1-\frac{TotFr_E}{TotFr_D}\right)^{-TotFr_E}$$

where $p_D$ is the frequency $\frac{F}{TotFr_D}$ of the term in the collection. Therefore, the binomial distribution $B(TotFr_E, F_E, p_D)$ of Formula 2.8 can be taken as a limiting form of the hypergeometric distribution when the population $TotFr_D$ is very large and the size of the sample is very small, that is $\frac{TotFr_E}{TotFr_D}\sim 0$. Indeed the binomial is used directly to obtain weighting scores in the expanded queries in Chapter 8.

## 2.4 Bose-Einstein statistics

In this Section we assume that we randomly place $F$ balls into $N$ recipients. The action of allocating a ball into an urn is the reverse operation of extracting a ball from an urn. Therefore, the allocation process can be easily reversed and transformed into a sequence of ball extractions from the urns. So the model of allocating balls into urns and that of extracting balls from urns possess the same mathematical properties. However, in order to introduce Bose-Einstein statistics it is easier thinking of allocating balls into the urns rather than extracting them. Once the random allocation of balls is completed, this event is completely described by its occupancy numbers: $k_1,\ldots,k_N$ where $k_i$ stands for the frequency of the balls in the $i$-th recipient.

Bose-Einstein statistics assumes that the balls of the same colour are all indistinguishable so that all possible arrangements generating the same ordered sequence of occupancy numbers become equivalent. Hence, with Bose-Einstein statistics we do not have a Bernoulli process of independent trials with a constant probability of success $p$.

The main difference of the Bose-Einstein statistics with the binomial is the assumption that all balls are indistinguishable.

The Bose-Einstein statistics computes the probability of obtaining the frequency $k$ in a recipient by counting the possible combinations consistent with the occupancy numbers

in the rest of the recipients, conditioned to all possible combinations consistent with the occupancy numbers of all recipients. More precisely, a possible configuration of the occupancy problem satisfies the equation [37]

$$k_1 + \ldots + k_N = F \tag{2.24}$$

The number $s_1$ of solutions of Equation 2.24 corresponds to all possible combinations consistent with the occupancy problem. This number $s_1$ is given by the binomial coefficient:

$$s_1 = \left( \begin{array}{c} N+F-1 \\ F \end{array} \right) = \frac{(N+F-1)!}{(N-1)!F!} \tag{2.25}$$

Similarly, let $k$ be the ball frequency in the $i$-th bin. A random allocation of the remaining $F-k$ tokens in the rest of the collection of $N-1$ bins is described by the same Equation 2.24 but with $N-1$ bins instead of $N$ ones:

$$k_1 + \ldots + k_{i-1} + k_{i+1} + \ldots + k_N = F - k \tag{2.26}$$

As before, the number $s_2$ of solutions of Equation 2.26 is:

$$s_2 = \left( \begin{array}{c} N-1+(F-k)-1 \\ F-k \end{array} \right) = \frac{(N+F-k-2)!}{(N-2)!(F-k)!} \tag{2.27}$$

Finally, the probability $P(k)$ that an arbitrary bin contains exactly $k$ occurrences of the ball $t$ is the ratio $\frac{s_2}{s_1}$. That is:

$$P(k) = \frac{\left( \begin{array}{c} N-F-k-2 \\ F-k \end{array} \right)}{\left( \begin{array}{c} N+F-1 \\ F \end{array} \right)} = \frac{(N+F-k-2)!F!(N-1)!}{(F-k)!(N-2)!(N+F-1)!} \tag{2.28}$$

Equation 2.28 is a cumbersome formula and some approximations are needed for the implementation. These approximations will be displayed in Section 2.4.1 and 2.4.2.

### 2.4.1 The geometric distribution approximation

After simplification, Equation 2.28 reduces to

$$P(k) = \frac{(F-k+1)\cdot \ldots \cdot F \cdot (N-1)}{(N+F-k-1)\cdot \ldots \cdot (N+F-1)}$$

Both numerator and denominator of Equation 2.29 are made up of a product of $k+1$ terms. We can divide both numerator and denominator by the product $N^{k+1}$ and distribute it over the terms:

$$P(k) = \frac{\left(\frac{F}{N} - \frac{k-1}{N}\right) \cdot \ldots \cdot \frac{F}{N} \cdot \left(1 - \frac{1}{N}\right)}{\left(1 + \frac{F}{N} - \frac{k+1}{N}\right) \cdot \ldots \cdot \left(1 + \frac{F}{N} - \frac{1}{N}\right)} \tag{2.29}$$

In $IR$ we may in general assume that $N \gg k$. With this assumption

$$\frac{k-i}{N} \sim 0, \text{ and } \frac{i+1}{N} \sim 0$$

for all $i$ with $i = 0, \ldots, k$.

We obtain a limiting form of Equation 2.29

$$\begin{aligned} P(k) \sim & \frac{\frac{F}{N} \cdot \ldots \cdot \frac{F}{N} \cdot 1}{(1 + \frac{F}{N}) \cdot \ldots \cdot (1 + \frac{F}{N})} \\ = & \frac{\left(\frac{F}{N}\right)^k}{\left(1 + \frac{F}{N}\right)^{k+1}} \\ = & \left(\frac{1}{1 + \frac{F}{N}}\right) \cdot \left(\frac{\frac{F}{N}}{1 + \frac{F}{N}}\right)^k \end{aligned} \tag{2.30}$$

Let $\lambda = \frac{F}{N}$ be the mean of the frequency of the ball $t$ in all bins. The probability that a ball occurs $k$ times in a bin is

$$P(k) \sim \left(\frac{1}{1+\lambda}\right) \cdot \left(\frac{\lambda}{1+\lambda}\right)^k \tag{2.31}$$

The right hand side of Equation 2.31 is known as the *geometric distribution* with probability $p = \frac{1}{1+\lambda}$.

### 2.4.2 Second approximation of the Bose-Einstein statistics

The second useful approximation of the Bose-Einstein statistics is generated by the Stirling formula. We will exploit a logarithmic function of Formula 2.28, therefore it is more convenient and easier to rewrite the Bose-Einstein statistics as follows:

$$-\log_2 P(k) = \log_2 \frac{(N+F-k-2)!F!\,(N-1)}{(F-k)!\,(N+F-1)!}$$

$$= -\log_2(N-1) - \log_2(e) + \tag{2.32}$$
$$+f(N+F-1, N+F-k-2) - f(F, F-k)$$

where

$$f(n,m) = (m+0.5)\cdot\log_2\left(\tfrac{n}{m}\right) + (n-m)\cdot\log_2 n$$

## 2.5 Fat–tailed distributions

Fat–tailed (or heavy-tailed) distributions are encountered in many different linguistic, sociological, biological and economic phenomena. Examples of the phenomena fitted by the fat–tailed distributions are classification of terms by frequencies, cities by population, biological genera by numbers of species, scientists by number of published papers, income by size, files by size [29, 55]. Among fat-tailed distributions there are the family of Pareto's distributions [7], which were originally introduced to model income distributions, Champernowne's lognormal distribution [20], the Waring distribution [58], the Yule distribution [102], the generalized inverse Gaussian distribution [101, 100].

As first applications of fat-tailed distributions in linguistics, we should mention the early works of Estoup, Willis and Zipf [35, 125, 134]. They introduced an empirical relationship between the frequency and the rank of the terms which are used in ordinary discourse. Such an empirical law is commonly known as Zipf's law. Zipf's law says that if we rank terms in the decreasing ordering of their relative frequencies and plot the logarithmic values of these relative frequencies $p$ against the logarithmic values of the term position in the ranking, then we approximately get a linear relation:

$$-\log p \sim \alpha\cdot\log(\text{rank})$$

The information content $-\log p$ of the terms is thus highest for terms lower in the ranking[1] and is proportional to the log of the rank. The slope $\alpha$ provides a measure of richness of the vocabulary. If the vocabulary is poor then $\alpha$ goes to 0 and the information content $-\log p$ (or, if we prefer Mandelbrot's terminology [74], the cost for the signal transmission) of all terms becomes a constant.

For example, the TREC 10 collection [56, 9] containing about 1,692,000 documents, has a vocabulary $V$ of about 3,097,466 stemmed terms occurring $TotFr_D = 666,447,515$

[1]The terms which are highly rare are also put in the stop list.

times in the whole collection. Classifying the terms by their frequencies, we can count about 8,826 categories. The last category, that is the class of terms occurring only once in the collection, contains about 1,420,000 terms. After the position $r_0 \sim 2^9 = 512$ (that is with values of $x$ greater than $\log_2 r_0 = 9$ in Figure 2.5) the curve is approximately linear with $\alpha = 1.36$ and it can be approximated by the relation:

$$\log_2 F_t \sim -1.365 \cdot \log_2(\text{rank}) + 29.31$$

of Figure 2.5, where $\log_2 TotFr_D = 29.31$.

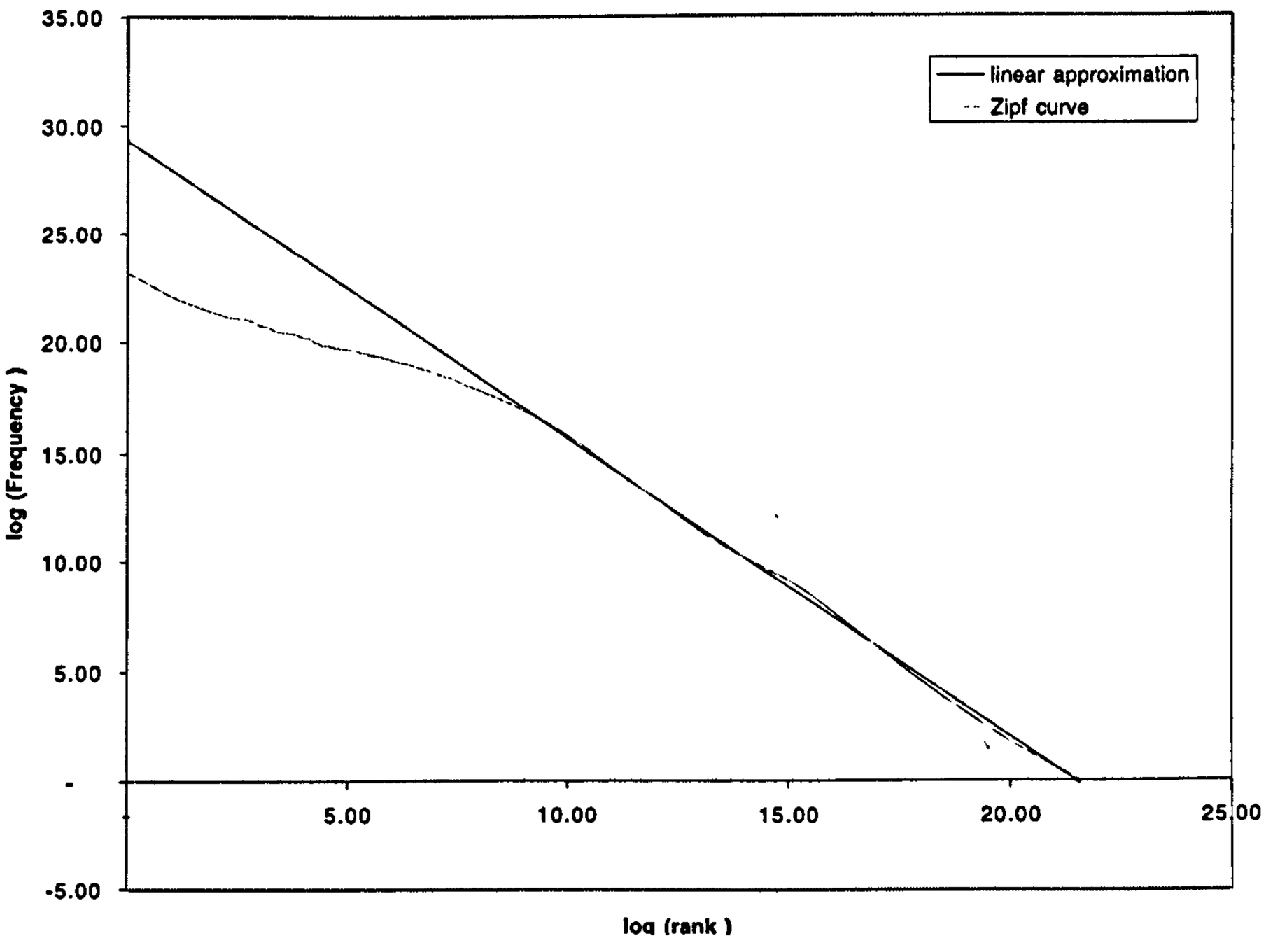


Figure 2.1: Relation between the logarithms of term rank and term-frequency in TREC-10 collection.

Similarly for the TREC-8 collection we get the Relation of Figure 2.5

$$\log_2 F_t \sim -1.399 \cdot \log_2(\text{rank}) + 27.14$$

The Zipf's law can be also regarded as a Pareto distribution, because the Zipf distribution is the discrete version of the Pareto distribution. An alternative frequency term distribution law was created by Champernowne who used the lognormal distribution,

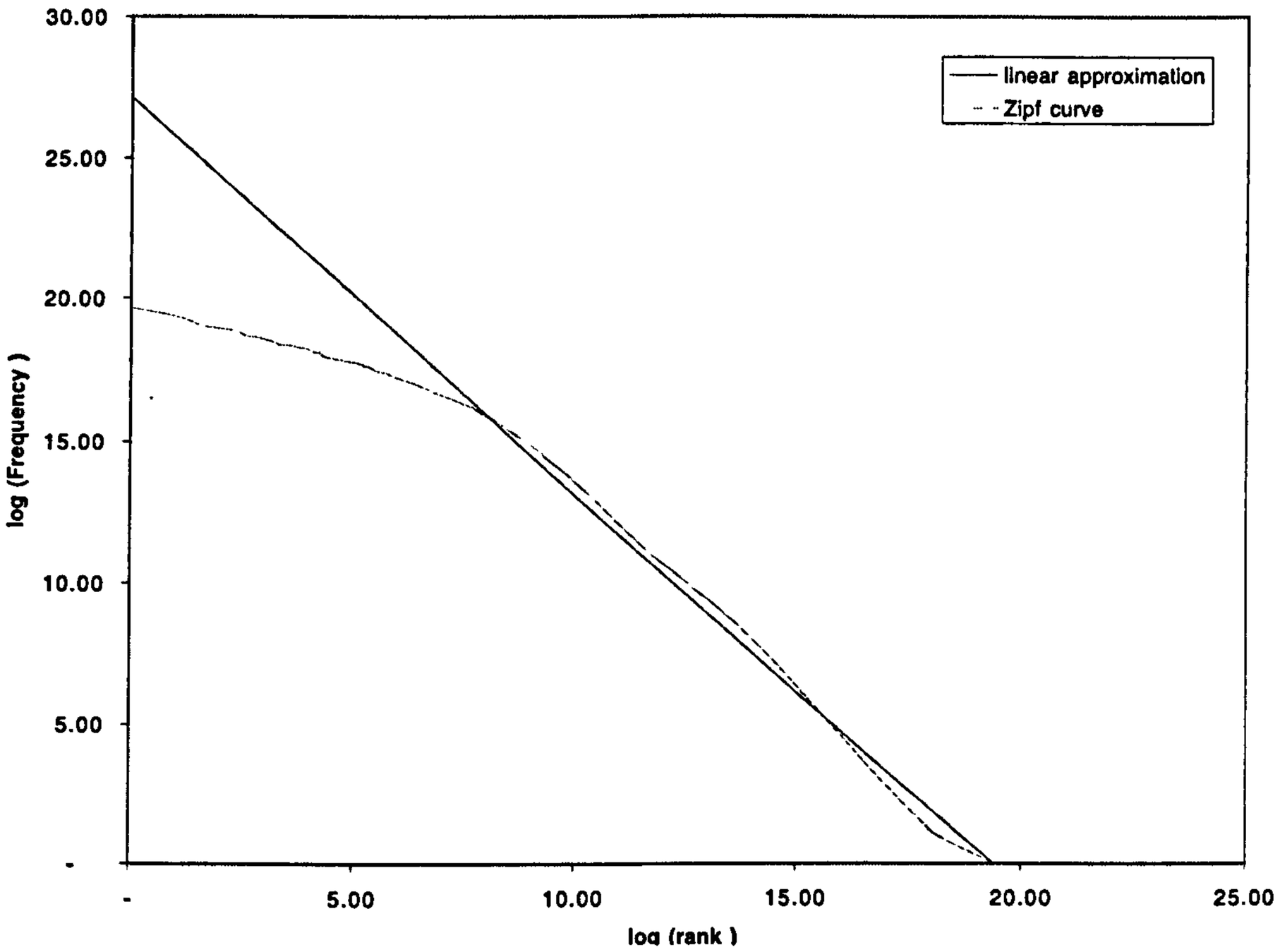


Figure 2.2: Relation between the logarithms of term rank and term-frequency in TREC-8 collection.

that is the normal distribution of the logarithmic values of the random variable. The use of Champernowne's distribution was rejected by Mandelbrot [74] in favour of Simon's proposal [102] which endeavored to use the Yule distribution, a generalization of the Zipf distribution, as a unified model to derive many fat–tailed distributions. He applied the Yule distribution, see Equation B.10 to the term distribution in prose sample.

As we see, there is a plethora of fat-tailed distributions and it would be impossible to give them a unifying definition or provide a unifying methodology able to derive all possible fat-tailed distributions. Indeed, the most general unifying proposal was made by Feller [38] with its family of Feller-Pareto distributions, which are introduced in Section 2.5.1.

Notwithstanding the impossibility of fully characterizing fat-tailed distributions, we may follow Arnold's hint [7]. We have already encountered a number of distributions and we discovered that for example the Bose-Einstein statistics and also the hypergeometric

distribution do not differ very much from the binomial distribution when the sample is very large. Indeed, all distributions which obey the law of large numbers may be reduced to the normal distribution and thus they do not lack the existence of their moments. In other words, mean, variance and higher order moments exist and are all finite. In contrast, fat–tailed distributions having a heavy tail cannot be reduced to the normal distribution and they lack some of their moments (for example the Zipf distribution does not even possess a finite mean). It is the lack of finite moments which makes fat in some sense their tail.

### 2.5.1 Feller-Pareto distributions

We will see that the two most used versions of the Pareto distributions, the classical and standard Pareto distributions, can be derived from Feller-Pareto's family. More precisely, they are examples of the generalized Pareto distributions which all belong to the family of distributions, that is, a generalized Pareto distribution can be seen as a linear combination of a power of the inverse of the Beta distribution [7] which is the general form of the Feller-Pareto's representation for fat–tailed distributions.

**Definition 2** Let $U = Y^{-1} - 1$ be a random variable where $Y$ has the Beta distribution with parameters $\alpha > 0$ and $\beta > 0$. $U$ is said to have a *Feller-Pareto distribution.*

With the Feller-Pareto distributions we are able to introduce the generalized, the classical and the standard Pareto distributions. The standard Pareto distribution is the continuous analogue of the standard discrete Zipf distribution.

By definition the Feller-Pareto probability density function derives from the Beta distribution B.6[see Appendix B.1] by substituting $(1+U)^{-1}$ for $Y$, that is

$$f_Y(y,\alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} y^{\alpha-1}(1-y)^{\beta-1}\Big|_{Y=(1+U)^{-1}} =$$

$$\begin{aligned} f_U(u,\alpha,\beta) &= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\left(\frac{1}{u+1}\right)^{\alpha-1}\left(\frac{u}{u+1}\right)^{\beta-1}\frac{-dY}{dU} \\ &= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\left(\frac{1}{u+1}\right)^{\alpha-1}\left(\frac{u}{u+1}\right)^{\beta-1}(u+1)^{-2} \\ &= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}(u+1)^{-(\alpha+\beta)}\,u^{\beta-1} \end{aligned} \tag{2.33}$$

with $u > 0$

Equation 2.33 follows from $\frac{dY}{dU} = -(U+1)^{-2}$, $U$ being a decreasing function with respect to $Y$.

### The generalized Pareto distribution

The *generalized Pareto distribution* of a random variable $W$ is obtained from a linear combination of a power of $U$, where $U$ is the Feller-Pareto distribution with probability density function of Equation 2.33, as follows:

$$W(\mu, \sigma, \gamma, \alpha, \beta) = \mu + \sigma U^{\gamma} \tag{2.34}$$

If $W = \mu + \sigma U^{\gamma}$ then

$$U = \left(\frac{W-\mu}{\sigma}\right)^{\frac{1}{\gamma}}$$

The probability distribution $P(w)$ of the random variable $W$ is thus given by the probability of the event $W > w$ that is:

$$P(W > w) = P(U > \left(\frac{W-\mu}{\sigma}\right)^{\frac{1}{\gamma}}) \tag{2.35}$$
$$\text{with } w > \mu$$

Deriving $U = \left(\frac{W-\mu}{\sigma}\right)^{\frac{1}{\gamma}}$ with respect to $W$ we get:

$$\frac{dU}{dW} = \frac{1}{\gamma\sigma}\left(\frac{W-\mu}{\sigma}\right)^{\frac{1}{\gamma}-1}$$

From this derivative and Equation 2.33, we easily obtain the probability density function of the generalized Pareto distribution:

$$\begin{aligned} f_W(w) = & \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\left(\left(\frac{w-\mu}{\sigma}\right)^{\frac{1}{\gamma}}+1\right)^{-(\alpha+\beta)} \cdot \\ & \cdot\left(\frac{w-\mu}{\sigma}\right)^{\frac{\beta-1}{\gamma}}\frac{1}{\gamma\sigma}\left(\frac{w-\mu}{\sigma}\right)^{\frac{1}{\gamma}-1} \\ = & \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)\gamma\sigma}\left(\left(\frac{w-\mu}{\sigma}\right)^{\frac{1}{\gamma}}+1\right)^{-(\alpha+\beta)}\left(\frac{w-\mu}{\sigma}\right)^{\frac{\beta}{\gamma}-1} \end{aligned} \tag{2.36}$$
$$\text{with } w > \mu$$

### The classical Pareto distribution

The *classical Pareto distribution* is obtained from the generalized Pareto distribution $W(\sigma, \sigma, 1, \alpha, 1)$, that is with:

$$\mu = \sigma \text{ and } \gamma = \beta = 1$$

The classical Pareto distribution has the probability density function

$$\begin{aligned} f_W(w) = & \frac{\Gamma(\alpha+1)}{\Gamma(\alpha)\sigma}\left(\frac{w}{\sigma}\right)^{-(\alpha+1)} & \text{with } w > \sigma \\ (2.37) \qquad = & \frac{\alpha}{\sigma}\left(\frac{w}{\sigma}\right)^{-(\alpha+1)} & \text{with } w > \sigma \end{aligned}$$

The classical Pareto distribution is then

$$P(W > x) = \int_{\sigma}^{x} \frac{\alpha}{\sigma}\left(\frac{w}{\sigma}\right)^{-(\alpha+1)} dw = 1 - \left(\frac{x}{\sigma}\right)^{-\alpha}$$

### The discrete analogue of the classical Pareto distribution: Zipf's law

Suppose that the random variable $X$ takes the discrete values $0, 1, 2, \ldots$ or $1, 2, \ldots$ and that $X$ has a fat-tailed distribution. Consider a sample made up of $n$ observations. Suppose that there are $V$ possible outcomes and that their frequency is such that $p(r+1) = P(X = r+1) \leq p(r) = P(X = r)$. According to [7] Zipf distributions are discretized Pareto distributions. The discrete analogous of Formula 2.36 is defined as:

$$(2.38) \qquad P(X \geq r) = \left(1 + \left(\frac{r - r_0}{\sigma}\right)^{\frac{1}{\gamma}}\right)^{-\alpha} \quad r \geq r_0$$

For $\gamma = \alpha = \sigma = 1$ and $r_0 = 0$ we obtain the *standard Zipf* distribution:

$$(2.39) \qquad P(X \geq r) = (1+r)^{-1} \quad r \geq 0$$

Note that

$$(2.40) \qquad P(X = r) = P(X \geq r) - P(X \geq r+1) = r^{-1}(1+r)^{-1}$$

For $\gamma \neq 1$ we obtain the Zipf distributions which are the discrete analogues of the classical Pareto distributions:

$$(2.41) \qquad P(X \geq r) = \left(1 + \frac{r - r_0}{\sigma}\right)^{-\alpha} \quad r \geq r_0$$

### The standard Pareto distribution

The *standard Pareto distribution* of a random variable $Z$ is obtained from the generalized Pareto distribution $W(0,1,1,1,1)$. Its probability density function is:

$$f_Z(z) = (1+z)^{-2} \quad \text{with } z > 0 \tag{2.42}$$

## 2.6 Mixing and compounding distributions

Many distributions can be constructed from different distributions by a process defined as *compounding*. We are using here the terminology of [67]. Let $X$ have the the probability distribution $P(X|Y)$ and $Y$ is another random variable which instead has probability distribution $Prob(Y)$. Then the *compounding of $P$ with $Prob$* is

$$C(X) = \int_{-\infty}^{+\infty} P(X|c \cdot Y) dProb(Y) \tag{2.43}$$

where $c$ is a constant.

Let $P_i$ be a set of probability distributions and $f_i$ a set of values such that

$$\sum_i f_i = 1$$

then the *the mixture of the probability distributions $P_i$* is

$$\sum_i f_i P_i \tag{2.44}$$

### 2.6.1 Compounding the binomial with the Beta distribution

In Section 3.3.5 of Chapter 5 we compound the binomial with the Beta distribution and therefore we show here, as an example, how to compound the binomial distribution of Equation 2.8 with the Beta distribution of Equation B.6 assuming that the prior $p$ is the parameter $Y$ in the compounding Relation 2.43.

$$P(X = k|p) = \binom{F}{k} p^k q^{F-k}$$

$$Prob(p) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\alpha-1} q^{\beta-1} \tag{2.45}$$

Note that if $Y = p$ then $dProb(p) = Prob(p)dp$. Hence, the compounding is

$$C(X = k) = \int_0^1 \left( \begin{array}{c} F \\ k \end{array} \right) p^k q^{F-k} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\alpha-1} q^{\beta-1} dp$$

which reduces to

$$C(X = k) = \left( \begin{array}{c} F \\ k \end{array} \right) \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \int_0^1 p^{k+\alpha-1} q^{F-k+\beta-1} dp$$

That is

$$C(X = k) = \left( \begin{array}{c} F \\ k \end{array} \right) \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \frac{\Gamma(k+\alpha)\Gamma(F-k+\beta)}{\Gamma(F+\alpha+\beta)} \tag{2.46}$$

We already know that $C(X = k)$ is a probability distribution for De Finetti theorem (see Theorem 5). However, it is of some utility to prove it directly. In order to show that 2.46 is a probability distribution we must verify that $\sum_{k=0}^{F} C(X = k)$ is equal to 1.

Let $x$ be any real number and $r$ be a positive integer. Let $\left( \begin{array}{c} x \\ r \end{array} \right)$ denote the real number

$$\frac{x \cdot (x-1) \cdots (x-r+1)}{r!}$$

If $a > 0$ then [37, Problem 20 of Chapter II]

$$\left( \begin{array}{c} -a \\ r \end{array} \right) = (-1)^r \frac{\Gamma(a+r)}{r!\Gamma(a)}$$

where $\Gamma$ is the Gamma function of Equation B.1 in Appendix B.1. Relation 2.46 can be rewritten as a generalized form of the hypergeometric distribution, see Relation 2.22, with binomials containing negative real numbers:

$$C(X = k) = \frac{(-1)^k \left( \begin{array}{c} -\alpha \\ k \end{array} \right) (-1)^{F-k} \left( \begin{array}{c} -\beta \\ F-k \end{array} \right)}{(-1)^F \left( \begin{array}{c} -(\alpha+\beta) \\ F \end{array} \right)} \tag{2.47}$$

From [37, Problem 9 of Chapter II]

$$\sum_{k=0}^{F} \left( \begin{array}{c} -\alpha \\ k \end{array} \right) \left( \begin{array}{c} -\beta \\ F-k \end{array} \right) = \left( \begin{array}{c} -(\alpha+\beta) \\ F \end{array} \right)$$

which proves that $\sum_{k=0}^{F} C(X=k) = 1$

## 2.7 Summary and Conclusions

We have presented different probability spaces of IR: the Bernoulli model and its limiting forms, the hypergeometric distribution, Bose-Einstein statistics and its limiting forms, the compound of the binomial distribution with the beta distribution, and the fat-tailed distributions. The components of the models of divergence from randomness are based on these distributions.

# Chapter 3

# The estimation problem in IR

The parameter estimation in Information Retrieval was first stated by Van Rijsbergen relatively to the probabilistic term-weighting model [118]. The interdependencies between parameter estimation and the properties of probabilistic models are also studied in [42]. Van Rijsbergen extended the Robertson and Sparck Jones weighting formula to a *linear discriminant function* and to a *non-linear discriminant function* when the terms are not assumed independent. These discriminant functions involve several parameters which need to be estimated from a small sample of relevant documents as well as from the whole collection of documents. Van Rijsbergen anticipated and recommended the use of the relevant statistical theory needed to address the estimation problem for Information Retrieval. We fully discuss and apply that proposal. We find that the content of Good's book [44] gives an excellent combination of historical discussion and technical details for a fruitful application of the estimation rules to the problem in Information Retrieval. We have developed an understanding that Information Retrieval can be abstractly redefined by suitable models drawing balls from or distributing balls into urns. De Finetti's Theorem and Bayes' rule are the central relationships in this abstract reading. Their applications allow us to introduce some "subjectivism" or "arbitrariness" in the parameter estimation problem. Fortunately, Information Retrieval is an empirical science and evaluation of the newly built term–weighting models can tell us how well the different methods perform in terms of precision measures.

An alternative approach to the parameter estimation is followed by Steinhaus [112]. Steinhaus minimises the loss function $(p' - p)^2$ where $p'$ is the estimate and $p$ is the

unknown probability of the event. This approach will be considered in Section 3.3.4

## 3.1 Sampling from different populations

We can imagine sampling as the experiment of drawing balls of several colours from different urns. If the ball selection uses a single urn then we would have a single population. We call the experiment of drawing from a single urn as *a sampling of Type I* (we use the terminology of Good [44]). We may use a "super"-population of urns, each of them having a distribution of Type I, and we may then choose one urn and perform sampling from this urn, obtaining a second type (*Type II*) distribution. Thus we can define infinitely many types (*Type III*, *Type IV* etc.) of sampling by iterating this construction indefinitely. The probability estimation becomes more and more complex as long as the type complexity of the sampling increases.

## 3.2 Type I sampling

Frequentist approach deals mainly with binary sampling of Type I. We have a single urn and after randomly selecting a ball we observe its colour and we replace it into the urn. If the ball is of the given colour then we have a *success*, otherwise a *failure*. For each ball colour $t$ we have an expected frequency value, the mean frequency $\lambda_t$, which is the number of successes $r$ divided by the total number of successes and failures in the sampling, e.g. the size $n = r + s$ of the sample. We assume that every sequence having $r$ successes and $s$ failures is equiprobable. We also assume that the sequence having $r$ successes and $s$ failures is the outcome of repetitive drawings. Each trial is assumed to be independent from previous ones. Such sequences are called *permutable* or *exchangeable*.

If the prior $p$ is known, then the probability of having a sequence with $r$ successes and $s$ failures is

$$Prob(r, s|p, d) = \quad p^r(1-p)^s \tag{3.1}$$

The probability of having a permutable sequence with $r$ successes and $s$ failures can be

obtain by multiplying Equation 3.1 by the binomial coefficient:

$$\binom{r+s}{r} p^r (1-p)^s \tag{3.2}$$

The symbol $d$ in 3.1 is to recall that the values $r$, $s$ and $p$ depend on the "document model" $d$. This notation may be used to denote other possible random variables which are observable in the given document regarded as empirical data. We may also extend Relation 3.1 considering the statistics relative to the entire collection $C$, and in this case we may use the notation $Prob(r, s|p, \ldots, d, C)$.

If the *a priori* probability of drawing $t$ is known and is equal to $p$, and the size of the sample tends towards very large, then the mean $\frac{r}{r+s}$ converges to $p$. This comes from the theorem of large numbers:

**Theorem 3 (Theorem of Large Numbers)** *Assume that the event A has probability p. Let us carry out a sequence of identical independent experiments. Then in the first n experiments the frequency r of successes is stochastically convergent to p, that is for all* $\varepsilon > 0$

$$\lim_{n\to\infty} Prob\left(|r - p \cdot n| > n \cdot \varepsilon\right) = 0 \tag{3.3}$$

However, if the *a priori* probability $p$ is unknown or the size of the sample is small, then the estimation of the probability $p$ cannot be set simply to the mean value $\lambda_t$ of the sample. In these cases, priors are parameters and Bayes' theorem can be used to solve the problem of a Type I probability estimation. According to Bayes' theorem, the probability of having $p$ as prior is:

$$\begin{aligned} Prob(p|r,s,d) &= \frac{Prob(r,s|p,d) \cdot Prob(p|d)}{Prob(r,s|d)} \\ &= \frac{p^r(1-p)^s \cdot Prob(p|d)}{Prob(r,s|d)} \end{aligned} \tag{3.4}$$

where the denominator is

$$\begin{aligned} Prob(r,s|d) &= \int_0^1 Prob(r,s|p,d) \cdot Prob(p|d)dp \\ &= \int_0^1 p^r(1-p)^s \cdot Prob(p|d)dp \\ &= \int_0^1 p^r(1-p)^s \cdot dProb(p|d) \end{aligned} \tag{3.5}$$

and the priors satisfy the condition:

$$\int_0^1 Prob(p|d)dp = 1 \tag{3.6}$$

The probability of Equation 3.5 can be regarded as the compounding probability of the binomial with the Beta distribution that has been studied in Section 2.6.1 of Chapter 2. The estimation of the unknown prior $p$ as given in Relation 3.4 depends on the priors of Relation 3.6. Chernoff's bounds tell us that, when the sample is very large, the priors are less and less important in providing an estimate for $p$.

**Theorem 4 (Chernoff Bounds)** *Assume the same situation as Theorem 3.3. Then, for all* $0 \leq \varepsilon \leq 1$

$$Prob\left(|r - p \cdot n| > \varepsilon pn\right) \leq 2e^{-\varepsilon^2 pn/3} \tag{3.7}$$

In this case this estimate is closer and closer to the maximum likelihood $\lambda_t$. Therefore, the Bayesian approach would not give a very different estimate from that if we had assumed a frequentist approach or, equivalently, a Type I distribution.

## 3.3 Type II sampling: De Finetti's Theorem

If we observe carefully the mathematical form of relation 3.5, then we may assert that Bayes' Theorem has made it possible to transform a Type I sampling into a Type II distribution. Indeed, we may read the application of Bayes' theorem as if we were sampling balls from several urns where the probability of having a frequency $p$ has the initial distribution function $Prob(p|d)$. More precisely, this is the content of De Finetti's theorem:

**Theorem 5 (De Finetti)** *A sample generated by a permutable random binary sequence can always be regarded as a binomial sampling in which the a priori probability p (Type I probability) has a Type II prior distribution function* $Prob(p|d)$.

In mathematical formalism De Finetti's Theorem can be regarded as the compounding of the binomial distribution with the distribution provided by the priors as introduced in Section 2.6 of Chapter 2. As already observed, if the sample is very large, then the priors $Prob(p|d)$ become less important and the process reduces to a Type I probability

distribution. In [57] there is a simple proof of this theorem which provides also an extension of the theorem to the case that the prior distribution is defined on a finite set.

### 3.3.1 Estimation of the probability with the posterior probability

Bayes' theorem provides a posterior probability of $p$ over the empirical data compounding two probability distributions of $p$, the likelihood and the prior distribution. We still have to solve the problems of choosing the best estimate for $p$. One way is to derive the most probable value of $p$, that is the value for $p$ that maximizes the posterior probability. A second way is to consider the expected value of $p$, that is the mean of $p$ over the posterior distribution. In other words, we can choose either the value for $p$ that satisfies the equation

$$\frac{dProb(p|r,s,d)}{dp} = 0 \tag{3.8}$$

or the expectation $E(p)$ of the random variable $p$ with respect to the posterior probability distribution

$$E(p) = \int_0^1 p \cdot Prob(p|r,s,d)dp = \int_0^1 p \cdot Prob(r,s|p,d)Prob(p|d)dp$$

These two approaches produce different results as it can be easily verified when the prior distribution is uniform. In such a case the solution of Equation 3.8 also maximises the likelihood $Prob(r,s|p,d)$ and therefore coincides with the maximum likelihood $\frac{r}{r+s}$ (see Section 3.3.3), whilst the expectation $E(p)$ of $p$ establishes the so–called Laplace's Law of Succession (see Section 3.3.2).

### 3.3.2 Bayes-Laplace estimation

Let us suppose that in the Bayes' relation the Type II distribution function of $p$ is uniform, that is

$$Prob(p|d) = c$$

where $c$ is a constant.

Equation B.7 (see Appendix) and Equation 3.4 give the Laplace's Law of Succession. To derive it we observe that:

$$E(p) = \frac{1}{Prob(r,s|d)} \int_0^1 p \cdot Prob(r,s|p,d) \cdot Prob(p|d)dp$$

$$
\begin{aligned}
&= \frac{c}{Prob(r,s|d)} \int_0^1 p^{r+1}(1-p)^s \cdot dp \\
&= \frac{c}{Prob(r,s|d)} \frac{\Gamma(r+2)\Gamma(s+1)}{\Gamma(r+s+3)} \\
&= \frac{c}{Prob(r,s|d)} \frac{r+1!s!}{r+s+2!}
\end{aligned} \tag{3.9}
$$

Similarly, exploiting Equation B.7 again:

$$
\begin{aligned}
Prob(r,s|d) = \int_0^1 Prob(r,s|p,d) \cdot Prob(p|d)dp = \\
= c\int_0^1 p^r(1-p)^s \cdot dp = \\
= c\frac{r!s!}{r+s+1!}
\end{aligned} \tag{3.10}
$$

Both Equations 3.9 and 3.10 imply Laplace's Law of succession:

$$E(p) = \frac{r+1}{r+s+2} \tag{3.11}$$

### 3.3.3 Maximum likelihood estimation

An alternative method to Bayes-Laplace is the maximum likelihood method. Relation 3.4 can be regarded as

$$\text{Posterior Probability} \propto \text{Prior Probability} \cdot \text{Likelihood} \tag{3.12}$$

If the prior probability function $Prob(p|d)$ is assumed uniform, then:

$$\text{Posterior Probability} \propto \text{Likelihood} \tag{3.13}$$

To obtain the posterior probability of $p$, the likelihood function $Prob(r,s|p,d)$ is maximized. The maximum likelihood estimate $\frac{r}{r+s}$ is given by the solution of the first derivative of the likelihood:

$$\frac{d}{dp}Prob(r,s|p,d) = p^{r-1}(1-p)^{s-1}(r-(r+s)p) \tag{3.14}$$

$\frac{r}{r+s}$ is the value for which the likelihood is maximised, that is when $\frac{d}{dp}Prob(r,s|p,d) = 0$.

An algorithm for computing maximum likelihood estimates, the so-called *EM* algorithm, from incomplete data is presented by Dempster, Laird and Rubin in 1977 [32, 79]. In situations where data are complete the maximum likelihood estimate is easy to compute. When data are incomplete some unknown parameters can be introduced to make

the data complete. The *EM* algorithm is an iterative computation of the maximum likelihood made into two steps. In the E-step the expectation of the unknown parameters $\Theta$ is computed. The M-step finds the estimates of the parameters which maximize this expectation. An application of the *EM* algorithm to the computation of the probability mixture parameter of the language model can be found in [59].

### 3.3.4 Estimation with the loss function

We apply the methodology used by Steinhaus [112] to the problem of estimation. Suppose we have a binary sample where the probability of success is $p$. As stated in De Finetti's theorem, see Section 3.3, suppose we also have a prior probability distribution for $p$. The loss function $I(x,r)$ is $(p-x)^2$ where $x$ is our estimate and $p$ is the unknown probability $p$. Thus starting from Bayes' rule 3.5 with $r+s$ trials the loss function is:

$$\begin{aligned} I(x,r) = & \int_0^1 Prob(r,s|p,d) \cdot Prob(p|d)(p-x)^2 dp \\ & = \int_0^1 p^r(1-p)^s(p-x)^2 \cdot dProb(p|d) \end{aligned} \tag{3.15}$$

It is easy to observe that

$$\begin{aligned} I(x,r) = & \quad E(p^2 - 2xp + x^2) \\ = & \quad E(p^2) - 2xE(p) + x^2 \end{aligned}$$

Therefore $I(x,r)$ is minimised when

$$\frac{dI(x,r)}{dx} = 0 \Leftrightarrow x = E(p)$$

Thus, minimising the loss function turns to be equivalent to the decision of choosing as an estimate the expected value $E(p)$ of $p$ as defined in Section 3.3.1.

### 3.3.5 Small binary samples

If the sample is not significantly large, then results obtained from Theorem 5 depend heavily on the priors and thus its use needs an estimation of the priors $Prob(p|d)$. In such a case the maximum likelihood $\lambda_t$ for the Type I distribution in Relation 3.1 is not very relevant. Although it may be considered for its limiting property provided

by the theorem of large numbers as less subjective estimate of the priors than other possible estimates. More generally the Type II distribution function can be given by a law similar to 3.1, that is by means of the Beta distribution B.7 of Appendix B.1 instead of the uniform distribution, with parameters $A$ and $B$ instead of $r$ and $s$.

$$Prob(p|d) = \frac{p^{A-1}(1-p)^{B-1}}{\int_0^1 p^{A-1}(1-p)^{B-1}dp} \tag{3.16}$$

Since

$$\int_0^1 p^{A-1}(1-p)^{B-1}dp = \frac{\Gamma(A)\Gamma(B)}{\Gamma(A+B)} \tag{3.17}$$

where $\Gamma$ is the Gamma function (see equation B.1 in the Appendix), then

$$Prob(p|d) = \frac{\Gamma(A+B)}{\Gamma(A)\Gamma(B)}p^{A-1}(1-p)^{B-1} \cdot dp \tag{3.18}$$

The distribution 3.18 is the Beta distribution with parameters $A$ and $B$. The *a posteriori* probability distribution which turns out from this Type II distribution, after conditionalizing on $Prob(t|d)$, takes the same form of Relation 3.18, that is:

$$Prob(p|t,d) = \frac{\Gamma(A+B+r+s)}{\Gamma(A+r)\Gamma(B+r)}p^{r+A-1}(1-p)^{r+B-1}dp \tag{3.19}$$

With a similar derivation to Equation 3.11 and similarly to the derivation used in compounding the binomial with the Beta distribution in the example of Section 2.6 of Chapter 2, the expectation of $p$ is Bayes-Laplace

$$E(p) = \frac{r+A}{r+s+A+B} \tag{3.20}$$

### 3.3.6 Multinomial selection and Dirichlet's priors

In this section we assume we have several urns containing balls of different colours (terms). In each urn (document) we extract $l$ balls (tokens) and for each colour $t_i$ we observe $tf$ successes. Thus, each document is regarded as a small sample from an unknown population. This time we observe different terms and thus we need to generalize from the binary sampling of the previous section. The generalization of the Beta distribution, the Dirichlet distribution, is used to assign the priors in the multinomial case. Priors of a Dirichlet distribution have a set of parameters $A_1, \ldots, A_n > 0$, one for

each colour $t_i$.

$$(3.21) \qquad p(p_1,\ldots,p_n,A_1,\ldots,A_n) = \frac{\Gamma(A)}{\Gamma(A_1)\cdots\Gamma(A_n)} p_1^{A_1-1}\cdots p_n^{A_n-1}$$

$$A = \sum_{i=1}^{n} A_i$$

$$\sum_{i=1}^{n} p_i = 1$$

The mean of $p_i$ is $\frac{A_i}{A}$, the variance is $\frac{A_i(A-A_i)}{A^2(A+1)}$. Obviously if all parameters are equal, then the Dirichlet distribution is uniform. If there are only two parameters, $A_1$ and $A_2$, then the Dirichlet distribution is that of Section 3.3.5 obtained with the Beta distribution of Formula B.6 over $A_1$ and $A_2$.

# Chapter 4

# Models of IR based on divergence from randomness

This Chapter defines the *basic Information Retrieval models* under the following alternative assumptions:

1. A document $d$ is a sample and a document collection is a set of samples over different unknown populations. Our experiment thus consists in drawing balls from a set of urns. In other words, the occurrence of a word $t$ at the $k$-th place in a document $d$ is abstractly equivalent to the observation of a ball of colour $t$ at the $k$-th trial of the experiment relative to the sample $d$.

2. An alternative view, but identical in mathematical properties, is having a bag of balls of different colours which need to be placed into different bins or cells (the documents). An experiment thus consists in placing $TotFr_D$ balls of different colours $t$ into $N$ different cells. Each cell has the property of "attracting" the balls of colour $t$ with different probability $p_t$.

The use of one of the two types of models depends on the problem we need to formalize. The ball extraction type is more useful to formalise the apparent aftereffect in sampling (see Section 4.6) as well as to define the class of hypergeometric distributions (see Sections 2.3, 8.6 and 2.6.1), while the occupancy type is suitable, as we have seen in Section 2.4, to describe the Bose-Einstein statistics.

In the second part of the chapter the basic Information Retrieval models are normalized. We observed in the introductory Chapter (see discussion on page 31), the nonnormalized score distribution of the informative content against document rank decreases very rapidly (see Figure 1.1 on page 32). In fact, Manmatha, Rath and Feng [75] observe that, in general, the score distributions produced by good retrieval models are initially exponential and after follow a normal distribution (see Figure 1.3 on page 36).

We have seen that the informative content is additive, so that when the query is made up of two or more terms the informative content of their conjunction is the sum of the single informative content-weights. The informative content, in general, assumes very large values. For example, with the query "What is a *prime factor*?", while the highest value of the informative content of the term *prime* is 854.0 obtained with a document $d_1$, the highest value of the term *factor* is 2419.9 obtained with a different document $d_2$. The two terms of the query co-occur only in $d_1$ but not in $d_2$. Notwithstanding the additivity property, if we had used the informative content as document score we would have obtained the document $d_2$ as first retrieved document, but $d_2$ is not relevant because it does not contain the term *prime*. Therefore the informative content—weight does not work well under the *term independence assumption* since the slope of the informative content is initially very steep.

The role of the first normalization of the informative content is to resize suitably the informative content of a term by using only a small part of it. The part which is left as term–weight is proportional to the risk we take in choosing the term as a descriptor of the document. Risk and gain are related by the standard law of utility theory, that is

$$\frac{\text{gain}}{\text{gain} + \text{loss}} = \text{risk}$$

In the previous example, the relevant document $d_1$ positions rightly at the first place with the gain–weight. The way we compute the risk and the gain is described in Section 4.6. Before that, we introduce the notion of informative content of a term.

## 4.1 Basic Models

One of the most influential ideas on our our thoughts when we proposed the models of $IR$ based on divergence from randomness is the notion of *empirical* or *informative content*

*of a theory* which comes from the book "The logic of Scientific discovery" written by Karl Popper in 1934 [83]. Popper proposed to regard the informative content of a theory as its testability. Theories may have degrees of testability. Some theories may have more potential falsifiers than other theories, and the theories with higher degree of falsifiability are also less likely to be true. For this reason Popper calls the *logical probability* of a theory the proportion of the complementary set of all falsifiers of the theory. In other words, the logical probability $p_t$ of a statement $t$ is complementary to its degree of falsifiability $Inf(t)$. Popper gives other names to the notion of informative content, such as *the degree of confirmation* or *the degree of corroboration.* He thus proposed the following mathematical relationship between logical probability and informative content

$$Inf(t) = 1 - p_t \tag{4.1}$$

The logical reasoning adopted for falsifying theories is the *modus tollens*:

> If the theory holds, then the consequent; but the consequent is not; therefore the theory is falsified.

Kemeny, Good and Hamblin [1] independently suggested the definition of the degree of confirmation as

$$Inf(t) = -\log_2 p_t \tag{4.2}$$

Shannon [98, 99] in his Mathematical Theory of Communication also used the logarithmic measure for measuring the information contained in a message. Base 2 for the log corresponds to the choice of the binary digit as the unit of information. Shannon's Theory of Communication and Popper's ideas on the nature of information have influenced a number of philosophers and scientists [18, 10, 11, 33, 62, 61, 63] who coined the term, *Semantic Information Theory*, to denote the studies in Logic and Philosophy on the usage of the term *information*,

> in the sense in which it is used of whatever it is that meaningful sentences and other comparable combinations of symbols convey to one who understands them. [63]

[1] Popper cites Hamblin's unpublished thesis *Language and the Theory of Information* of 1955 and Good's report on Popper's paper in *Mathematical Review*, 17, page 367.

Conventional information theory is not concerned with the semantic aspects of communication which are irrelevant to the engineering problem of signal transmission. Semantic Information Theory investigates the axiomatization of logical principles for assigning probabilities to sentences, and it studies the relationship between informative content and probability.

Willis and Solomonoff [124, 108] use Equation ( 4.2) as a measure of the amount of information carried by an event and Goldman [43] develops information theory starting from Equation ( 4.2).

A second influence on our work is the notion of *randomness* as it is conceived in the notion of Kolmogorov complexity [71]. Kolmogorov complexity provides a definition of a random (finite or infinite) sequence. Regular sequences can be easily compressed, while random sequences do not possess shorter descriptions. The key theorem of Kolmogorov complexity is the existence of a universal Turing machine $U$ which computes all and only all partial recursive prefix functions. A prefix function is able to encode arbitrary sequences of natural integers using prefix-codes. An example of prefix code is the $\delta$–code used for example to compress inverted files of Information Retrieval[126]. We saw that in a Bernoulli process if the number of trials increases indefinitely, then the maximum likelihood approaches the prior probability of success. For short sequences the priors are too conclusive that we cannot perform an "objective" or "justified" probabilistic inference. In case that we have small empirical data Solomonoff[106, 107] proposed to assign a universal prior probability p which satisfies the following Coding Theorem

$$-\log_2 \mathrm{p}(x) = -\log_2 \sum_{U(p)=x} 2^{-l(p)} = K(x) \tag{4.3}$$

The equality holds up to an additive constant.

$2^{-K(x)}$ is the probability of having a prefix of complexity $K(x)$, that is the length $l(p)$ of the minimal prefix string $p$ such that the string $x$ is the output of the universal prefix machine $U$ on $p$.

The probability $2^{-K(x)}$ thus represents the cost of encoding the string $x$ by the shortest program $p$; $\sum_{U(p)=x} 2^{-l(p)}$ can be reduced to the computation of the probability of this representative up to some constant which does not depend on the string $x$. Random sequences thus have a large algorithmic complexity $K(x)$ and, complementarily, a

smaller prior probability $\mathbf{p}(x)$.

Thus the Coding Theorem states a powerful and theoretically appealing fact. Inductive inference can be solved using the algorithmic complexity $K$. For example in our framework, we may assign in principle the universal prior probability distribution to our Bernoulli trials. In practice, we have to specify an encoding language to represent our problem, which should be chosen to be as "optimal" as possible in relation to its compressibility power, and then we try to compute the encoding cost according to the chosen representation language.

We may define, for example, the probability relative to the encoding cost of a binary Bernoulli process of $l$ trials, to be the number of possible consistent configurations out of all configurations (as defined by Equation 2.3 or Equation 2.4). Therefore, the most informative sequences are those with the highest encoding cost, that is those sequences with the smallest binomial probability $B(l, tf, p)$ (as defined by Equation 2.1 or Equation 2.8).

Now, we saw that $B(l, tf, p)$ is maximised when $p$ is the maximum likelihood $\frac{tf}{l}$. This is equivalent to the fact that the divergence $D$ of $\frac{tf}{l}$ from $p$, as defined by Equation 2.14 of Section 2.2.2, is minimised. In case that $B(l, tf, p)$ is minimised, we observe that the frequencies in the sample diverge from those we would get by choosing a sample at random. In other words, the successful trials do not occur in the number predicted by the Bernoulli distribution.

*We here mean by "random" that the trials of the sample follow a Bernoulli process.* This notion of randomness differs from the notion of randomness given in algorithmic complexity theory, which is similar to Popper's notion of *objective disorder* or irregularity. In our case, by contrast a random selection of the sample brings about the regularity in the frequencies anticipated by the priors.

Nevertheless, the way we compute the informative content is the same, that is

$$-\log_2 p_t$$

where the probability $p_t$ is given by different probabilistic models of the sample space.

## 4.2 The informative content $Inf_1$ in the basic probabilistic models

In this section we assume that for each pair of a term and document, the following four random variables are given:

1. the total number of term tokens $TotFr_D$ in the collection $D$;
2. the term-frequency $F_t$ in the collection;
3. the cardinality $N$ of the collection;
4. the term-frequency $tf$ in the document $d$;

Let $Prob_1$ be a probability distribution over the sample space, and let $X$ be the random variable counting the occurrences of the term in the documents. With $Prob_1(tf|F_t, TotFr_D, N)$ we denote the probability that $X = tf$ with respect to the empirical data.

**Definition 6** The *informative content* of a term $t$ in a document $d$ is

$$Inf_1(tf|F_t, TotFr_D, N) = -\log Prob_1(tf|F_t, TotFr_D, N)$$

In the rest of this chapter the shorter expressions $Inf_1(tf)$ and $Prob_1(tf)$ will denote $Inf_1(tf|F_t, TotFr_D, N)$ and $Prob_1(tf|F_t, TotFr_D, N)$.

Our objective is the definition of the sample space over this population together with the assignment of a suitable probability distribution $Prob_1$.

## 4.3 The basic binomial model

We make the assumption that the $F_t$ tokens of a non–informative word $t$ distribute over $N$ documents according to the binomial law. The situation is abstractly equivalent to having a sequence of $F$ trials with $N$ possible outcomes, the documents, at each trial. The prior probability of having a given document as outcome is

$$p = \frac{1}{N}$$

We are now interested to determine the probability of observing $tf$ occurrences of $t$ in a document rather than the probability of the given configuration

$$tf_1 + \ldots + tf_N = F_t$$

which we have already encountered in Section 2.1.3 (see Equation 2.4). The probability of observing $tf$ occurrences in an arbitrary document out of $F$ Bernoulli trials is thus given by Equation 2.8, that is

$$(4.4) Prob_1(tf) = B(N, F, tf) = \begin{pmatrix} F \\ tf \end{pmatrix} p^{tf} q^{F-tf} \quad \text{where } p = \frac{1}{N} \text{ and } q = \frac{N-1}{N}$$

The maximum likelihood frequency of the term in the collection is

$$(4.5) \qquad \lambda = \frac{F}{N}$$

while the expectation $E(p)$ of $p$ is $\frac{tf+1}{F+2}$ (see the Bayes-Laplace relation 3.11)

Equation 4.4 was used by Harter to define his 2-Poisson model[52].

We saw in Section 2.1.3 that the configuration problem can also be seen as a scheme for placing a number $TotFr_D$ of $V$ terms in a collection of $N$ documents. For each term $t \in V$ a possible configuration of term placements satisfies Condition 2.4 and all terms satisfy the further condition (see Equation 2.5)

$$F_1 + \ldots + F_V = TotFr_D$$

So, the probability of a possible configuration that is consistent with the empirical data is the product of two multinomial distributions (see Equation 2.6 and 2.7), each deriving from the conditions 2.4 and 2.5.

The informative content of $t$ in a document $d$ is thus given by

$$(4.6) \qquad Inf_1(tf) = -\log_2 \left[ \begin{pmatrix} F \\ tf \end{pmatrix} p^{tf} q^{F-tf} \right] \quad \text{with } p = \frac{1}{N}$$

The reader may notice that the document-frequency $n$ (the number of different documents containing the term) is not used in this model.

### 4.3.1 The model $P$

The informative content 4.6 can be approximated using the Poisson distribution. We may use this approximation when the mean $p \cdot F$ of the Poisson distribution, which is the maximum likelihood $\lambda$ of 4.5, is small or moderate in magnitude. Therefore, we assume that the number $F$ of occurrences of the term is smaller than the number $N$ of documents in the collection. The informative content is that defined by the estimate 2.13. This formula generates the basic probabilistic model $P$. The acronym $P$ stands for Poisson.

$$Inf_1(tf) = tf \cdot \log_2 \frac{tf}{\lambda} + \left(\lambda + \frac{1}{12 \cdot tf} - tf\right) \cdot \log_2 e + 0.5 \cdot \log_2(2\pi \cdot tf) \quad [\text{ model } P]$$

$$\text{with } \lambda = \frac{F}{N} \text{ and } F \ll N \tag{4.7}$$

### 4.3.2 The model $D$

We have approximated the binomial distribution using the information theoretic divergence $D$ and Stirling's formula in Section 2.2.2.

From Equations 2.15 and 2.14 we derive the *basic probabilistic model D*. The acronym $D$ stands for Divergence of the mean $\phi = \frac{tf}{F}$ from $p = \frac{1}{N}$.

$$Inf_1(tf) = F \cdot (D(\phi, p) + 0.5 \log_2 (2\pi \cdot \phi \cdot (1 - \phi))) \quad [\text{ model } D] \tag{4.8}$$

$$\text{with } \phi = \frac{tf}{F} \text{ and } p = \frac{1}{N}$$

We will see that the models $P$ and $D$ do not produce different experimental results, so they are experimentally indistinguishable.

## 4.4 The basic Bose-Einstein model

In Section 2.4 we have seen that the Bose-Einstein statistics differs from the binomial distribution because all configurations satisfying Equation 2.4 are indistinguishable. The informative content $Inf_1$ derives from the probability in Equation 2.28

$$Inf_1(tf) = -\log_2 \frac{(N + F - tf - 2)!F!(N - 1)!}{(F - tf)!(N - 2)!(N + F - 1)!} \tag{4.9}$$

### 4.4.1 The model $G$

In Section 2.4.1 we have approximated Equation 2.28 with the geometric distribution.

From Equation 2.31 we derive the *basic probabilistic model G*

$$Inf_1(tf) = \log_2(1+\lambda) + tf \cdot \log_2\left(1+\frac{1}{\lambda}\right) \quad [\text{ model } G] \tag{4.10}$$
$$\text{with } \lambda = \frac{F}{N} \text{ and } F \ll N$$

The symbol $G$ stands for the geometric distribution. This approximation was obtained with $N$ large, that is with $F \ll N$.

### 4.4.2 The model $B_E$

The second approximation of the Bose-Einstein statistics comes from a direct application of the Stirling formula (see Equation 2.33):

$$Inf_1(tf) = -\log_2(N-1) - \log_2(e) + f(N+F-1, N+F-k-2) - f(F, F-k)$$
$$f(n,m) = (m+0.5)\cdot\log_2\left(\frac{n}{m}\right) + (n-m)\cdot\log_2 n \text{ and } F \ll N \text{ [ model } B_E] \tag{4.11}$$

We will see that the models $G$ and $B_E$ do not produce different experimental results, so they in practice coincide.

## 4.5 The tf-idf model

Let us choose a different sample space from binomial and Bose-Einstein statistics. We assign to each possible term-frequency $tf$ in a document $d$ the probability $p^{tf}q$ of the geometric distribution, where $p$ is the prior probability that $t$ occurs in the document. In fact the probability

$$Prob_1(tf>0) = \sum_{tf=1}^{\infty} q\cdot p^{tf} = q\cdot p\sum_{tf=0}^{\infty} p^{tf} = q\cdot p\frac{1-\lim_{n\to\infty} p^{tf+1}}{1-p} = p$$

Therefore, the *tf-idf probabilistic model* is a generalization of the model $G$ defined in Section 4.4.1. The tf-idf model coincides with $G$ in the case that the prior $p$ is set equal to $\frac{\lambda}{1+\lambda}$ with $\lambda = \frac{F}{N}$.

An alternative way to assign the prior $p$ exploits the document-frequency $n_t$, that is the number of documents of the collection containing the term $t$. The probability $p$ is the relative document-frequency $\frac{n_t}{N}$. Substituting these priors in the geometric distribution

we obtain

$$inf_1(tf) = \log_2 \frac{N}{N - n_t} + tf \cdot \log_2 \frac{N}{n_t}$$

Since in general $n_t \ll N$, we can assume $N \sim N - n_t$ that is

$$inf_1(tf) = tf \cdot \log_2 \frac{N}{n_t}$$

We saw that if the prior is given by the Bayes-Laplace estimate then

$$inf_1(tf) = tf \cdot \log_2 \frac{N+2}{n_t+1}$$

We also saw that, if the prior is assumed to be of the beta form with parameters $A$ and $B$, see Equation 3.20, then the Bayes rule generates

$$inf_1(tf) = tf \cdot \log_2 \frac{N+A+B}{n_t+A} \tag{4.12}$$

### 4.5.1 The model $I(n)$

From Equation 4.12 we can generate a class of models by varying the parameters $A$ and $B$. In the INQUERY Information retrieval system [2] the parameter values $A$ and $B$ are set to 0.5. In the absence of evidence, that is when the collection is empty with $N = n_t = 0$, for $A = B = 0.5$ the *a posteriori* probability $p$ has the maximum uncertainty value 0.5. We also set the values of $A$ and $B$ to 0.5. The basic probabilistic model $I(n)$ is thus defined as

$$inf_1(tf) = tf \cdot \log_2 \frac{N+1}{n_t+0.5} \quad [\text{model } I(n)] \tag{4.13}$$

The symbol $I(n)$ stands for "model with the Inverse of the document-frequency $n$".

### 4.5.2 The model $I(n_e)$

As in the previous section, we derive the equation similar to that 4.12 using the prior $p = \frac{n_e + A}{N + A + B}$, where $n_e$ is the number of expected documents containing the term according to the binomial law 2.8.

It is easy to compute this estimate from the binomial 2.8.

$$n_e = N \cdot (1 - B(F_t, 0, p)) = N \cdot \left(1 - q^{F_t}\right) \tag{4.14}$$

A new basic probabilistic model $I(n_e)$ is thus defined

$$inf_1(tf) = tf \cdot \log_2 \frac{N+1}{n_e+0.5} \quad \text{[model } I(n_e)\text{]} \tag{4.15}$$

The name $I(n_e)$ stands for "model with the Inverse of the expected document-frequency $n_e$".

### 4.5.3 The model $I(F)$

Note that $1 - B(N, F, 0) \sim 1 - e^{-\frac{F}{N}}$ by the Poisson approximation of the binomial, and that $1 - e^{-\frac{F}{N}} \sim \frac{F}{N}$ with an error of order $O(\left(\frac{F}{N}\right)^2)$. Using this approximation and assuming that $\frac{F}{N}$ is small, another basic probabilistic model $I(F)$ is thus derived from the model $I(n_e)$ (see Formula 4.15)

$$inf_1(tf) = tf \cdot \log_2 \frac{N+1}{F+0.5} \quad \text{[model } I(F)\text{]} \tag{4.16}$$

The name $I(F)$ stands for "model with the Inverse of the term-frequency $F$". A generalization of the $I(F)$ was given by Kwok [68] with the ICTF Weights (ICTF stands for the Inverse Collection Term Frequency Weights), in the context of the standard probabilistic model using relevance feedback information [90]. Kwok reported that the ICTF performed much better than Salton's inverse document-frequency model[94]. We show that in our experiments $I(F)$ and $I(n_e)$ behave similarly and irrespective to the other normalization components.

## 4.6 First normalization of the informative content

Starting with this section we resume the systematic exposition of the second component of the retrieval models introduced in Chapter 1.

We have abstractly reproduced the Information Retrieval process as we had coloured balls to place in or extract from urns. If sampling is with replacement, then the population is not changed and the probability of extracting the same ball in a successive trial is invariant. The conditional probability of having a sequence, for example, of $\{red, red\}$ when the first trial is *red* is the same as the prior probability of having $\{red\}$.

In a Bernoulli process the "holding time" between the appearance of two balls of the same colour does not depend on how long the ball has not appeared in the past. This

situation is described by a process with a complete lack of memory, that is past outcomes do not modify the expectation of the event.

A different situation is when we are searching for tokens of a term and after a long unsuccessful search we find a few of then in a portion of a document. It is quite likely that we have finally reached a document for which we can expect an increased rate of success in continuing our search. The more occurrences we find the higher is our expectation.

We assume that this expectation is measured by $Prob_2(tf)$ in Formula 1.1. The probability $Prob_2(tf)$ has been called by statisticians an apparent *aftereffect* of future sampling [37, pp118-125].

The intuition underlying the aftereffect in IR is that the greater the term-frequency $tf$ of a term in a document, the more the term is contributing to discriminating that document.

There are several models for modelling the aftereffect, one of these is the law of succession of Laplace [44] discussed in Section 3.3.2. Simulating aftereffect with our urn model consists in replace the extracted ball with other balls of the same or different colour. A second possibility is to use different urns having a prior probability of being selected, but once the urn has been selected the balls continue to be drawn from the chosen urn.

Feller[37, page 119] pictures the urn models of aftereffect as the results of a super-human game of chance. If at each trial the chance of a rare event, like the occurrence of an accident, remains constant in time as expected then the population of the urn continues to be the same. But we may assume that an occurrence of that rare event has an aftereffect in the fact that it increases rapidly the chance to occur soon again. The more the rare event is observed at regular time intervals the more the chance of new occurrences of this event increases. *In other words, the probability of the event in the successive interval of time is conditioned by its frequency $tf$ in previous intervals of time.*

Turning back to the IR context, if $tf$ is large then the probability that the term may select a relevant document is high. The fact that $tf$ is large depends also on the length of the document. Moreover, relevant documents may have different lengths and we cannot predict the length of a relevant document. Therefore we assume for the moment that the

length of a relevant document is of arbitrary and large in size. In Section 6.2 we show how to normalize the actual document length $l$ to a standard length. When enlarging the actual size of a relevant document to an arbitrary large size, the chance of encountering a new token of the observed term increases in accordance with the size $tf$ of already observed tokens.

We thus assume that the probability that the observed term contributes to the selection of a relevant document is high, if the probability of encountering one more token of the same term in a relevant document is similarly high. We reason that a high expectation of encountering one more occurrence will be due to some underlying semantic cause and will not be simply random. The probability of a further success in encountering a term is thus a conditional probability which approaches 1 as $tf$ increases and becomes large. On the other hand, if successes were brought about by pure chance, then the conditional probability would tend to approach 0 as $tf$ increases and becomes large. We need however, a method to estimate our conditional probability.

We assume that the probability $Prob_2(tf)$ is related only to the "elite set" of the term, which is defined to be the set $E_t$ of all documents containing the term. We also assume that the probability $Prob_2(tf)$ in Formula 1.1 is obtained by a conditional probability $p(tf+1|tf,d)$ of having one more occurrence of $t$ in the document $d$ and that $p(tf+1|tf,d)$ is obtained from an aftereffect model.

This probability is computed in the next two Sections.

### 4.6.1 The first normalization $L$

The first model of $Prob_2(tf)$ is given by Laplace's law of succession. The law of succession in this context is used when we have no "advance knowledge" of how many tokens of a term should occur in a relevant document of arbitrary large size. The Laplace model of aftereffect is explained by Feller [37]. Feller shows that the probability $p(tf+1|tf,d)$ is close to $\frac{tf+1}{tf+2}$ and does not depend on the document length.

Laplace's law of succession is thus obtained by assuming that:

i) the probability $Prob_2(tf)$ modelling the aftereffect in the elite set in Formula 1.1 is given by the conditional probability of having one more token of the term in the document, that is passing from $tf$ observed occurrences to $tf+1$, and

ii) the length of a document is very large.

A Bayesian derivation of Laplace's law of succession was given in Section 3.3.5 with Formula 3.20 at page 3.20. According to this formula we would have

$$Prob_2(tf) = \frac{tf + A}{tf + A + B} \tag{4.17}$$

Similarly, if $tf \geq 1$ then $Prob_2(tf)$ can be given by the conditional probability of having $tf$ occurrences assuming that $tf - 1$ have been observed. If $A = B = 0.5$ we get the following Equation

$$Prob_2(tf) = \frac{tf + 0.5}{tf + 1} \tag{4.18}$$

Equations 1.1 and 4.18 give the normalization $L$:

$$\begin{aligned} weight(t,d) &= \frac{B}{tf + A + B} \cdot Inf_1(tf) \\ &\propto \frac{1}{tf + 1} \cdot Inf_1(tf) \end{aligned} \tag{4.19}$$

The constant $B = 0.5$ is ignored being independent from the term and thus not influential to the ranking. In our experiments, which we do not report here for sake of space, the parameters $A = B = 0.5$ of Relation 4.17 performs better than other values, therefore we refer to the Formula 4.19 as *First Normalization* $L$ of the informative content.

### 4.6.2 The first normalization $B$

The *second* model of $Prob_2(tf)$ is slightly more complex than that given by Relation 4.18.

The conditional probability of Laplace's law directly computes the aftereffect on future sampling. The hypothesis about aftereffect is that any newly encountered token of a term in a document is not obtained by accident. If we admit that randomness is not the cause of encountering new tokens then the probability of encountering a new token must increase (or decrease) with respect to the probability which is expected by randomness. Hence, the aftereffect on the future sampling is obtained by a process in which the probability of obtaining a newly encountered token is *inversely related* to that which would be obtained by accident.

In other words, the aftereffect of sampling from *the elite set* yields a distribution which departs from one of the "ideal" schemes of randomness we described before. Therefore, we can model the aftereffect process by Bernoulli.

However, a sequence of Bernoulli trials is a process characterized by a complete lack of memory (lack of aftereffect).

It is known that previous successes or failures do not influence successive outcomes. The lack of memory does not allow us to use Bernoulli trials, as for example in the ideal urn model defined by Laplace, since the conditional probability $p(tf+1|tf,d)$ would be constant for all $tf$.

To obtain the estimate $Prob_2$ with Bernoulli trials we use the following urn model.

We add a new token of the term to the collection, having thus $F+1$ tokens instead of $F$. We then compute the probability $B\left(F+1,tf+1,\frac{1}{n}\right)$ that this new token falls into the observed document, thus having a within–document term–frequency $tf+1$ instead $tf$.

The probability $B\left(F+1,tf+1,\frac{1}{n}\right)$ is thus that of obtaining *by accident* one more token of the term $t$ in the document $d$ out of all $n$ documents in which $t$ occurs when a new token is added to the elite set.

The comparison

$$\frac{B\left(F+1,tf+1,\frac{1}{n}\right)}{B\left(F,tf,\frac{1}{n}\right)}$$

of the new probability $B\left(F+1,tf+1,\frac{1}{n}\right)$ to the previous one $B\left(F,tf,\frac{1}{n}\right)$ tells us whether the probability of encountering a new occurrence is increased or diminished by sampling from our urn model.

Therefore, we may talk in this case of an *incremental rate* $\frac{\Delta B}{B}$ of term-occurrence in the elite set rather than of probability $Prob_2$ of term-occurrence in the elite set.

We suppose that the incremental rate of occurrence is

$$\frac{\Delta B}{B} = \frac{B\left(F,tf,\frac{1}{n}\right)-B\left(F+1,tf+1,\frac{1}{n}\right)}{B\left(F,tf,\frac{1}{n}\right)} = 1-\frac{B\left(F+1,tf+1,\frac{1}{n}\right)}{B\left(F,tf,\frac{1}{n}\right)} \tag{4.20}$$

If the ratio of two Bernoulli processes

$$\frac{B\left(F+1,tf+1,\frac{1}{n}\right)}{B\left(F,tf,\frac{1}{n}\right)} \tag{4.21}$$

is smaller than 1, then the probability of having received at random the newly added token increases.

Since the binomial decreases very rapidly at increasing values of the term-frequency, the larger the $tf$ the less accidental one more occurrence of the term is. Therefore, accepting the term as a descriptor of a potentially relevant document is less risky. Equation 4.21 is a ratio of two binomials as given by Equation 4.4 (but using the elite set with $p = \frac{1}{n}$ instead of $p = \frac{1}{N}$):

$$\frac{\Delta B}{B} = 1 - \frac{B\left(F+1, tf+1, \frac{1}{n}\right)}{B\left(F, tf, \frac{1}{n}\right)} = \tag{4.22}$$

$$= 1 - \frac{F+1}{n \cdot (tf+1)}$$

The equations 1.1 and 4.22 give

$$weight(t,d) = \frac{B\left(F+1, tf+1, \frac{1}{n}\right)}{B\left(F, tf, \frac{1}{n}\right)} \cdot Inf_1(tf) = \frac{F+1}{n \cdot (tf+1)} \cdot Inf_1(tf) \tag{4.23}$$

## 4.7 Relating the aftereffect probability $Prob_2$ to $Inf_1$

In this section we set out a formal derivation of Formula 1.1, which describes the relationship between the elite set and the statistics of the whole collection, which involves showing how the probabilities $Prob_2$ and $Prob_1$ are combined. The use of the gain as term-weighting is completely new, as well as its relation to the apparent aftereffect of sampling. We stress that also the idea of using $Inf_1$ for defining basic Information Retrieval models is original, although we saw that the standard tf–idf term-weighting can be interpreted as informative content (see the model $I(n)$).

Let us assume that a term $t$ belongs to a query $q$. We assume that if the term $t$ also occurs in a document then *we accept it as a descriptor* for a potentially relevant document (relevant to the query $q$). A gain and a loss are thus achieved by accepting the query term $t$ as a descriptor of a potentially relevant document. The gain is the amount of information we will get if the document turns out to be actually relevant. The gain is thus a fraction of $Inf_1(tf)$. What remains of $Inf_1(tf)$ is the loss, and the loss is produced in the case that the document turns out not to be relevant. This translates

into the equation:

$$\text{gain} + \text{loss} = Inf_1(tf) \tag{4.24}$$

We weight the term by computing only the expected gain, namely

$$weight(t,d) = \text{gain}$$

The conditional probability $Prob_2(tf)$ of occurrence of the term $t$ is related to the odds in the standard way (the higher its probability the smaller the gain):

$$Prob_2(tf) = \frac{\text{loss}}{\text{gain} + \text{loss}} \tag{4.25}$$

From Equation 4.25 the loss is

$$\text{loss} = Prob_2(tf) \cdot Inf_1(tf) \tag{4.26}$$

For scoring documents we use only the gain, which from 4.24 and 4.26 is

$$\begin{aligned} weight(t,d) = \text{gain} = & \quad Inf_1(tf) - \text{loss} \\ = & \quad (1 - Prob_2(tf)) \cdot Inf_1(tf) \end{aligned} \tag{4.27}$$

**Example 1** *As an example, let us consider the term "progress" which occurs* 22,789 *times in a collection containing 567,529 documents. Let us use the Poisson model P for computing the amount of information* $Inf_1$ *and use Laplace's law of succession to compute the loss and the gain of accepting the term as a descriptor for a potentially relevant document. We distinguish two cases: the term-frequency in the document is equal to* 0 *or not. In the second case suppose* $tf = 11$ *as an example. We construct the following contingency table:*

| | *Accept* ($tf = 11$) | *Not accept* ($tf = 0$) |
|---|---|---|
| *d is relevant* | $\text{gain}_1 = 6.9390$ | $\text{loss}_0 = 0.04015$ |
| *d is not relevant* | $\text{loss}_1 = 69.3904$ | $\text{gain}_0 = 0$ |
| | $Inf_1 = 76.3295$ | $Inf_0 = 0.04015$ |

*First we compute the amount of information* $Inf_1 = 76.3295$ *as given by the formula 2.9 with* $tf = 11$ *and* $1 - Prob_2(tf) = 1 - \frac{10}{11} = 0.0909$ *from 4.18, then* $\text{gain}_1$ *is obtained by multiplying these two values. Similarly,* $\text{loss}_1 = 0.9090 \cdot 76.3295 = 69.3904$.

*When $tf = 0$ we reject the term, that is the term is considered not to be a descriptor of a potentially relevant document, so by rejecting the term we have a gain when the term "progress" is not important for predicting the relevance of the document. According to Laplace's law of succession the gain is* 0*, while the loss is very small.*

## 4.8 First Normalized Models of Divergence from Randomness

The first normalization factor of $Inf_1(tf)$ of Equation 4.19 is denoted by $L$ (for Laplace), while the normalization of Equation 4.23 is denoted by $B$ (for Binomial). First Normalized Models of IR are obtained from the basic models

- $P$ of Equation 4.7 on page 86;
- $D$ of Equation 4.8 on page 86;
- $G$ of Equation 4.10 on page 87;
- $B_E$ of Equation 4.11 on page 87;
- $I(n)$ of Equation 4.13 on page 88;
- $I(n_e)$ of Equation 4.15 on page 89;
- $I(F)$ of Equation 4.16 on page 89

applying the first normalization

- $L$ of Equation 4.19 on page 92 or
- $B$ of Equation 4.23 on page 94.

For the sake of completeness we here list the 14 first normalized models of divergence from randomness which we test in this dissertation.

### 4.8.1 Model $PL$

$$w(tf) = \frac{1}{tf+1}\left(tf \cdot \log_2 \frac{tf}{\lambda} + \left(\lambda + \frac{1}{12 \cdot tf} - tf\right) \cdot \log_2 e + 0.5 \cdot \log_2(2\pi \cdot tf)\right) \quad [\text{ model } PL] \tag{4.28}$$

$$\text{with } \lambda = \frac{F}{N} \text{ and } F \ll N$$

### 4.8.2 Model $PB$

$$w(tf) = \frac{F+1}{n\cdot(tf+1)}\left(tf\cdot\log_2\frac{tf}{\lambda}+\left(\lambda+\frac{1}{12\cdot tf}-tf\right)\cdot\log_2 e+0.5\cdot\log_2(2\pi\cdot tf)\right) \quad [\text{ model } PB] \tag{4.29}$$

$$\text{with } \lambda=\frac{F}{N} \text{ and } F\ll N$$

### 4.8.3 Model $DL$

$$w(tf) = \frac{1}{tf+1}\left(F\cdot\left(D(\phi,p)+0.5\log_2\left(2\pi\cdot\phi\cdot(1-\phi)\right)\right)\right) \quad [\text{ model } DL] \tag{4.30}$$

$$\text{with } \phi=\frac{tf}{F}, D \text{ as in Equation 2.14 and } p=\frac{1}{N}$$

### 4.8.4 Model $DB$

$$w(tf) = \frac{F+1}{n\cdot(tf+1)}\left(F\cdot\left(D(\phi,p)+0.5\log_2\left(2\pi\cdot\phi\cdot(1-\phi)\right)\right)\right) \quad [\text{ model } DB] \tag{4.31}$$

$$\text{with } \phi=\frac{tf}{F}, D \text{ as in Equation 2.14 and } p=\frac{1}{N}$$

### 4.8.5 Model $GL$

$$w(tf) = \frac{1}{tf+1}\left(F\cdot\left(\log_2(1+\lambda)+tf\cdot\log_2\left(1+\frac{1}{\lambda}\right)\right)\right) \quad [\text{ model } GL] \tag{4.32}$$

$$\text{with } \lambda=\frac{F}{N} \text{ and } F\ll N$$

### 4.8.6 Model $GB$

$$w(tf) = \frac{F+1}{n\cdot(tf+1)}\left(\log_2(1+\lambda)+tf\cdot\log_2\left(1+\frac{1}{\lambda}\right)\right) \quad [\text{ model } GB] \tag{4.33}$$

$$\text{with } \lambda=\frac{F}{N} \text{ and } F\ll N$$

### 4.8.7 Model $B_EL$

$$w(tf) = \frac{1}{tf+1}\left(-\log_2(N-1)-\log_2(e)+f(N+F-1,N+F-k-2)-f(F,F-k)\right)$$

$$\text{with } f \text{ as in Relation 4.11 } \textit{and } F\ll N \text{ [ model } B_EL] \tag{4.34}$$

### 4.8.8 Model $B_E B$

$$w(tf) = \frac{F+1}{n \cdot (tf+1)} \left(-\log_2 (N-1) - \log_2(e) + f\left(N+F-1, N+F-k-2\right) - f\left(F, F-k\right)\right) \tag{4.35}$$

with $f$ as in Relation 4.11 *and* $F \ll N$ [ model $B_E B$]

### 4.8.9 Model $I(n)L$

$$w(tf) = \frac{1}{tf+1} \left(tf \cdot \log_2 \frac{N+1}{n_t + 0.5}\right) \quad \text{[model } I(n)L\text{]} \tag{4.36}$$

### 4.8.10 Model $I(n)B$

$$w(tf) = \frac{F+1}{n \cdot (tf+1)} \left(tf \cdot \log_2 \frac{N+1}{n_t + 0.5}\right) \quad \text{[model } I(n)L\text{]} \tag{4.37}$$

### 4.8.11 The model $I(n_e)L$

$$w(tf) = \frac{1}{tf+1} \left(tf \cdot \log_2 \frac{N+1}{n_e + 0.5}\right) \quad \text{[model } I(n_e)L\text{]} \tag{4.38}$$

where $n_e$ is as in equation 4.14.

### 4.8.12 The model $I(n_e)B$

$$w(tf) = \frac{F+1}{n \cdot (tf+1)} \left(tf \cdot \log_2 \frac{N+1}{n_e + 0.5}\right) \quad \text{[model } I(n_e)B\text{]} \tag{4.39}$$

where $n_e$ is as in equation 4.14.

### 4.8.13 Model $I(F)L$

$$w(tf) = \frac{1}{tf+1} \left(tf \cdot \log_2 \frac{N+1}{F + 0.5}\right) \quad \text{[model } I(F)L\text{]} \tag{4.40}$$

### 4.8.14 Model $I(F)B$

$$w(tf) = \frac{F+1}{n \cdot (tf+1)} \left(tf \cdot \log_2 \frac{N+1}{F + 0.5}\right) \quad \text{[model } I(F)B\text{]} \tag{4.41}$$

# Chapter 5

# Related IR models

Before we develop the last component of the probabilistic Information Retrieval models based on divergence from randomness, it is useful and important to examine the existing models for Information Retrieval, that have influenced our work. By models we include both weighting functions for document retrieval and term-indexing. A complete and detailed presentation of the probabilistic models can be found in [41, 24]. Therefore, we here do not try to survey all of them but to introduce and discuss the most important aspects of them to make a tight comparison with our work and we try and we take a technical view of them as close as possible to that used in our investigation.

In particular the language modelling, described in Sections 5.4 and 5.5, are also applied to the models of divergence from randomness in Chapter 6.

## 5.1 The vector space model of IR

The vector space model is a class of models rather than a single model. Many models can be traced back to that implemented by the well known, the SMART system[93, 96], one of the first experimental systems of Information Retrieval. The basic matching function between the document $d$ and the query $q$ is obtained by considering both queries and documents as vectors $\mathbf{d}, \mathbf{q} \in \mathbf{R}^V$ and computing the similarity of documents and queries with the cosine function of the corresponding vectors. $\mathbf{R}^V$ is the product space of the vocabulary $V$. The cosine is the inner product normalized by the norms of the vectors:

$$sim(q, d) = \frac{(\mathbf{d}, \mathbf{q})}{\|\mathbf{d}\| \|\mathbf{q}\|} \tag{5.1}$$

The term-weights making up the document vector $\mathbf{d} = (w_1, w_2, \ldots, w_V)$ define a variant of the vector space model. The simplest term-weighting function is [96, see Chapter 3] the *Inverse Document Frequency Weight*

$$w_k = tf\left(\log_2 \frac{N}{n_t} + 1\right) \tag{5.2}$$

where $\frac{n_t}{N}$ is the relative document frequency of the $k$-th term $t$ , that is the ratio of the number $n_t$ of documents in which the term occurs and the number $N$ of documents in the collection, and $\|\mathbf{d}\| = \sqrt{\sum_{t_k \in d} w_k^2}$ is the norm.

When the within–document term-frequencies $tf$, relative to the terms of the query, are all the same in two or more documents, the longest document receives the smallest similarity score, since its norm $\|\mathbf{d}\|$ is larger. In the vector space model there is thus an implicit normalization of the term-frequency with respect to the document length. This problem in Information Retrieval is called *length normalization.* For many years the length normalization performed by the similarity induced by the cosine function in the vector space model has been considered very robust. But, on the other hand, the cosine normalization requires the computation of the norm of the documents which can be heavy in implementation. A retrieval model in order to be a genuine alternative to the vector space model must use a different method from the cosine function for normalizing the term-frequency $tf$. For example, it could consider the document length $l_d$ as an explicit random variable of the probability space. This is what the $BM25$ formula does: the $BM25$ abandons the normalization provided by the cosine for the inclusion of the document length in the weighting formula. We discuss more extensively the length normalization problem in Chapter 6.

## 5.2 The standard probabilistic model of IR

Information Retrieval models, vector space model included, are based on probability theory [90, 51, 119, 91, 40, 127, 117, 115, 25, 128, 82], but the meaning of the term "probabilistic" pertains to the explicit usage of relevance as an element *rel* of the algebra of events. In other words, a probabilistic model ranks documents satisfying the so called "Probability Ranking Principle" ($PRP$) [89]. The $PRP$ asserts that documents should be ranked according to the decreasing ordering established by the probability of relevance

$p(rel, d|q)$ with respect to a given query $q$. Bayes' theorem relates the *a posteriori* probability $p(rel, d|q)$ with the *likelihood* (also called the *a priori* probability) $p(q, d|rel)$ and the priors $p(rel), p(t)$.

First, *term independence assumption* simplifies the computation of the likelihood

$$p(q, d|rel) = \prod_{t \in q} p(t, d|rel)$$

where

$$p(t, d|rel) = \begin{cases} p(t|rel) & \text{if } t \in d \\ 0 & \text{otherwise} \end{cases}$$

Then, Bayes' theorem is

$$p(rel, d|t) = \frac{p(t, d|rel) \cdot p(rel)}{p(t)} \tag{5.3}$$

In the binary independence indexing model of Maron–Kuhns and Fuhr [78, 40] Bayes's theorem is applied with a different reading, that is

$$p(rel|t, d) = \frac{p(t|rel, d) \cdot p(rel|d)}{p(t|d)} \tag{5.4}$$

All terms in the observed document $d$ of Equation 5.4 are regarded as items of evidence in Bayes' rule. This is equivalent to assuming that we are mainly sampling from a single document, like the sampling in the language model shown in Chapter 3. As a consequence the prior $p(rel|d)$ should be assigned considering the document $d$ as an item of evidence, whilst Equation 5.3 provides a prior probability to relevance which is not conditioned by the observed document. This view is actually implemented by the standard probabilistic model. Notwithstanding this discrepancy, it comes out indeed that the results are practically the same for both formalizations and we may equally draw the same final weighting formula. For the above considerations, we prefer to assume Formula 5.3 as generating formula of the standard probabilistic model.

In the Croft and Harper model [27] the prior $p(t)$ can be regarded as an approximation of the probability of occurrence of the term $t$ in only non-relevant documents $p(t|\overline{rel})$. This assumption leads to the relation

$$p(rel, d|t) = \frac{p(t, d|rel) \cdot p(rel)}{p(t|\overline{rel})}$$

Since $p(rel)$ is a constant and observing that the Boolean event "$t, d$" in Equation 5.3 stands for $t \in d$, then we derive the posterior probability distribution of relevance:

$$p(rel, d|q) \propto \prod_{t \in q \cap d} \frac{p(t|rel)}{p(t|\overline{rel})} \tag{5.5}$$

The Robertson Sparck–Jones[90] model uses instead the cross product ratio:

$$p(rel, d|q) \propto \prod_{t \in q \cap d} \frac{p(t|rel) \cdot p(\bar{t}|\overline{rel})}{p(t|\overline{rel}) \cdot p(\bar{t}|rel)} \tag{5.6}$$

If the probabilities involved in relation 5.6 are estimated by using the counting measure over the set of documents and if $R$, $r$ and $n$ respectively denote the cardinalities of the set of relevant documents, the set of relevant documents in which the term $t$ occurs and the set of documents in which the term $t$ occurs, then we obtain (see the contingency Table 5.1):

Table 5.1: The contingency table in the probabilistic model.

| | $t$ | $\bar{t}$ | |
|---|---|---|---|
| $rel$ | $r$ | $R-r$ | $R$ |
| $\overline{rel}$ | $n-r$ | $N-n-R+r$ | N-R |
| | $n$ | $N-n$ | $N$ |

$$p(rel, d|q) \propto \prod_{t \in q \cap d} \frac{r \cdot (N - R - n + r)}{(n - r) \cdot (R - r)} \tag{5.7}$$

We can transform this probability into an additive weighting formula using the monotonic function log:

$$p(rel, d|q) \propto \sum_{t \in q \cap d} \log \frac{r \cdot (N - R - n + r)}{(n - r) \cdot (R - r)} \tag{5.8}$$

In the circumstances that one or more components of the cross product ratio becomes null, a smoothing constant, for example 0.5, is added to each component of the cross product

$$p(rel, d|q) \propto \sum_{t \in q \cap d} \log \frac{(r + 0.5) \cdot (N - R - n + r + 0.5)}{(n - r + 0.5) \cdot (R - r + 0.5)} \tag{5.9}$$

Therefore, when no knowledge on relevance is available, that is when $R = r = 0$, the probability of relevance becomes proportional to:

$$p(rel, d|q) \propto \sum_{t \in q \cap d} \log \frac{(N - n + 0.5)}{(n + 0.5)} \tag{5.10}$$

The term–frequency within the document or the query is not yet taken into account in the Robertson–Sparck Jones model. Regardless of the number $tf$ of occurrences of the term in the document this model assigns the same weight to every document containing the term. The 2-Poisson model of Harter suggests a way of extending the Relation 5.9 and including the statistics about the observed document[52, 53, 54]. We show how to modify the model in Section 5.2.2 but first we introduce the 2-Poisson model.

### 5.2.1 The 2-Poisson model

The 2-Poisson model is a probabilistic model of keyword indexing. It cannot be regarded directly as a retrieval model. The purpose of Harter's work is to identify the keywords likely to be informative for an arbitrary document. Such words are called *specialty* words by Harter in contraposition to the other ones, the *non-specialty* ones, which instead are considered to occur in documents at random. The origin of the 2-Poisson model can be traced back to Maron, Damerau, Edmundson and Wyllys[73, 77, 34, 30]. In these works it is observed that the divergence between the rare usage of a word across the document collection and the contrasting relative within-document frequency constitutes a revealing indication of the informative status of a word. Damerau suggests selecting the high status words by making the assumption that the Poisson distribution describes these frequencies. If the probability results in a very small value, then the word is marked as an index term. Obviously, not all words always fall either into one class or into the other. But none the less, many word tokens occur randomly in many documents while they occur more densely and nonrandomly in a few documents. This set of documents is the *Elite set* of the term, and is the set of documents which extensively connects with the concept or the semantics related to the term. The Elite set $E$ attracts the tokens with an expected rate $\lambda_E$. Tokens fall randomly into the other class with a lower rate $\lambda_{\overline{E}}$. The final probability of occurrence of the term in a document is given by the mixture of these two Poisson distributions [67]:

$$prob(tf) = \alpha \cdot \frac{e^{-\lambda_E}\lambda_E^{tf}}{tf!} + (1-\alpha) \cdot \frac{e^{-\lambda_{\overline{E}}}\lambda_{\overline{E}}^{tf}}{tf!} \tag{5.11}$$

The probability of a term $t$ to appear $tf$ times in a document $d$ belonging to the Elite set $E$ is given by the conditional probability

$$prob(X = tf, d \in E|tf) = \frac{prob(X = tf, d \in E)}{prob(tf)}$$

$$= \frac{\alpha \cdot \frac{e^{-\lambda_E}\lambda_E^{tf}}{tf!}}{\alpha \cdot \frac{e^{-\lambda_E}\lambda_E^{tf}}{tf!} + (1-\alpha) \cdot \frac{e^{-\lambda_{\overline{E}}}\lambda_{\overline{E}}^{tf}}{tf!}}$$

That is

$$prob(tf, d \in E|tf) = \frac{1}{1 + \beta \cdot e^{(\lambda_E - \lambda_{\overline{E}})}\left(\frac{\lambda_{\overline{E}}}{\lambda_E}\right)^{tf}} \tag{5.12}$$

where $\beta = \frac{1-\alpha}{\alpha}$ and $\alpha = \lambda_E = 0$ in the case that the word belongs to the non-specialty class. The conditional probability of Equation 5.12 is used by Harter to generate a ranking of the most informative words. However, the 2-Poisson model requires the estimation of 3 parameters for each word of the vocabulary, and this is a real drawback for the direct practical application of his model to term selection or term-weighting problems.

A last remark concerns the N-Poisson model, the generalization of the 2-Poisson model. Any probability distribution on $(0, \infty)$ can be defined as a mixing distribution of Poissons [84]. Therefore, it is true that each word distributes following a N-Poisson distribution for some N. N-Poisson models thus have a practical application only when N assumes a very small value, such as in the case of the 2-Poisson model or the 3-Poisson model.

In the next section we see how as much as possible was taken from the 2-Poisson model to solve the term-weighting problem.

### 5.2.2 The $BM25$ matching function

Interestingly, eliteness is a hidden variable of the 2-Poisson model. This is reflected in the fact that three parameters $\alpha$, $\lambda_E$ and $\lambda_{\overline{E}}$ need to be estimated in Equation 5.11. The matter becomes more complicated, if we want to exploit the 2-Poisson indexing model to enhance the probabilistic use of relevance in the retrieval model. The combination

of Equation 5.6 of the probabilistic model with Equation 5.11 of the 2-Poisson model needs reasonable approximations for making the mixture a workable retrieval model.

If we simplify the relationship among terms which compound the given query assuming that the terms are stochastically independent, then the combination of the notion of eliteness with that of relevance generates the Robertson, van Rijsbergen and Porter Equation[91]:

$$w = \ln \frac{(p_1 {\lambda_E}^{tf} e^{-\lambda_E} + (1-p_1){\lambda_{\overline{E}}}^{tf} e^{-\lambda_{\overline{E}}})(p_2 e^{-\lambda_E} + (1-p_2)e^{-\lambda_{\overline{E}}})}{(p_2 {\lambda_E}^{tf} e^{-\lambda_E} + (1-p_2){\lambda_{\overline{E}}}^{tf} e^{-\lambda_{\overline{E}}})(p_1 e^{-\lambda_E} + (1-p_1)e^{-\lambda_{\overline{E}}})} \tag{5.13}$$

where

$$p_1 = prob(d \in E|rel) \text{ and } p_2 = prob(d \in \overline{E}|rel)$$

Equation 5.13 is equivalent to

$$w = \ln \frac{\left(p_1 + (1-p_1)\left(\frac{\lambda_{\overline{E}}}{\lambda_E}\right)^{tf} e^{\lambda_E - \lambda_{\overline{E}}}\right)\left(p_2 e^{-\lambda_E + \lambda_{\overline{E}}} + (1-p_2)\right)}{\left(p_2 + (1-p_2)\left(\frac{\lambda_{\overline{E}}}{\lambda_E}\right)^{tf} e^{\lambda_E - \lambda_{\overline{E}}}\right)\left(p_1 e^{-\lambda_E + \lambda_{\overline{E}}} + (1-p_1)\right)} \tag{5.14}$$

Let

$$w(tf) = \ln C(tf) + \ln C_0$$

be the Equation 5.14 where $C_0$ is the ratio of the two components of the cross product not containing the variable $tf$. The first derivative with respect to the variable $tf$ is

$$w' = \frac{(p_2 - p_1) \cdot e^{\lambda_E - \lambda_{\overline{E}}} \cdot \left(\frac{\lambda_{\overline{E}}}{\lambda_E}\right)^{tf} \cdot \ln\left(\frac{\lambda_{\overline{E}}}{\lambda_E}\right)}{C(tf)}$$

Note that $\ln\left(\frac{\lambda_{\overline{E}}}{\lambda_E}\right) < 0$ because $\lambda_{\overline{E}} < \lambda_E$ in the 2-Poisson model. Therefore $w$ is a monotonically increasing function under the hypothesis that $p_2 < p_1$, which obviously holds since the size of the elite set of a term is assumed to be small in Harter's model.

The limiting form of Equation 5.14 for $tf \to \infty$ is

$$w = \frac{p_1\left(p_2 e^{-\lambda_E + \lambda_{\overline{E}}} + (1-p_2)\right)}{p_2\left(p_1 e^{-\lambda_E + \lambda_{\overline{E}}} + (1-p_1)\right)}$$

Since $e^{-\lambda_E + \lambda_{\overline{E}}} \sim 0$, this limit is very close to

$$\frac{p_1(1-p_2)}{p_2(1-p_1)} \tag{5.15}$$

Without loss of generality, since the weighting function is monotonic with respect to the within–document term-frequency $tf$, when we rank documents according to the weight 5.14 we may assume that the topmost documents of the ranking have their $tf$ value higher than that of the documents lower in the ranking. Hence, the limiting form 5.15 can be taken as the actual score of the topmost documents.

Robertson and Walker [87] define as an approximation of Equation 5.15 the product

$$w = \frac{tf}{tf + K} \cdot \frac{p(t|rel)p(\bar{t}|\overline{rel})}{p(t|\overline{rel})p(\bar{t}|rel)} \tag{5.16}$$

Indeed, both Equations 5.14 and 5.16 have Formula 5.15 as limit for large $tf$. Varying the parameter $K$ and using Relation 5.9 we obtain the so-called $BM$ weighting formulae ($BM$ stands for Best Match):

$$w = \frac{tf}{k + tf} \ln \frac{(r + 0.5) \cdot (N - R - n + r + 0.5)}{(n - r + 0.5) \cdot (R - r + 0.5)} \quad [BM\text{'s family}]$$

The $BM25$ matching function is:

$$(5.17 \sum_{t \in q} \frac{(k_1 + 1)tf}{(K + tf)} \cdot \frac{(k_3 + 1) \cdot tf_q}{(k_3 + tf_q)} \log_2 \frac{(r + 0.5) \cdot (N - R - n + r + 0.5)}{(n - r + 0.5) \cdot (R - r + 0.5)} \quad [BM25]$$

The unexpanded $BM25$ matching function, that is when $R = r = 0$, is:

$$\sum_{t \in q} \frac{(k_1 + 1)tf}{(K + tf)} \cdot \frac{(k_3 + 1) \cdot tf_q}{(k_3 + tf_q)} \log_2 \frac{N - n + 0.5}{n + 0.5} \tag{5.18}$$

where

i) $K$ is $k_1((1 - b) + b(\frac{l}{avg\_l}))$.

ii) $k_1$ and $b$ are set by default to 1.2 and 0.75 respectively, $k_3$ to 1000 [88].

By using these default parameters, the unexpanded baseline $BM25$ ranking function, that is the $BM25$ applied in the absence of information about relevance, is:

$$\sum_{t \in q} \frac{2.2 \cdot tf}{0.3 + 0.9\frac{l}{avg\_l} + tf} \cdot \frac{1001 \cdot tf_q}{1000 + tf_q} \log_2 \frac{N - n + 0.5}{n + 0.5} \quad [BM25 - unexp] \tag{5.19}$$

## 5.3 Inference Network Retrieval

Turtle and Croft [116] introduced the use of inference networks to support document retrieval. The Bayesian inference network model is at the basis of the INQUERY system [28]. Information retrieval is viewed as an inference or evidential reasoning process in which the probability of one or more queries is computed with documents as items of "evidence". Bayesian inference networks are used to specify the dependence between queries and documents as a mechanism for propagating and inferring the probabilistic relationship between the query and the document.

The query $Q$ in an Inference Network Retrieval is deemed as a Boolean propositional formula. It is known that any Boolean formula can be equivalently expressed in the disjunctive normal form, which is a disjunction of conjunctions $C_i$ (the constituents) of atomic or negation of atomic formulas (terms) $t_k$. Formally

$$Q = \bigvee_{i \in I} C_i$$

Let us apply the theorem of complete or total probability. For a set of mutually exclusive events $C_i$, with $i \in I$, which is also a covering of the event $Q$, the probability of the event $Q$ is:

$$P(Q) = \sum_{i \in I} P(Q|C_i) \cdot P(C_i) \tag{5.20}$$

As an example, $Q$ can be the conjunction $C$ of terms as defined in the standard vector space model, in the circumstances that $I$ has only one element. $P(Q|C_i)$ may range in the unit interval $[0, 1]$ and its value can be given by the user during the query formulation. The crucial point in the inference network model is how assigning the prior probability $P(C_i)$. At this aim, we can iterate the theorem of complete probability but using the set $D$ of documents as basic space. In order to have $D$ as event space, the documents must be mutually exclusive events. Two arbitrary documents may be indeed regarded to be atomic and thus mutually exclusive events, in symbol $d \cap d' = \emptyset$. In fact if we interpret the logical conjunction of two arbitrary documents as their juxtaposition, then the result of this fusion does not, in general, generate a sensible and consistent text.

Therefore

$$P(C_i) = \sum_j P(C_i|d_j) \cdot P(d_j) \tag{5.21}$$

We substitute the last formula in ( 5.20):

$$P(Q) = \sum_{i \in I} P(Q|C_i) \cdot (\sum_j P(C_i|d_j) \cdot P(d_j)) \tag{5.22}$$

that is:

$$P(Q) = \sum_j \left( \sum_{i \in I} P(Q|C_i) \cdot P(C_i|d_j) \right) \cdot P(d_j) \tag{5.23}$$

With a different computation, we obtain a second relation for $P(Q)$ using the Theorem of complete probability:

$$P(Q) = \sum_j P(Q|d_j) \cdot P(d_j) \tag{5.24}$$

By equations ( 5.23) and ( 5.24) we derive:

$$\sum_j P(Q|d_j) \cdot P(d_j) = \sum_j \sum_{i \in I} P(Q|C_i) \cdot P(C_i|d_j) \cdot P(d_j) \tag{5.25}$$

and therefore:

$$\sum_j P(Q|d_j) \cdot P(d_j) = \sum_j \sum_{i \in I} P(Q|C_i) \cdot P(C_i|d_j) \cdot P(d_j) \tag{5.26}$$

To obtain the probability of a constituent $C$, we suppose the probabilistic term independence. Actually, the *term independence assumption* is extended to the negation of terms:

$$P(C_i|d_j) = \prod_{k \in K} P(t_k|d_j) \cdot \prod_{k \notin K} P(\neg t_k|d_j) \tag{5.27}$$

provided that $C_i = \bigwedge_{k \in K} t_k \wedge \bigwedge_{k \notin K} \neg t_k$, $\neg t_k$ meaning that $t_k$ does not occur in $C_i$ and $P(\neg t_k|d_j) = 1 - P(t_k|d_j)$.

Turtle and Croft use the following formula for assigning the posterior probability $P(t_k|d_j)$

$$P(t_k|d_j) = \gamma + \delta \cdot idfn(t_k) \cdot tfn(t_k, d_j) \tag{5.28}$$

$$P(\neg t_k|d_j) = \delta(1 - idfn(t_k) \cdot tfn(t_k, d_j)) \tag{5.29}$$

where $idfn(t_k)$ is the normalized inverse document frequency

$$idfn(t_k) = \frac{\log \frac{N}{n_{t_k}}}{\log N}$$

$$\gamma + \delta = 1$$

and the normalized term-frequency of a term in a document is

$$tfn(t_k, d_j) = \frac{tf(t_k, d_j)}{\arg_{d \in D} \max tf(t_k, d)}$$

Beyond the idea of combining evidence from multiple sources, inference networks first introduced the use of a nonzero default probability for term–weights. The same feature holds in language modelling as we see in the next sections.

## 5.4 The language model

The language modelling approach was first proposed by Ponte and Croft [82]. The general idea underpinning language modelling in IR is that documents and queries are sequences of words and that the document retrieval score is computed by the probability of producing the query from the document regarded as evidence. More precisely, a document $d$ which is relevant to a query $q$ constitutes a plausible evidence which, among many other pieces of evidence (documents), maximizes the conditional probability $p(q|d)$.

The document is treated as "a model" which should be estimated by the available data, namely the statistics within the document and the statistics of the whole collection.

When estimating the "document model" from data, the model may over–fit the data, that is it may model the possible word sequences which can be extracted in the document, but may not fit other word sequences. Thus, if for example a term does not appear from the document, it will be assigned a zero probability, and similarly all sequences containing this term will have a zero probability, and the absence of the term is particularly likely if the text length of the document is very short. A perfect model for a document would compute unity probability to the document itself (considered as word sequence) and zero probability to all other word sequences. For example, such a perfect model would give zero probability to all word sequences that do not contain words belonging to the vocabulary of the observed document.

In order to overcome the problem of assigning zero probabilities to terms not belonging to the observed document, smoothing probabilities can be used. Smoothing is a method to avoid over-fitting the data, and in Bayesian statistics it is connected with the assignment of the so-called prior distribution. In the Bayesian views it is assumed that

there is some true, but unknown, distribution of probabilities over the events. These probabilities are called the *priors*, or the *a priori probability distribution*, and they are regarded as parameters to be estimated from the *a posteriori* probabilities. Beside the non-empirical probabilities of the priors, the available empirical data provide the most likely probability estimates for the events. Bayes' method supplies a way of combining the two distributions and calculating a posterior or inferred estimate for the events. When, empirical data are large in number an accurate value for the priors is not necessary. However, when the empirical data are not significantly large in number, then Bayes' method requires values for the priors, and sometimes they have to be provided from a subjective starting position. We choose an initial distribution and then we replace it when a better one is determined.

### 5.4.1 Ponte and Croft's model

The first example of the language modelling approach to $IR$ is the model created by Ponte and Croft[82]. The language model offers a uniform approach to both indexing and weighting schemes, while in the standard probabilistic approach these processes use two different models [109]. In language modelling the term "model" has acquired a twofold nature: it can be interpreted as "a probabilistic model" for the empirical data, and can also be used as a "retrieval model". Unlike the 2-Poisson model defining the $BM25$ formula, in which the probability of relevance is a hidden variable, the Ponte and Croft model starts from the "raw" maximum likelihood of terms in the given document as the "model" of the language:

$$p(t|d) = p_d(t) = \frac{tf}{l_d} \tag{5.30}$$

and

$$p(q|d) = \prod_{t \in q} p(t|d) \tag{5.31}$$

The probability of the terms not occurring in the document are then computed by using a default value, that is the raw term-frequency in the collection:

$$p(t|d) = p_D = \frac{F_t}{TotFr_D} \tag{5.32}$$

In order to ensure the fundamental condition $\sum_{t \in V} p(t|d) = 1$ summing up the probabilities of Equations 5.30 and 5.32, a normalization factor is needed. We then assume that the

sum of the two probabilities holds up to a normalization factor.

While the language model based on the Dirichlet Priors directly combines these two probabilities (see next Sections and particularly Section 3.3.6), Ponte and Croft try to make the probability of Equation 5.30 more robust mixing it with a different probability established with a larger estimation, that is considering the set $E_t$ of documents containing the term $t$:

$$\overline{p}(t|d) = \frac{1}{n_t} \sum_{d \in E_t} p(t|d) \tag{5.33}$$

The mixing is obtained by introducing a risk probability function $\widehat{R}$ which is the geometric distribution 2.4.1:

$$\widehat{R}_{t,d} = \left(\frac{1}{1+\bar{f}_{t,d}}\right) \times \left(\frac{\bar{f}_{t,d}}{1+\bar{f}_{t,d}}\right)^{tf} \tag{5.34}$$

where

$$\bar{f}_{t,d} = \overline{p}(t|d) \times l_d.$$

and then:

$$p(q|d) = \begin{cases} p_d(t)^{(1-\widehat{R}_{t,d})} \times \overline{p}(t|d)^{\widehat{R}_{t,d}} & \text{if } tf > 0 \\ p_D & \text{otherwise} \end{cases}$$

The purpose of using the risk function is to enable a choice of probability close to either the value of the maximum likelihood or to the mean frequency of the term, according to the size of the relative term-frequency. If the term-frequency in the document is high then the risk is minimal and the probability of the term can be reduced to its maximum likelihood $p_d(t)$. If $tf$ is small, then the maximum likelihood estimate is less reliable, and in this case, the risk function is high, and the probability of the term reduces mainly to the mean frequency in the set $E_t$ of documents containing the term $t$. Similarly, if the length of the document is very large then the maximum likelihood estimate is more reliable, the risk function accordingly becomes small, and the probability of the term reduces back to the maximum likelihood in the document. Finally, if the term does not occur in the document then the probability of the term is chosen to be the maximum likelihood $p_D$ in the collection $D$.

## 5.5 Language model: Dirichlet's prior for IR

In this section we resume the situation of Section 3.3.6, in which we have several urns containing balls of different colours (terms) and in each urn (document) we extract $l$ balls (tokens). We have derived the fundamental Formula 3.21 assuming Dirichlet's priors:

$$p(p_1,\ldots,p_n,A_1,\ldots,A_n) = \frac{\Gamma(A)}{\Gamma(A_1)\cdots\Gamma(A_n)}p_1^{A_1-1}\cdots p_n^{A_n-1}$$

$$A = \sum_{i=1}^{n} A_i$$

$$\sum_{i=1}^{n} p_i = 1$$

The Formula 3.21 is maximized according to Bayesian statistics when $p_i$ is equal to the mean $\frac{A_i}{A}$. Let us denote the mean by $\lambda_i$.

**Example 2** *Let $D$ be a collection of documents containing $FreqTotColl$ tokens over the vocabulary $V$. Let us count the frequencies $F_{t_i}$ of each term in the collection. Since the sample of documents is significative large, instead of assigning priors with the Dirichlet distribution, we may use the multinomial distribution as defined in Formula B.11 of the Appendix. We instantiate the distribution B.11 with the following parameters:*

$$n_i = F_{t_i} \tag{5.35}$$

*The parameter $n$ is the sum $FreqTotColl$ of all tokens in the collection $FreqTotColl$. The expected relative frequency is $\lambda_i = \frac{F_{t_i}}{FreqTotColl}$.*

**Example 3** *Let us count the frequency $tf_i$ of terms in the document $d$. We instantiate the Formula 3.21 again, giving this time a term the* a priori *probability $\frac{1}{|V|}$ of occurring in a document, where $|V|$ is the total number of terms in the collection.*

$$A_i = tf_i + 1 \tag{5.36}$$

*The parameter $A$ is the sum of the length $l_d$ of the document and $|V|$. The expected relative frequency for all other terms $t_i$ is:*

$$\lambda_i = \frac{tf_i + 1}{l_d + |V|} \tag{5.37}$$

To answer a query $Q = \{t_1, \ldots, t_k\}$ we assume that terms are independent:

$$p(Q|d) = p(t_i, \ldots, t_k) = \prod_{i=1}^{k} \lambda_i = \tag{5.38}$$
$$= \prod_{i=1}^{k} \left( \frac{tf_i + 1}{l_d + |V|} \right)$$

**Experiment 7** [Dirichlet priors]The run on TREC–10 with the retrieval function 5.38 yields an average precision of 0.1262.

**Example 4** *Let us count the frequency $tf_i$ of terms in the document d. We instantiate the Formula 3.21 with the following parameters:*

$$A_i = tf_i + \mu \cdot \frac{F_{t_i}}{FreqTotColl} \tag{5.39}$$

*The parameter A is the sum of the length $l_d$ of the document and the parameter $\mu$. The expected relative frequency for all other terms $t_i$ is:*

$$\lambda_i = \frac{tf_i + \mu \cdot \frac{F_{t_i}}{FreqTotColl}}{l_d + \mu} \tag{5.40}$$

Dirichlet priors created by the formula 5.39 are used by Lafferty and Zhai to define their language model [70, 133]. To answer a query $Q = \{t_1, \ldots, t_k\}$ it is assumed that terms are independent:

$$p(Q|d) = p(t_i, \ldots, t_k) = \prod_{i=1}^{k} \lambda_i =$$
$$= \prod_{i=1}^{k} \left( \frac{tf_i + \mu \cdot \frac{F_{t_i}}{FreqTotColl}}{l_d + \mu} \right) \tag{5.41}$$

We can express Relation 5.41 in additive form applying the monotonic logistic function, but before we divide $p(Q|d)$ by the value $\prod_{i=1}^{k} \mu \cdot \frac{F_{t_i}}{FreqTotColl}$, that does not affect the ranking because it is independent of the document.

$$p(Q|d) \propto \log \frac{\prod_{i=1}^{k} \lambda_i}{\prod_{i=1}^{k} \mu \cdot \frac{F_{t_i}}{FreqTotColl}} =$$
$$= \sum_{i=1}^{k} \log \left( \frac{FreqTotColl \cdot tf_i}{\mu F_{t_i}} + 1 \right) - k \cdot \log(l_d + \mu) \tag{5.42}$$

**Experiment 8** [Dirichlet] In [132] it is reported that the best match with the Dirichlet priors gives an average precision 0.2560 for TREC–8 (short queries) with $\mu = 800$. We ran the same experiment and found the best match for $mu = 400$ with an average precision of 0.2541 in comparison to the precision 0.2600 of our $I(n_e)B2$. For long queries of TREC–8 it is reported in [132] that the best match has an average precision of 0.2600 with $\mu = 2000$ (in comparison to an average precision 0.2841 of our $I(n_e)B2$). However, our experiment with Dirichlet's priors gave a lower average precision (0.1914). We can get the mean average precision to rise to 0.2661 by including the frequency $tfq_i$ of the term in the query into the weighting formula:

$$(5.43)\quad p(Q|d) \propto \sum_{i=1}^{k} tfq_i \cdot \log\left(\frac{FreqTotColl \cdot tf_i}{\mu F_{t_i}} + 1\right) - \sum_{i=1}^{k} tfq_i \cdot \log(l_d + \mu)$$

In general, the Dirichlet model gives a good performance with short queries. For comparisons between the $BM25$ and the Dirichlet priors see Tables 7.10, 7.11 and 7.12. The best performing values are in bold. Dirichlet priors implemented from Formula 5.43 is denoted by the model $LM(\mu = \ldots)$ with $\mu$ chosen as the best performing value for the MAP (Mean Average Precision) for each of the TREC-8, TREC-9 and TREC-10 data.

## 5.6 Language model: mixtures of probability distributions

In Example 2 we have associated a probabilistic model to a single document. It may be possible to combine $r$ possible models of the collection $D$ with a mixture of these probability models. We can even combine the single document models with the model of the collection as explained in Example 2. For each model we thus associate a weight $\mu_i$ to its probability function and we mixture them with a linear combination of all probability functions $p(p_1, \ldots, p_n, A_1^d, \ldots, A_n^d)$. If the weights sum up to 1 then the resulting function is also a probability density function:

$$(5.44)\quad p(p_1, \ldots, p_n) = \sum_{d=1}^{r} \mu_d \cdot p(p_1, \ldots, p_n, A_1^d, \ldots, A_n^d)$$

$$(5.45)\quad \sum_{d=1}^{N} \mu_d = 1$$

The main problem of estimating the weights of the mixture remains. However, since the expectation of the frequency for each term $i$ is a linear operator on the density functions

and the density function of the mixture is a linear combination of density functions then the mean of each $p_i$ is the linear combination of the means:

$$\lambda_i = \sum_{d=1}^{r} \mu_d \cdot \lambda_i^d \tag{5.46}$$

For example, combining two multinomials, the first defined by a document $d$ while the second by the collection $D$ (see example 2), we obtain the following probability for the term $i$:

$$\lambda_i = \mu_d \cdot \frac{tf}{l} + (1-\mu_d) \cdot \frac{F_t}{FreqTotColl} \tag{5.47}$$

This relation was introduced by Hiemstra in [60].

$$p(Q) = p(t_i, \ldots, t_k) = \prod_{i=1}^{k} \lambda_i =$$

$$= \prod_{i=1}^{k} \left( \mu_d \cdot \frac{tf_i}{l} + (1-\mu_d) \cdot \frac{F_{t_i}}{FreqTotColl} \right) = \tag{5.48}$$

$$= (1-\mu_d) \prod_{i=1}^{k} \frac{F_{t_i}}{FreqTotColl} \prod_{i=1}^{k} \left( 1 + \frac{\mu_d}{(1-\mu_d)} \cdot \frac{tf_i \cdot FreqTotColl}{l \cdot F_{t_i}} \right) \tag{5.49}$$

The parameter $\mu_d$ can be set to a constant $\mu$, that is we assume that $\mu_d$ is independent of both the document and the term. The first product in Formula 5.48 is therefore a common factor of the score of each document and then:

$$p(Q) \propto \prod_{i=1}^{k} \left( 1 + \frac{\mu}{(1-\mu)} \cdot \frac{tf_i \cdot FreqTotColl}{l \cdot F_{t_i}} \right) \tag{5.50}$$

**Experiment 9** [Mixture] We ran an experiment with TREC–10 data using the parameter $\mu_d = 0.15$ as suggested by Hiemstra for the TREC–8 collection and we obtained the average precision of 0.1201[59]. Varying the parameter $\mu_d$ we obtained the best match with $\mu_d = 0.75$ and average precision 0.1465 for TREC–10 data. It is easier to implement the logistic version of Formula 5.50, because the terms which belong to the query but not to the document do not contribute to the following sum:

$$p(Q) \propto \sum_{i=1}^{k} \log_2 \left( 1 + \frac{\mu}{(1-\mu)} \cdot \frac{tf_i \cdot FreqTotColl}{l \cdot F_{t_i}} \right) \tag{5.51}$$

# Chapter 6

# Term-frequency normalization

In this chapter we study the probability distribution $P(tf, l)$ of two random variables over the elite set $E_t$, which is called a *bivariate discrete distribution* [67], where $tf$ is the within–document term-frequency and $l_d$ the length of the document $d$. An observation from sampling from a bivariate population consists of a pair of measurements. For a bivariate distribution the correlation coefficient is a useful function to measure the degree to which the two variables vary together. If the bivariate distribution is normal then sampling shows a linear relationship between the two variables [31].

The *correlation coefficient* is the covariance of the normalized variables of $tf$ and $l_d$, that is

$$\upsilon = E\left(\frac{tf - \overline{tf}}{\sigma_{tf}} \cdot \frac{l - \bar{l}}{\sigma_l}\right)$$

where $E$ is the expectation, $\overline{tf} = \dfrac{F_t}{n_t}$ is the mean term-frequency in the elite set of the term, $\bar{l}$ is the average length in the elite set, $\sigma_l^2$ and $\sigma_{tf}^2$ are the variance of the length and the variance of the term-frequency in the elite set respectively.

The value $-1 \leq \upsilon \leq 1$ indicates the degree of the linear dependence between the two random variables. When $\upsilon = 0$ the correlation coefficient indicates that the two random variables are independent. When $tf = a \cdot l_d + b$ for some $a, b$ [31], the correlation coefficient is $-1$ or $1$. For this reason the value $\upsilon$ provides a measure of the extent to which the two variables are linearly related.

Harter was not in a position to draw a general relationship between $tf$ and $l_d$, because the correlation coefficient gave ambiguous results. His experiment however dealt with a

sample of highly informative terms from a small text collection of technical abstracts. The interval of confidence values for the correlation coefficient was around $\upsilon = 0$ for some words, positive or negative for others.

Harter concluded that the relationship between term-frequency and length in a document is not obvious.

We ran a similar experiment with the TREC collection w2Tg of 2GB using as sample of terms the set of all terms of the queries of TREC-7 and TREC–8, and with the TREC collection WT10g of 10GB using as sample of terms the set of terms of the queries of TREC–10. An excerpt of the results are shown in Table 6.1.

We used Fisher's method [39] to derive the confidence interval of the correlation coefficient. Fisher's method consists in deriving the confidence interval of the transform

$$Z = 0.5 \ln \frac{1+\upsilon}{1-\upsilon}$$

which is normally distributed with mean

$$0.5 \ln \frac{1+\overline{\upsilon}}{1-\overline{\upsilon}}$$

and standard deviation

$$\frac{1}{\sqrt{n-3}}$$

regardless the value of $\upsilon$.

The size of the set of documents from which we get the correlation factor is in general very large with the exception of a few cases, such as the stemmed word "postmenopaus" as shown in Table 6.1 .

Our findings do not coincide with Harter's conclusions. A positive correlation can be established between term-frequency and words. *Although the value of the correlation coefficient is relatively small, small values of the correlation factor are regarded very meaningful in large samples* [111]. In our sample terms appear in many thousand of documents so that the samples can be considered very large. The situation is different when a small value of the correlation coefficient is observed using very small samples. In such cases, we can hardly conclude something, since the results should be considered neither meaningful nor statistically significant [111]. It is worth noticing that the most frequent terms in the collection,which are the terms with the largest test samples, are mainly those which possess the greatest correlation coefficient (see Table 6.2).

| Query | Term | $n_t$ | Confidence interval of $\upsilon$ (95%) | |
|---|---|---|---|---|
| 351 | explor | 10012 | 0.266 | 0.266 |
| 351 | falkland | 400 | -0.057 | -0.047 |
| 351 | petroleum | 8245 | 0.289 | 0.289 |
| 352 | british | 42153 | 0.132 | 0.132 |
| 352 | chunnel | 33 | 0.034 | 0.157 |
| 352 | impact | 29197 | 0.340 | 0.340 |
| 353 | antarctica | 207 | 0.051 | 0.070 |
| 353 | explor | 10012 | 0.266 | 0.266 |
| 354 | journalist | 10395 | 0.166 | 0.166 |
| 354 | risk | 29230 | 0.294 | 0.294 |
| 355 | ocean | 6343 | 0.723 | 0.723 |
| 355 | remot | 5049 | 0.439 | 0.439 |
| 355 | sens | 21424 | 0.338 | 0.338 |
| 356 | britain | 24919 | 0.106 | 0.106 |
| 356 | estrogen | 61 | -0.079 | -0.013 |
| 356 | postmenopaus | 7 | -0.389 | 0.319 |
| 357 | disput | 17882 | 0.122 | 0.122 |
| 357 | territori | 21765 | 0.378 | 0.378 |
| 357 | water | 31578 | 0.330 | 0.330 |
| 358 | alcohol | 5049 | 0.042 | 0.042 |
| 358 | blood | 8010 | 0.205 | 0.205 |
| 358 | fatal | 3888 | 0.298 | 0.298 |
| 359 | fund | 66160 | 0.517 | 0.517 |
| 359 | mutual | 12482 | 0.212 | 0.212 |
| 359 | predictor | 98 | -0.096 | -0.056 |
| 360 | benefit | 41067 | 0.249 | 0.249 |
| 360 | drug | 20757 | 0.126 | 0.126 |
| 360 | legal | 35355 | 0.280 | 0.280 |
| 361 | cloth | 9704 | 0.093 | 0.093 |
| 361 | sweatshop | 93 | 0.374 | 0.410 |
| 362 | human | 26099 | 0.286 | 0.286 |
| 362 | smuggl | 2996 | 0.069 | 0.071 |

Table 6.1: The Correlation coefficient between length and term-frequency with terms of the first 12 queries of TREC -7

| Term | $n_t$ | $v$ | Term | $n_t$ | $v$ | Term | $n_t$ | $v$ |
|---|---|---|---|---|---|---|---|---|
| ocean | 6343 | 0.72 | treatment | 15587 | 0.24 | mainstream | 2757 | 0.06 |
| equip | 33896 | 0.63 | sick | 3660 | 0.22 | rain | 5166 | 0.06 |
| us | 46905 | 0.61 | mutual | 12482 | 0.21 | dismantl | 3438 | 0.05 |
| organ | 57608 | 0.55 | blood | 8010 | 0.21 | mental | 5974 | 0.05 |
| fund | 66160 | 0.52 | hydrogen | 817 | 0.20 | nativ | 6199 | 0.05 |
| insur | 24625 | 0.46 | medic | 17632 | 0.18 | prize | 5217 | 0.05 |
| radioact | 2286 | 0.44 | europ | 54138 | 0.18 | alcohol | 5049 | 0.04 |
| remot | 5049 | 0.44 | price | 78637 | 0.18 | merci | 1553 | 0.04 |
| oceanograph | 128 | 0.43 | food | 27841 | 0.17 | nino | 153 | 0.03 |
| transport | 31436 | 0.41 | journalist | 10395 | 0.17 | winner | 9711 | 0.03 |
| disast | 7073 | 0.40 | children | 23549 | 0.16 | piraci | 406 | 0.03 |
| sweatshop | 93 | 0.39 | court | 39586 | 0.16 | orphan | 614 | 0.03 |
| territori | 21765 | 0.38 | hybrid | 1063 | 0.15 | syndrom | 1221 | 0.02 |
| el | 12194 | 0.37 | illeg | 12194 | 0.15 | rabi | 128 | 0.01 |
| wast | 13262 | 0.36 | home | 60333 | 0.14 | euro | 4017 | 0.01 |
| technolog | 33280 | 0.35 | smoke | 5407 | 0.14 | holist | 118 | 0.00 |
| build | 65160 | 0.35 | british | 42153 | 0.13 | vitro | 268 | 0.00 |
| world | 101523 | 0.35 | robot | 790 | 0.13 | cigar | 503 | -0.01 |
| impact | 29197 | 0.34 | cyanid | 223 | 0.13 | bulimia | 25 | -0.02 |
| sens | 21424 | 0.34 | medicin | 6338 | 0.13 | nobel | 1035 | -0.02 |
| enhanc | 12777 | 0.34 | drug | 20757 | 0.13 | postmenopaus | 7 | -0.04 |
| automobil | 4760 | 0.34 | disput | 17882 | 0.12 | estrogen | 61 | -0.05 |
| water | 31578 | 0.33 | opposi | 24231 | 0.12 | falkland | 400 | -0.05 |
| law | 58831 | 0.32 | amazon | 512 | 0.12 | predictor | 98 | -0.08 |
| export | 27950 | 0.31 | casino | 1171 | 0.11 | anorexia | 41 | -0.09 |
| altern | 23917 | 0.31 | space | 17734 | 0.11 | nervosa | 11 | -0.31 |
| commerci | 38903 | 0.31 | britain | 24919 | 0.11 | | | |
| fatal | 3888 | 0.30 | encryp | 84 | 0.10 | | | |
| energi | 26336 | 0.30 | ill | 11023 | 0.10 | | | |
| soil | 4482 | 0.30 | teach | 7261 | 0.10 | | | |
| risk | 29230 | 0.29 | chunnel | 33 | 0.10 | | | |
| petroleum | 8245 | 0.29 | recal | 12474 | 0.10 | | | |
| human | 26099 | 0.29 | cloth | 9704 | 0.09 | | | |
| transfer | 22539 | 0.28 | tourism | 4245 | 0.09 | | | |
| legal | 35355 | 0.28 | tunnel | 3152 | 0.09 | | | |
| fuel | 15183 | 0.27 | car | 30309 | 0.09 | | | |
| american | 53460 | 0.27 | obes | 191 | 0.09 | | | |
| forest | 6656 | 0.27 | disabl | 5152 | 0.09 | | | |
| explor | 10012 | 0.27 | kill | 25557 | 0.08 | | | |
| fertil | 2636 | 0.26 | school | 36008 | 0.07 | | | |
| health | 32856 | 0.25 | smuggl | 2996 | 0.07 | | | |
| benefit | 41067 | 0.25 | moon | 1612 | 0.07 | | | |
| vessel | 5424 | 0.25 | arsen | 1789 | 0.06 | | | |
| station | 25565 | 0.25 | antarctica | 207 | 0.06 | | | |

Table 6.2: Terms of TREC -7 in decreasing ordering of term-frequency–document length correlation.

On the other hand, for very rare terms independence or negative correlation can be observed. We thus agree with Harter that highly technical terms, such as specific terms from chemistry, may tend to occur independently from the length of the document.

We have found an average correlation of 0.186 for the queries of TREC-7, 0.147 for the queries of TREC-8 and 0.146 for the queries of TREC-10.

In conclusion, our findings enforce our intuition that term-frequency normalization is an important component of the retrieval model. If we had not found a positive correlation between length and term-frequencies we would have found difficult to motivate our normalization functions.

Indeed a term-frequency normalization function substitutes the actual term-frequency $tf$ for the new term-frequency value $tfn$ computed on the basis of the document length. A positive correlation factor means that the longer the document is, the bigger the term-frequency is.

After this discussion, we are now in position to formulate the problem of the *Term-Frequency Normalization.* It is the process of predicting the number $tfn$ of occurrences a term would have in a document if this document were of a standard length $\Delta l$. Once the law is determined, in practice, we will substitute $tfn$ for each occurrence of $tf$ in our model generating Formula 1.1.

When a new term-frequency normalization function is defined, a new value of the correlation coefficient is established

$$v = E\left(\frac{tfn - \overline{tfn}}{\sigma_{tfn}} \cdot \frac{l - avg\_l}{\sigma_l}\right) = E\left(\frac{tfn \cdot l}{\sigma_{tfn} \cdot \sigma_l}\right) - \frac{avg\_l \cdot \overline{tfn}}{\sigma_{tfn} \cdot \sigma_l}$$

It is natural to test first the linear dependence assumption. It is easy to instantiate the parameters $a$ and $b$ in such a case. From

$$\sum_d tf = F_t = a \cdot \sum_d l_d + b \cdot N$$

linearity is when $a = \frac{F_t}{TotFr_{E_t}}$ and $b = 0$, which holds when terms are distributed uniformly over their elite set, and we observed from our experiment that this is likely to happen when elite sets are large.

A linear correlation can also be assumed to hold among different pieces of an homogeneous text. For example, we can split any document into a number of fragments of

different length. The existence of a linear correlation becomes equivalent to the assumption of a uniform distribution of the frequencies within single documents.

**H1** The distribution of a term is uniform in the document (see a further development of this Hypothesis **H1** on page 127).

However, it is a matter of fact that the term-frequency normalization based on a logarithmic relationship between term-frequency and length provides a better performance than the uniform distribution **H1** in all our experimental results.

**H2** The relative term-frequency $\frac{tf}{l}$ within the document is a decreasing function of the length $l$ (see a further development of this Hypothesis **H2** on page 127).

Another issue is the connection of the variation of document length to the relevance. This problem is called *the length normalization problem.* We already observed in Section 5.1 that the cosine matching function of the vector space model produces an implicit document length normalization. The emphasis of some works [104, 103, 97] has been on the advantage that long documents have to be retrieved in the vector space model over the short ones. The term-frequency normalization, or the length normalization, has been thus thought as the empirical problem of penalizing the term-weights in long documents.

According to Singhal, Buckley and Mitra [103, 105] however the cosine normalization of the vector space model penalizes too much long documents. This conclusion was drawn by comparing the retrieval curve to the relevance curve against document length. The comparison shows that the relevance curve was above the retrieval curve for long documents, but it was below for short documents. These experiments indicate that the length of retrieved documents is related to the relevance. According to Singhal, Buckley and Mitra a good score function should retrieve documents of different lengths, but their chance of being retrieved should be also similar to their likelihood of relevance. But the drawback of a relevance-based term-frequency normalization would be the introduction of unknown parameters to be estimated with the relevance data. In addition, relevance depends on the type of retrieval task and task requirements can modify the typical length of the document. For example, short documents should be preferred to long ones in topic distillation, whose task is to select the pages containing the main WEB resources on a topic.

The Mandelbrot-Paretian-Zipf law is the third proposal explored and studied in this dissertation (see Section 6.3) and, the term-frequency normalization based on language model is the last proposal. We have chosen Dirichlet's priors as representative of the class of language models (see Section 6.4), but any other language model could have been equally used.

We have seen that when the sample is very large a Mandelbrot-Pareto-Zipf law exists between the size of the corpus and the term frequencies. Potential applications of the rank-frequency law have been hardly explored in IR. Aalbersberg substituted the ranks for frequencies in the term-weights of the vector space model and showed that the performance is still comparable with that of the standard vector space model [1]. Blair instead used the Zipf law to measure the effectiveness of retrieval systems[12].

In Section 6.3 we show how to explore the Mandelbrot-Pareto-Zipf law and obtain from it a *parameter free* term-weighting function for IR. This model of normalization gives results that are quite robust and in general superior to the linear correlation, suggested by Harter (see [52, page 23] and page 35). The main content of this Chapter is published in [6, 5]. Last investigation concerns the use of probabilities as they are assigned in the language modelling approach. We employ Dirichlet's priors to normalize the term-frequency within a document. The application from language modelling to the divergence from randomness models is straightforward. Any probability $p(t|d)$ based on the language model can be transferred as it is into the term-frequency normalization component. It is sufficient to observe that the expected number of tokens of a term $t$ in a document is $p(t|d) \cdot l$. When we compare the term tokens to a standard length $\Delta l$, the term-frequency normalization of a language model is

$$tfn = p(t|d) \cdot \Delta l$$

However, the hypotheses **H2** seems to be the most robust and effective proposal for term-frequency normalization.

## 6.1 Related works on term-frequency normalization

Comparing the document length to the average document length has been shown to enhance the effectiveness of IR systems. For example, the $BM25$ matching function of Okapi has an implicit form of normalization:

$$\sum_{t \in Q} \frac{(k_1+1)tf}{(K+tf)} \cdot \frac{(k_3+1)\cdot qtf}{(k_3+qtf)} \log_2 \frac{(r+0.5)(N-n-R+r+0.5)}{(R-r+0.5)(n-r+0.5)} \tag{6.1}$$

where

- $R$ is the number of documents known to be relevant to a specific topic,
- $r$ is the number of relevant documents containing the term,
- $qtf$ is the frequency of the term within the topic from which $Q$ was derived
- $l$ and $avg\_l$ are respectively the document length and average document length.
- $K$ is $k_1((1-b)+b(\frac{l}{avg\_l}))$,
- $k1$, $b$ and $k_3$ are parameters which depend on the nature of the queries and possibly on the database;
- $k_1$ and $b$ are set by default to 1.2 and 0.75 respectively, $k_3$ is often set to 1000 (effectively infinite). In TREC 4, [88] $k_1$ was in the range $1 \leq k_1 \leq 2$ and $b$ in the interval $0.6 \leq b \leq 0.75$ respectively.

Using the default parameters above ($k_1 = 1.2$ and $b = 0.75$), the unexpanded BM25 ranking function is defined from Equation 5.9 without using the information about relevance, that is setting $R = r = 0$.

$$\sum_{t \in Q} \frac{2.2 \cdot tf}{0.3+0.9\frac{l}{avg\_l}+tf} \cdot \frac{1001 \cdot qtf}{1000+qtf} \log_2 \frac{N-n+0.5}{n+0.5}$$

An evolution of the INQUERY ranking formula [2] uses the same normalization factor as the unexpanded BM25 with $k_1 = 2$ and $b = 0.75$, and $qtf = 1$:

$$\frac{tf}{tf+0.5+1.5\frac{l}{avg\_l}} \cdot \frac{\log_2 \frac{N+0.5}{n}}{log_2(N+1)} \tag{6.2}$$

Hence, the BM25 length normalization $tfn$ can be thought as the product

$$tfn = T \cdot tf \tag{6.3}$$

where $T$ is:

$$T = \frac{1}{tf + 0.3 + 0.9 \cdot \frac{l}{avg\_l}} \tag{6.4}$$

Our experiments will show that this normalization can be taken as a simple, powerful and robust type of normalization of $tf$. We will demonstrate that the BM25 length normalization is strictly related to the Equation 4.19.

Indeed,

$$tf + 0.3 + 0.9 \cdot \frac{avg\_l}{l} = tf + k_1$$

with $k_1 = 1.2$ when $l = avg\_l$.

If we use the normalization factor $T$ in Equation 6.4 as a combined way to perform both the normalization gain and the term-frequency normalization we may use it to generalize the BM25 with our basic models of divergence from randomness as follows:

$$weight(t,d) = T \cdot Inf_1(tf) \tag{6.5}$$

Moreover, if the basic model $Inf_1(tf)$ in Equation 6.5 is $I(n)$ or $I(F)$ of Equations 4.13 or 4.16 as shown in Table 1.2, then from the normalization $T$ we obtain the randomness model given in Equation 4.19 up to the parameter $k_1$. The formal derivation of the unexpanded BM25 formula is given in Section 7.1.

## 6.2 Term-frequency normalizations H1 and H2

Our next concern is to introduce suitable functions able to normalize the random variables $tf$ to a given length of document. In other words, we would like to obtain the expected number of tokens of a term in a document as if the lengths of the documents were all equal to a fixed value, for example to their average length.

The probabilistic models of randomness are based on the *term independence assumption*. We assume that an occurrence of a term cannot be conditioned by the presence of

other tokens in the observed text. According to the formal model, the length of a document should be the sum of finitely many independent random variables. However, when tokens of the same term occur densely within a portion of text it is possible to detect dependence. All terms which co-occur more often over the collection or within single documents are related. This dependence also extends to the occurrences of the same word. Indeed, the divergence from randomness measures such a dependence. Although we can explain and measure how *improbable* the density is by chance, the same models do not give us any insight to derive an expected document length.

It is difficult to express how improbable it is for us to obtain a specific length of observed document or why it should have that length. The comparison of $tf$ tokens in a document of length $l_1$ to $tf$ tokens in a document of length $l_2$ is not yet possible in a framework based on the *term independence assumption.*

We make some alternative hypotheses on how to compare different term frequencies and test them with a Bayesian methodology choosing the hypothesis which is best from an empirical point of view.

We make four assumptions on how to resize term-frequencies according to the length of the documents and we evaluate them. The first assumption is similar to the "verbosity hypothesis" of Robertson [87], which states that the distribution of term-frequencies in a document of length $l$ is a 2-Poisson with means $\lambda \cdot \frac{l}{avg\_l}$ and $\mu \cdot \frac{l}{avg\_l}$, where $\lambda$ and $\mu$ are the original means related to the observed term (as discussed in the Introduction) and $avg\_l$ is the average length of documents.

We first define a density function $\rho(l)$ of the term-frequency. Then, for each document $d$ of length $l(d)$ we compute the term-frequency on the interval $[l(d), l(d) + \Delta l]$ of given length $\Delta l$. We take this value as the normalized term-frequency. The magnitude $\Delta l$ of the interval is a crucial choice. It can be either the median, the mean $avg\_l$ of the distribution or their multiples. The mean minimizes the mean squared error function

$$\frac{1}{N}\sum_{i=1}^{N}(\Delta l - l(d))^2$$

while the median minimizes the mean absolute error function

$$\frac{1}{N}\sum_{i=1}^{N}(\Delta l - l(d))$$

**H1** The distribution of a term is uniform in the document. The term-frequency density $\rho(l)$ is a constant $\rho$

$$\rho(l) = c \cdot \frac{tf}{l} = \rho \tag{6.6}$$

where $c$ is a constant.

**H2** The term-frequency density $\rho(l)$ is a decreasing function of the length $l$.

We start with these two assumptions **H1** and **H2** on the density $\rho(l)$ but other choices are equally possible.

Experiments shows that the normalization with $\Delta l = avg\_l$ is the most appropriate choice for long queries (see Chapter 7). For short queries $\Delta l = c \cdot avg\_l$ with $c = 7$ is the most effective value. However Figures 6.2 and 6.3 show a great stability in performance with a very large interval of values of $c$.

According to hypothesis **H1** the *normalized term-frequency* $tfn$ is:

$$tfn = \int_{l(d)}^{l(d)+c\cdot avg\_l} \rho(l)dl = \rho \cdot c \cdot avg\_l = c \cdot tf \cdot \frac{avg\_l}{l(d)} \tag{6.7}$$

whilst, according to the hypothesis **H2**

$$tfn = \int_{l(d)}^{l(d)+c\cdot avg\_l} \rho(l)dl = c \cdot tf \cdot \int_{l(d)}^{l(d)+c\cdot avg\_l} \frac{dl}{l} = tf \cdot \ln\left(1 + \frac{c \cdot avg\_l}{l(d)}\right) \tag{6.8}$$

To determine the value of the constant $c$ in **H1** when the effective length of the document coincides with the average length, $l(d) = avg\_l$, we assume that the normalized term-frequency $tfn$ is equal to $tf$.

Therefore, the constant $c$ is 1 assuming **H1**.

$$tfn = tf \cdot \frac{avg\_l}{l(d)} \quad \textbf{[H1]} \tag{6.9}$$

$$tfn = tf \cdot \ln\left(1 + \frac{c \cdot avg\_l}{l(d)}\right) \quad \textbf{[H2]} \tag{6.10}$$

We substitute uniformly $tfn$ of Equations 6.9 or 6.10 for $tf$ in $weight(t, d)$ of Equations 4.19 and 4.23.

Note that **H1** is an approximation of **H2** when $l$ is large:

$$tfn \cdot l = tf \cdot l \cdot \ln\left(1 + \frac{\Delta l}{l}\right) = tf \cdot \ln\left(1 + \frac{\Delta l}{l}\right)^{l} \sim tf \cdot \ln e^{\Delta l} = tf \cdot \Delta l$$

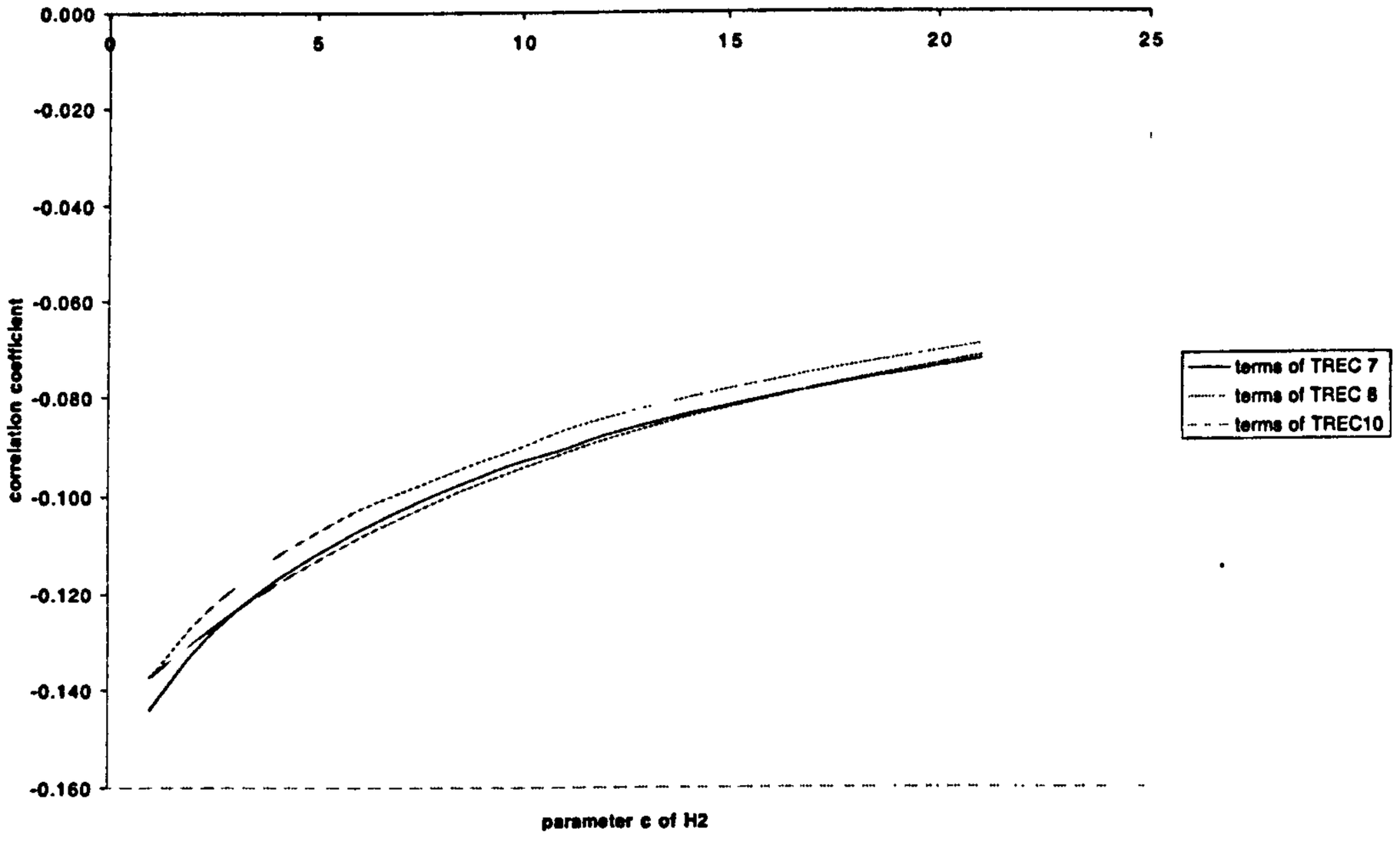


Figure 6.1: The average correlation coefficient between the document length and the term-frequencies normalized by **H2** of Formula 6.10. The sample of terms are from the queries of TREC–7, TREC–8 and TREC–10 respectively.

### 6.2.1 A discussion on the Second Normalization H2

We have observed that, with the exception of the vector space model, which possesses an implicit mechanism of length normalization, all models of IR, such as language models and probabilistic models, have parameters which need to be estimated. Best Match method is the easiest way to determine the optimal values of the parameters for a given collection. We observed from the experiments (see Tables 7.11, 7.10 and 7.12 and Table 8.4) that the parameter $\mu$ of the IR model with Dirichlet's priors depends both on the collection and on the length of the query. It is quite difficult to predict for an arbitrary collection an optimal value of the parameter of the language model. On the contrary the $BM25$ has its parameters quite stable for all collections, but it performs in general worse than language models. The optimal values for the parameters of the $BM25$ are close to the values which we formally derived from the model $I(n)L2$ when the parameter $c$ of **H2** is set to 1. This setting of the parameter $c$ corresponds to the normalization with the average length of the documents. However, we see that for short or moderately long queries the optimal values are located after $c = 1$. As shown in

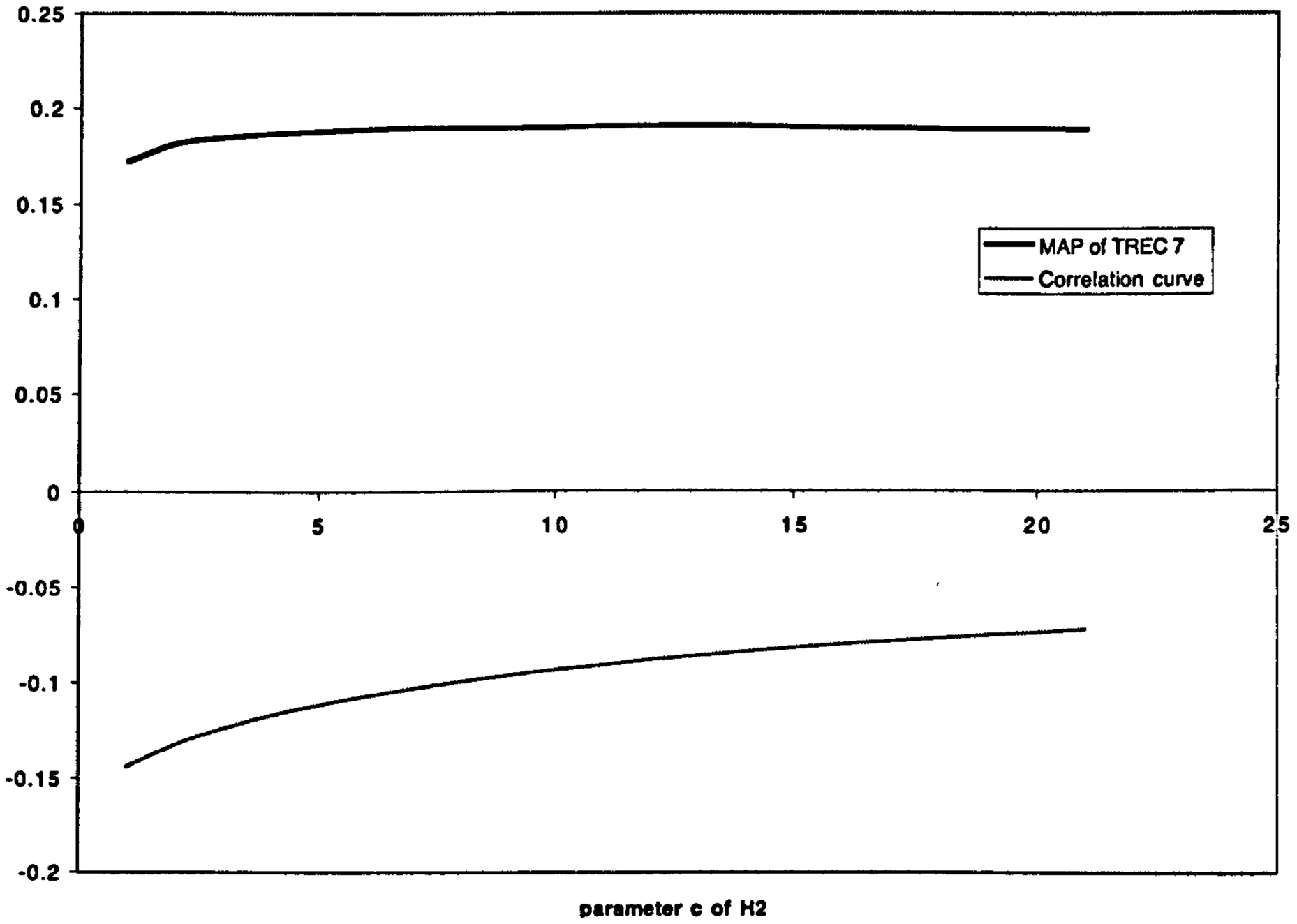


Figure 6.2: Comparison of the the correlation coefficient as in Figure 6.1 to the performance. The model is $I(n_e)B2$. The best matching value of MAP for TREC 7 data is 0.1904 at $c = 13$. Best Pr@10 is 0.4400 at $c = 8$.

Figures 6.2 and 6.3 the size of the interval $\Delta l = c \cdot avg\_l$ is relatively important since performance is almost constant for a relatively large interval of values of $c$. Therefore, the parameter $c$ can be set to any arbitrary value greater than 1 (around 7 is the best). For long queries, however, the best matching value of $c$ tends to converge to 1.

Although, Normalization **H2** is not parameter free, however the introduction of the parameter $c$ has been theoretically motivated and also its optimal matching values lie in a large interval of values. In addition for actual queries, that is when users submit a short query and the query expansion mechanism is activated, the normalization **H2** comes out to be very stable and robust.

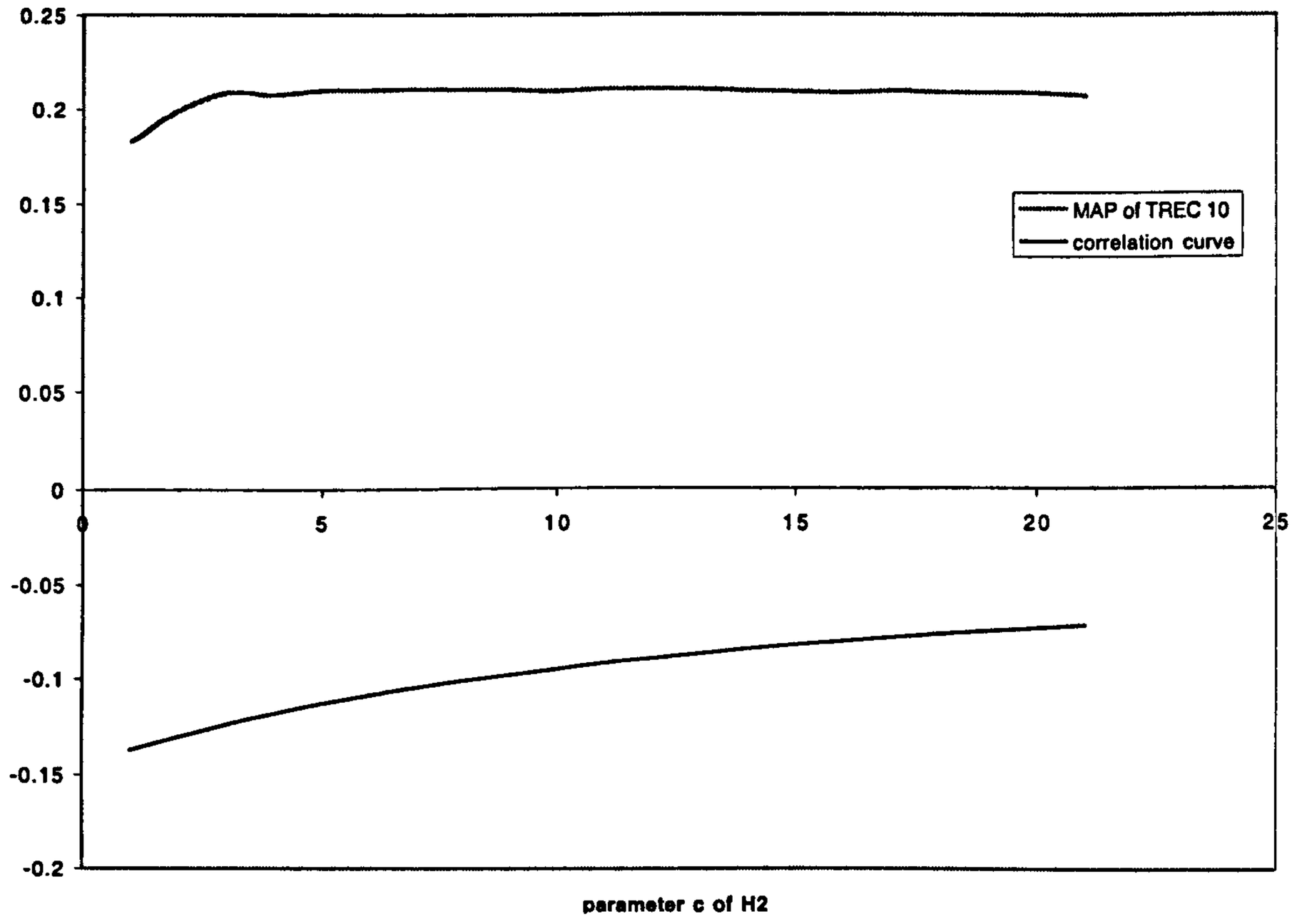


Figure 6.3: Comparison of the the average correlation coefficient as in Figure 6.1 to the performance. The model is $I(n_e)B2$. The best matching value of MAP for TREC 10 data is 0.2107 at $c = 12$. Best Pr@10 is 0.3720 at $c = 7$.

## 6.3 Term-frequency normalization based on the classical Pareto distribution

In Section 2.1 we have seen that an homogenous piece of text, like a document or a book of length $l$, can be seen as a sample of the population. The sample space is the product space $V^l$ and the probability distribution is the number of possible outcome configurations satisfying the empirical data. In the case of independent and equiprobable trials the process is Bernoullian with a probability distribution given by Equation 2.1. Let $l$ be the length of the text sample and let $V$ be the set of all terms contained in the text sample.

Mandelbrot derived several relationships among the vocabulary size, the text length and term-frequencies $F_t$, using a Estoup–Zipf like law [74]. According to Mandelbrot the

rank-frequency relationship is

$$p = BC^B(rank + C)^{-(B+1)} \text{ with } B > 0$$

where $C$ and $B$ are two unknown parameters. According to Mandelbrot $-\log_2 p$ is the "cost" of transmitting the signal with frequency $p$ in some optimal binary code. Indeed in our implementation of the direct file which is needed to implement the query expansion module we have used the $\delta$-code of the rank of the word with frequency $p$ [72, 126] as optimal code for encoding the words. This cost is roughly $-\log_2 p$, following the rank-frequency relationship above.

We now derive Mandelbrot's relationships using the Feller-Pareto distributions. The proof is similar to that used by Mandelbrot [74]. We assume that all terms $t_r$ are ordered by the decreasing ordering of their occurrence frequencies $F$, that is:

$$F_1 > F_2 > \ldots > F_V$$

The probability that the term $t_r$ occurs $F$ times is according to the binomial law 2.1:

$$p(F|l) = \binom{l}{F} p_r^F q_r^{l-F} \quad \text{where } p_r = P(X = r) \tag{6.11}$$

In Section 2.5.1 we have derived a relation between frequencies and number of outcomes of the sample space that possess the same frequencies. In other words the alternative outcomes of $V$ having the same rank in the occurrence ordering distribute according to the Paretian distribution. Let us thus assume that the classical Pareto density function of Equation 2.37 specifies the value of the probability $p = p(r)$.

$$p = p(X = r) = \frac{\alpha}{\sigma}\left(\frac{r}{\sigma}\right)^{-(\alpha+1)} \quad \text{with } r > \sigma \text{ and } \alpha > 0 \tag{6.12}$$

We take $\sigma$ as the first rank in the ordering for which the Paretian law begins to hold. We recall that $\alpha > 0$ and it is the parameter inherited by one of the two parameters of the Beta distribution. Let

$$A = (\alpha + 1)^{-1} \tag{6.13}$$

Notice that $\alpha = A^{-1} - 1 = \frac{1 - A}{A}$.

Since $\alpha > 0, 0 < A < 1$ and $A$ is a monotonically decreasing function of the parameter $\alpha$:

$$\lim_{\alpha \to 0} A = 1 \tag{6.14}$$

$$\lim_{\alpha \to +\infty} A = 0 \tag{6.15}$$

Extracting $r$ from Equation 6.12

$$r = \sigma \left( p \frac{\sigma A}{1 - A} \right)^{-A} \tag{6.16}$$

Hence deriving the function $r = r(p)$ with respect to the probability $p$:

$$\frac{dr}{dp} = -\frac{(\sigma A)^{1-A}}{(1 - A)^{-A}} \cdot p^{-(1+A)} \tag{6.17}$$

The rank $r$ should range from 0 to $V$, but we may assume that $r$ is continuous and ranges from 0 to $\infty$, since we here use the Paretian distribution which is the continuous analogue of Zipf's law.

The number of terms $n(F|l)$ that occur $F$ times in the text is given by the sum of $p(F|l, t_r)$ over all terms $t_r$. This number is equivalent to the integral of the function in Equation 6.11 with respect to the variable $r$:

$$n(F|l) = \int_0^{\infty} \binom{l}{F} p^F q^{l-F} dr \tag{6.18}$$

Let

$$C_0 = (\sigma A)^{1-A}(1 - A)^A \tag{6.19}$$

then the derivative of $r$ in Equation 6.17 can be rewritten as

$$\frac{dr}{dp} = C_0 \cdot p^{-(1+A)} \tag{6.20}$$

Substituting the derivative of $r$ of Equation 6.20 in the integral of Equation 6.18

$$n(F|l) = C_0 \int_0^1 \binom{l}{F} p^{F-A-1} q^{l-F} dp \tag{6.21}$$

Exploiting Relation B.7 of the Beta distribution with parameters $F - A$ and $l - F + 1$ and Relation B.3, i.e. $\Gamma(l - F + 1) = (l - F)!$,

$$\begin{aligned} n(F|l) = \; & C_0 \begin{pmatrix} l \\ F \end{pmatrix} \frac{(l-F)!\Gamma(F-A)}{\Gamma(l-A+1)} \\ = \; & C_0 \frac{l!\Gamma(F-A)}{F!\Gamma(l-A+1)} \end{aligned}$$

We now find an approximation of the ratio $C(a,n) = \dfrac{\Gamma(n+a)}{n!}$ assuming that $n$ is large. From the Stirling Formula 2.10 and the approximation of the Gamma function of Equation B.4, the ratio $C(a,n)$ is equivalent to

$$\frac{\sqrt{2\pi}e^{-n-a}(n+a)^{n+a-0.5}}{\sqrt{2\pi}\cdot e^{-n}n^{n+0.5}} =$$

$$= \frac{e^{-a}(n+a)^{n+a-0.5}}{n^{n+a-0.5}n^{-(a-1)}} =$$

$$= e^{-a}\left(1+\frac{a}{n}\right)^{n+a-0.5} n^{a-1} =$$

For $n$ large $\left(1+\dfrac{a}{n}\right)^{n+a-0.5} \sim e^a$ which implies

$$C(a,n) \sim n^{a-1} \tag{6.22}$$

Substituting the approximation of Equation 6.22 in $n(F|l)$

$$\begin{aligned} n(F|l) = \; & C_0 \frac{C(-A,F)}{C(-1+A,l)} \\ n(F|l) \sim \; & C_0 \frac{l^A}{F^{A+1}} \qquad \text{with } l \text{ and } F \text{ large} \end{aligned} \tag{6.23}$$

Equation 6.23 is the same relationship determined by Mandelbrot [74, see Appendix A] up to a proportional factor.

### 6.3.1 The relationship between the vocabulary and the text length

In the last section we derived the number $n(F|l)$ of different terms of the vocabulary $V$ having a given term-frequency $F$ in an arbitrary text of length $l$, based on the hypothesis that the distribution of the terms in $V$, classified according their frequencies, follows the Paretian law.

We want now to compute the number of unique terms in the text, that is the size of the vocabulary $V$. The derivation is similar to that obtained for $n(F|l)$. First we note that the number of terms occurring in the text can be obtained by summing the probabilities of occurrence over all terms. The probability of occurrence of an arbitrary term is $1 - q^l$ as defined in Equation 6.11.

$$V(l) = \sum_r (1 - q^l) = \int_0^{+\infty} (1 - q^l) dr$$

Since

$$1 - q^l = (1 - q) \sum_{x=0}^{l-1} q^x = \sum_{x=0}^{l-1} pq^x$$

we first approximate for every integer $x$, with $0 \leq x < l$

$$\int_0^{+\infty} pq^x dr$$

With the same steps as in the derivation of Equation 6.21 from the binomial law 6.18, we similarly derive

$$\begin{aligned} (6.24) \quad \int_0^{+\infty} pq^x dr = & \quad C_0 \int_0^1 p^{-A} q^x dp = C_0 \frac{\Gamma(1-A)\Gamma(x+1)}{\Gamma(x+2-A)} \\ = & \quad C_0 \Gamma(1-A) \frac{x!}{\Gamma(x+2-A)} = C_0 \Gamma(1-A) \frac{1}{C(2-A, x)} \\ (6.25) \qquad \sim & \quad \frac{C_0 \Gamma(1-A)}{x^{1-A}} \end{aligned}$$

The last equality comes from the approximation 6.22. Therefore

$$V(l) = \int_0^{+\infty} (1 - q^l) dr = \sum_{x=0}^{l-1} \int_0^{+\infty} pq^x dr \sim \sum_{x=0}^{l-1} \frac{C_0 \Gamma(1-A)}{x^{1-A}} = C_0 \Gamma(1-A) \sum_{x=0}^{l-1} \frac{1}{x^{1-A}}$$

Now

$$\sum_{x=0}^{l-1} \frac{1}{x^{1-A}} \sim \int_0^l x^{-1+A} dx = \frac{l^A}{A}$$

which implies

$$(6.26) \qquad V(l) \sim \frac{C_0 \Gamma(1-A) l^A}{A}$$

Equation 6.26 is the same as the relationship determined by Mandelbrot [74, see Appendix A] up to a proportional factor.

It is interesting to find the limiting values of the size of the vocabulary for $A \to 0$ (i.e. for $\alpha \to \infty$) and for $A \to 1$ (i.e. for $\alpha \to 0$). First, let us substitute Expression 6.19 for $C_0$ in $V(l)$.

$$V(l) \sim \frac{(\sigma A)^{1-A}(1-A)^{A}\Gamma(1-A)l^{A}}{A} \tag{6.27}$$

Also, note that $(1-A)^{A}\Gamma(1-A) = \frac{\Gamma(2-A)}{(1-A)^{1-A}} \to 1$ for both $A \to 0$ and $A \to 1$, since $\Gamma(2) = 1! = 0! = \Gamma(1)$. Then

$$\lim_{A\to 0} V(l) = \lim_{A\to 0} \frac{\sigma^{1-A}l^{A}}{A^{A}} = \sigma$$

and

$$\lim_{A\to 1} V(l) = \lim_{A\to 1} \frac{\sigma^{1-A}l^{A}}{A^{A}} = l$$

The parameter $A$ measures the expressiveness of the vocabulary or the specificity of the language. A rich vocabulary is characterized by a small value $\alpha$ or equivalently of a value of $A$ close to 1.

### 6.3.2 Example: the Paretian law applied

Let us show how Relation 6.27 has been used with the collection $Wt10g$ [56, 9]. We have to fit the Paretian model with the data thus determining the values for the parameters $\sigma$ and $A$. $\sigma$ is the number of the most frequent terms after the indexing process. These terms appear at the top places in the ranking and therefore these terms are added to the stop list. In Section 2.5 we derived the value $\alpha = 1.36$ for the TREC-10 collection (see Figure 2.5) and $\sigma = r_0 \sim 2^9 = 512$.

Without the use of Porter's stemmer but with a very large initial stop list we have determined a number $V$ of 293,484 unique terms occurring in a text collection of length $l = 469,493,061$. For $\alpha$ we found the value 0.82 and $\sigma = 230$. The corresponding value of $A = (\alpha+1)^{-1}$ is 0.55. This value for $\alpha$ is used to obtain a term-frequency normalization function in Section 6.3.3.

### 6.3.3 The Paretian term-frequency normalization formula

In the previous section, $l$ was the number of total tokens in the collection. This time let $l$ be the length of a document. With a smaller text the expressiveness of the language

that is the size of the vocabulary changes. Let us discuss this problem with some details. With a different and smaller text sample the values of the parameters $\sigma$ and $A$ characterizing the Paretian law should be different. They depend on the sample and not on the population. Obviously with a larger size $l$ of text a variation of expressiveness of the language, expressed in terms of the number of unique words used in the text, can be observed [100]. Many of the non-stop terms in a large text collection are *hapax legomena*. These terms are all terms that appear only once in the collection. In our example the relative frequency of their class is 45.99%. Sichel [100] observed that a proportion of $\sim 50\%$ has been often observed in real cases. It is also observed that the proportion of hapax legomena decreases with the increase of the size of the collection and should go slowly to 0. The proportion of hapax legomena according to Sichel measures the richness of the vocabulary. This is not far from what Paretian law establishes. With the Paretian law, the expressiveness rate of the language is instead measured by the parameter $\alpha$. As it has been shown in the limit cases, $\alpha$ should slowly increase and $A$ thus slowly decrease with a larger collection. However, we here make some reductive assumptions on $\alpha$ and $A$ in order to exploit the Paretian law and obtain a term-frequency normalization. We assume that:

1. The Paretian law is the same for all document samples and its parameters $\sigma$ and $A$ are given by fitting the classical Pareto's model with the empirical data.

2. In order to compare two different frequencies within two different documents we do not compare their maximum likelihood frequencies $\frac{tf}{l_d}$ but their frequency class, that is we compare the expected number $n(tf|l_d)$ of terms in the term class having a given frequency $tf$. This number is provided by Equation 6.23.

3. Under the two hypotheses 1 and 2, the unknown term-frequency $tfn$ satisfies thus the equation

$$n(tf|l) = n(tfn|avg\_l) \tag{6.28}$$

where $avg\_l$ is the average document length.

According to these, we obtain the relation

$$\frac{n(tf|l)}{n(tfn|avg\_l)} = \frac{l^A tf^{-(A+1)}}{avg\_l^A tfn^{-(A+1)}} = 1$$

Solving the equation, we get

$$tfn = \ tf \cdot \left(\frac{avg\_l}{l}\right)^{\frac{A}{A+1}} \tag{6.29}$$

Let

$$Z = \frac{A}{A+1}$$

Since $A$ and $\alpha > 0$ are related by Relation 6.13 and since $A$ ranges in the interval $(0,1)$, the parameter $Z$ ranges in the interval $(0, 0.5)$. For the collection Wt10g, whose parameters are found in Section 6.3.1, the value of $Z$ is $\sim 0.30$ which corresponds to $\alpha = 1.365$.

### Experiments with Paretian term-frequency normalization

We verify that the determined value $\sim 0.3$ of $Z$ in Equation 6.29 for the WT10G collection and that of 2 GB of TREC 7 and TREC-8 lies within the interval of the best match values for retrieval $(0.28 - 0.35)$ using the set of queries of TREC-8 and TREC-9.

The results reported in Table 6.3 are confined to the setting used in the official runs at TREC-10. We have not used the stemming algorithm leaving therefore a richer vocabulary in the set of the most informative terms than after reduction by stemming. We also eliminate some noise by not including the terms occurring less than 10 times. Notice that the performance of the model $B_E L$ with the Zipfian normalization, without query expansion and without stemming (second line of the Table 6.3) is slightly superior to the performance of the model $B_E L$ with **H2** normalization without query expansion (first line of the Table 6.3).

The results of Tables 6.4 and 6.5 show that Zipfian normalization is effective and that the derived parameter from the Pareto-Zipf is close to the experimental Best Match value.

## 6.4 Term-frequency normalization Dirichlet priors

We assume that

$$tfn = \mu \cdot \lambda_{D_i} \tag{6.30}$$

| Method | Parameter | AvPrec | Prec-at-10 | Prec-at-20 | Prec-at-30 |
|---|---|---|---|---|---|
| Model performance without query expansion | | | | | |
| $B_E L2$ | c=7 | 0.1788 | 0.3180 | 0.2730 | 0.2413 |
| $B_E L$ | Pareto Z =0.30 | 0.1824 | 0.3180 | 0.2700 | 0.2393 |
| $B_E L$ | Pareto Z =0.35 | 0.1813 | 0.3200 | 0.2590 | 0.2393 |
| $B_E L$ | Pareto Z =0.40 | 0.1817 | 0.3240 | 0.2670 | 0.2393 |

Table 6.3: Performance of $B_E L$ with different term-frequency normalizations on TREC-10 data.

| TREC 8. | | | | | | | |
|---|---|---|---|---|---|---|---|
| Models | MAP | MAP@10 | Pr @5 | Pr @10 | Pr @20 | R-Prec | RelRet |
| $I(n_e)BZ$, $Z = 0.28$ | 0.2598 | 0.3613 | 0.4960 | 0.4740 | 0.4210 | 0.2983 | 2891 |
| $I(n_e)BZ$, $Z = 0.2942$ | 0.2603 | 0.3608 | 0.4800 | 0.4740 | 0.4180 | 0.3001 | 2898 |
| $I(n_e)BZ$, $Z = 0.31$ | 0.2610 | 0.3615 | 0.4840 | 0.4740 | 0.4160 | 0.3017 | 2900 |
| $I(n_e)BZ$, $Z = 0.32$ | 0.2617 | 0.3617 | 0.4840 | 0.4740 | 0.4200 | 0.3033 | 2904 |
| $I(n_e)BZ$, $Z = 0.33$ | 0.2621 | 0.3599 | 0.4880 | 0.4680 | 0.4210 | 0.3026 | 2905 |
| $I(n_e)BZ$, $Z = 0.34$ | 0.2623 | 0.3582 | 0.4880 | 0.4660 | 0.4240 | 0.3021 | 2903 |
| $I(n_e)BZ$, $Z = 0.35$ | 0.2630 | 0.3567 | 0.4920 | 0.4620 | 0.4240 | 0.3039 | 2905 |
| $I(n_e)BZ$, $Z = 0.36$ | 0.2631 | 0.3584 | 0.4920 | 0.4640 | 0.4240 | 0.3037 | 2905 |
| $I(n_e)BZ$, $Z = 0.38$ | 0.2623 | 0.3589 | 0.4920 | 0.4640 | 0.4210 | 0.3041 | 2904 |

Table 6.4: The performance of the Pareto term-frequency normalization for TREC-8 data. The run $Z = 0.2942$ is that relative to the value of $Z$ corresponding to the slope $\alpha = 1.399$ for the 2 GB collection of TREC-8.

| TREC 9. | | | | | | | |
|---|---|---|---|---|---|---|---|
| Models | MAP | MAP@10 | Pr @5 | Pr @10 | Pr @20 | R-Prec | RelRet |
| $I(n_e)BZ$, $Z = 0.25$ | 0.1858 | 0.2225 | 0.2800 | 0.2480 | 0.2120 | 0.2220 | 1489 |
| $I(n_e)BZ$, $Z = 0.28$ | 0.1905 | 0.2246 | 0.2880 | 0.2440 | 0.2110 | 0.2226 | 1494 |
| $I(n_e)BZ$, $Z = 0.2972$ | 0.1924 | 0.2253 | 0.2840 | 0.2440 | 0.2090 | 0.2218 | 1498 |
| $I(n_e)BZ$, $Z = 0.31$ | 0.1926 | 0.2261 | 0.2800 | 0.2460 | 0.2110 | 0.2228 | 1498 |
| $I(n_e)BZ$, $Z = 0.32$ | 0.1920 | 0.2243 | 0.2840 | 0.2480 | 0.2130 | 0.2254 | 1500 |
| $I(n_e)BZ$, $Z = 0.35$ | 0.1917 | 0.2204 | 0.2800 | 0.2460 | 0.2140 | 0.2242 | 1494 |

Table 6.5: The performance of the Pareto term-frequency normalization for TREC 9 data. The run $Z = 0.2972$ is that relative to the value of $Z$ corresponding to the slope $\alpha = 1.365$ for the wt10g collection.

where $\lambda_{D_i}$ is given by the Dirichlet priors of Equation 5.40 in the models based on divergence from randomness. This is equivalent to

$$tfn = \lambda_i \cdot \mu = \frac{tf_i + \mu \cdot \frac{F_{t_i}}{FreqTotColl}}{l_d + \mu} \cdot \mu \quad \textbf{[H3]} \tag{6.31}$$

In the experiments we test Dirichlet's priors normalization only against the model $I(n_e)B$. The new model is denoted by $I(n_e)B3$.

# Chapter 7

# Normalized models of IR based on divergence from randomness

In previous chapters the three components of our theoretical framework have been entirely developed. We have several instances of each component which only need to be assembled together to obtain the full IR models. In Section 4.8 we have introduced the 14 First Normalized Models in which the term-frequency variable $tf$ was not yet normalized as shown in Chapter 6. We are now ready to provide the retrieval score of each document of the collection with respect to a query. The query is assumed to be a set of independent terms. *Term independence assumption* translates into the additive property of *gain* of Equation 4.27 over the sets of terms occurring in both the query and the observed document. We obtain the final matching function of relevant documents under the hypothesis of the uniform substitution of $tfn$ for $tf$ and the hypothesis $H1$ or $H2$:

$$R(q,d) = \sum_{t \in q} weight(t,d) = \sum_{t \in q} qtf \cdot (1 - Prob_2(tfn)) \cdot Inf_1(tfn) \tag{7.1}$$

where $qtf$ is the multiplicity of term-occurrences in the query.

We cannot here list all models because they are $14 \times 4 = 56$, being 14 the number of First Normalized Models and 4 the normalization techniques presented in Chapter 6. For the sake of completeness we now recapitulate how to instantiate the three components of the model

The weighting formulas are obtained as the product of two informative content func-

tions (see Formula 1.1 on page 28). The first function $Inf_1(tf|d,D) = -\log p_1(tf|d,D)$ is related to the collection $D$. The second $Inf_2 = 1 - p_2(tf|E_t,d)$ takes into account the elite set $E_t$ of the term. The weight

$$w(t|d) = \quad Inf_1(tfn|d,D) \cdot Inf_2(tfn|E_t,d) \tag{7.2}$$

Possible interpretations of $Inf_1$, $Inf_2$ and $tfn$ are:

$$Inf_1(tfn|d,D) = \begin{cases} -\log_2 \left( \dfrac{2^{F \cdot D(f,p)}}{(2\pi tfn(1-f))^{\frac{1}{2}}} \right) & [D] \\ tf \cdot \log_2 \dfrac{tf}{\lambda} + \left( \lambda + \dfrac{1}{12 \cdot tf} - tf \right) \cdot \log_2 e + & \\ +0.5 \cdot \log_2 (2\pi \cdot tf) & [P] \\ -\log_2 \left( \dfrac{1}{1+\lambda} \right) \cdot \left( \dfrac{\lambda}{1+\lambda} \right)^{tfn} & [B_E] \\ \log_2 (1+\lambda) + tf \cdot \log_2 \left( 1 + \dfrac{1}{\lambda} \right) & [G] \\ -\log_2 \left( \dfrac{n_e + 0.5}{N+1} \right)^{tfn} & [I(n_e)] \\ -\log_2 \left( \dfrac{n+0.5}{N+1} \right)^{tfn} & [I(n)] \\ tf \cdot \log_2 \dfrac{N+1}{F+0.5} & [I(F)] \end{cases} \tag{7.3}$$

$$Inf_2(tfn|d,E_t) = \begin{cases} \dfrac{1}{tfn+1} & [L] \\ \dfrac{F+1}{n(tfn+1)} & [B] \end{cases} \tag{7.4}$$

and

$$tfn = \begin{cases} tf \cdot \dfrac{avg\_l}{l} & [1] \\ tf \cdot \ln \left( 1 + \dfrac{c \cdot avg\_l}{l} \right) & [2] \\ \mu \cdot \lambda_{D_i} & [3] \\ tf \cdot \left( \dfrac{avg\_l}{l} \right)^z & [Z] \end{cases} \tag{7.5}$$

The factor $Inf_2$ of Equation (7.2) is the *First Normalization of the informative content* $Inf_1$. The Second Normalization is the uniform substitution of $tf$ for $tfn$ in Equation(7.2). Before presenting results from our experiments we would like to connect the $BM25$ formula to the model $I(n)L2$ assuming the value $c = 1$ in the Second Normalization **H2**.

## 7.1 A derivation of BM25 and INQUERY formula

The normalization of the term-frequency of the ranking formula $BM25$ can be derived by the normalization $L2$, and therefore both the $BM25$ and INQUERY [2] formulae are versions of the model $I(n)L2$:

$$I(n)L2: \quad \frac{tfn}{tfn+k_1} log_2 \frac{N+1}{n+0.5} \tag{7.6}$$

where

$$tfn = tf \cdot log_2(1 + \frac{avg\_l}{l}) \text{ and } k_1 = 1,2$$

Let $k_1 = 1$ and let us introduce the variable $x = \frac{l}{avg\_l}$. Then:

$$\frac{tfn}{tfn+1} = \frac{tf}{tf + \frac{1}{log_2(x+1) - log_2 x}}$$

Let us carry out the Taylor series expansion of the function

$$g(x) = \frac{1}{log_2(x+1) - log_2 x}$$

at the point $x = 1$. Its derivative is

$$g'(x) = \frac{log_2 e \cdot g^2(x)}{x(x+1)}$$

From $g(1) = 1$ and $g'(1) = log_2 e \cdot 0.5$ we obtain

$$\begin{aligned} \frac{tfn}{tfn+1} &= \frac{tf}{tf + 1 + log_2 e \cdot 0.5 \cdot (\frac{l}{avg\_l} - 1) + O((\frac{l}{avg\_l} - 1)^2)} \\ &= \frac{tf}{tf + 0.2786 + 0.7213 \cdot \frac{l}{avg\_l} + O((\frac{l}{avg\_l} - 1)^2)} \end{aligned} \tag{7.7}$$

The expansion of 7.7 in $\frac{tfn}{tfn+1}$ with error $O((\frac{l}{avg\_l} - 1)^3)$ gives

$$\begin{aligned} &\frac{tf}{tf + 1 + log_2 e \cdot 0.5 \cdot (\frac{l}{avg\_l} - 1) - \frac{1}{8} log_2 e \cdot (3 - 2log_2 e)(\frac{l}{avg\_l} - 1)^2} \\ &= \frac{tf}{tf + 0.2580 + 0.7627 \cdot \frac{l}{avg\_l} - 0.0207 \cdot \frac{l}{avg\_l}^2} \end{aligned}$$

The INQUERY normalization factor of Formula 6.2 is obtained with the parameter $k_1 = 2$ which corresponds to the application of Laplace's law of succession as stated in Formula 4.17 (with coefficients 0.5572 and 1.4426 instead of 0.5 and 1.5).

| Collection | TREC | avg_l | $\sigma$ | $\beta = \frac{avg\_l}{\sigma}$ | $\Phi(\beta)$ | documents : $\left\lvert \frac{l}{avg\_l} - 1 \right\rvert < 1$ |
|---|---|---|---|---|---|---|
| Disks 1,2 | 1,2,3 | 209.6 | 776.2 | 0.27 | 0.61 | 0.89 |
| Disks 4,5 | 6 | 265.5 | 1149.4 | 0.23 | 0.59 | 0.91 |
| Disks 4,5 (no CR) | 7,8 | 246.5 | 707.2 | 0.35 | 0.64 | 0.90 |

Table 7.1: The probability $\Phi(\beta)$ is the probability computed by the standard normal distribution that a random document has length $\left\lvert \frac{l}{avg\_l} - 1 \right\rvert < 1$ in a collection with mean $avg\_l$ and variance $\sigma^2$.

When $k_1$ is the default value 1.2 of the $BM25$, the coefficients become 0.3096 and 0.9152 instead of the empirical values 0.3 and 0.9 of the $BM25$ formula.

The $O((\frac{l}{avg\_l}-1)^2)$ in 7.7 is small when $|\frac{l}{avg\_l}-1| < 1$. It is interesting to estimate the probability that the length $l$ of a random document satisfies such a relation. By applying the Central Limit Theorem to the random variable $l$ with mean $avg\_l$ and variance $\sigma^2$, the discrepancy $l - avg\_l < \sigma \cdot \beta$ for every fixed value $\beta$ converges to the value $\Phi(\beta)$ given by the normal distribution $\Phi$. If we set $\beta = \frac{avg\_l}{\sigma}$ the relation $|\frac{l}{avg\_l} - 1| < 1$ is satisfied. Thus the approximation 7.7 should hold when the standard deviation $\sigma$ is close to the mean $avg\_l$. In practice, the expected number of documents satisfying the constraint $|\frac{l}{avg\_l} - 1| < 1$, given by the Central Limit Theorem, is smaller than the actual number, as shown in Table 7.1. The effectiveness of the approximation is confirmed by our experiments, not reported here, that have shown that the $BM25$ formula with its parameters set as in Formula 7.7 has the same performance as $I(n)L2$.

## 7.2 Experimental data

The data we used consisted of three test collections of TREC (Text REtrieval Conference). The first test collection was put on disks 1 and 2, the second collection, on disks 4 and 5. The third collection is the collection WT10g [9].

We also report here the results from the last TREC Conference TREC-11 with a new collection the ".GOV" collection. TREC-11 experiments were carried out by Glasgow

University [81].

Disks 1 and 2 for TREC-1, TREC-2 and TREC-3 experiments consists of about 2 Gbytes of data, of about 528,000 documents from the Department of Energy Abstracts, the Federal Register, the Associated Press Newswire and the Ziff-Davis collections. Disks 1 and 2 contain (after the use of the stop list) 138,743,975 pointers (a pointer is the unit piece of information of the inverted file, that contains the pair "term-document" information and the relative within-document term-frequency). We used the compression techniques of [126] to represent the inverted file in a compressed format. The space required by the compressed inverted file for disks 1 and 2 is 96 Mbytes, i.e. 11.4 bits per pointer. The average length of a document from disks 1 and 2 is 210 tokens (tokens from the stop list were not computed).

The TREC-6 test collection consists of about 2.1 Gbytes of data, of about 556,000 documents, from the Congressional Record, Financial Register, Financial Times, Foreign Broadcast Information Service and LA Times collections. Unlike TREC-6, in TREC-7 and TREC-8, the collection CR (about 28,000 transcripts from Congressional Record) was not indexed. Disks 4 and 5 contain 147,625,088 pointers . The space occupied by the compressed inverted file for disks 4 and 5 is 103 Mbytes, i.e. the inverted file needs 11.2 bits per pointer. The average length of a document on disks 4 and 5 is 265 tokens. This average length decreases to 246 without indexing the CR collection. Indeed, the CR document length average is much longer than the document average length of other collections (624 tokens per document).

The text in the fields that was human-assigned was not indexed for use in the experiments.

## 7.3 Experiments with long queries

For the first test collection we used the topics of TREC-1, TREC-2 and TREC-3 (50 topics each), while for the second collection we used the topics of TREC-6, TREC-7 and TREC-8 (50 topics each).

Each of the 50 topics consists of three fields: a title (from 1 to 3 words), a description (1 or 2 sentences), and a narrative (a paragraph listing specific criteria for accepting or rejecting a document). In our experiments we used all these three fields. We used

Porter's stemming algorithm and a stop list of 235 words.

We tested the basic models with first and second normalization and compared them with model $BM25$ of Okapi as defined by Formula 5.19. To find the non-interpolated average measure of precision (as proposed by Chris Buckley and first used in TREC-2 [49]), for each query and for each $i$-th retrieved relevant document the exact precision $Prob_i$ is computed (i.e. $\frac{i}{r}$, where $r$ is the document position in the rank), then the average precision for the query is obtained (i.e. $\frac{\sum_i Prob_i}{R}$, where $R$ is the number of relevant documents in the collection) and finally the mean of the average precision over all topics (see also Appendix A.1). The non interpolated average precision for the 11 levels of recall is shown in Tables 7.2, 7.3, 7.4, 7.5, 7.8 and 7.9 by MAP, the precision at 5, 10, 30, 100 and R (R-precision) retrieved documents, where R is the number of relevant documents for each query, denoted by Pr@5, Pr@10, Pr@30, Pr@100 and Pr@R respectively. We used $l$ and $avg\_l$ as the length of a document and the average number of tokens in a document in the collection respectively. The results from the experiments can be summarised as follows:

- Results from TREC–1 (see Table 7.2). $I(n_e)B2$ and its approximation $I(F)B2$ have the best average precision and precision at 5 documents retrieved. The two limiting forms of Bose-Einstein model, $GB2$ and $B_EB2$, have best precision at 10. BM25 has best precision for high recall.

- Results from TREC–2 (see Table 7.3). $I(n_e)L2$ and its approximation $I(F)L2$ have the best average precision and precision at 5 documents retrieved. The standard *idf-tf* model with Laplace's Law of Succession, $I(n)L2$, has the best precision at 30. BM25 has the best precision at high recall values and the highest precision at 10.

- Results from TREC–3 (see Table 7.4). $I(n_e)L2$ and its approximation $I(F)L2$ have the best average precision. The two approximations of the Bernoulli model, $PL2$ and $DL2$, have the highest precision at 5 documents retrieved. The standard *idf-tf* model with Laplace's Law of Succession, $I(n)L2$, has the best precision at 30. BM25 has the best precision at high recall values and the highest precision at 10.

- Results from TREC–6 (see Table 7.5). The standard *tf-idf* model with Laplace's Law of Succession, $I(n)L2$, has the highest precision at 5 documents retrieved. $I(n_e)B1$, namely the *idf* and Poisson mixture model together with the uniform distribution hypothesis on term-frequency $H1$ and the Bernoulli normalization $B$, has the best performance at higher recall values.

- Results from TREC–6 without the CR collection (see Table 7.6). Removing long documents from the collection has positive effects on the approximation $G$ of the Bose-Einstein model and on the term-frequency normalization $B$.

- Results from TREC–7 (see Table 7.8). $I(n_e)L2$ and its approximation $I(F)L2$ have the highest precision at different recall levels.

- Results from TREC–8 (see Table 7.9).Similarly to TREC-7, $I(n_e)L2$ and its approximation $I(F)L2$ have the highest precision at different recall levels, except for the Poisson model $PL2$ which has the highest precision at 5.

### 7.3.1 Results from experiments with long queries

Our results show that all the models are robust with respect to different data sets. We have used a parameter-free version of the term-frequency normalization **H2**, that is with $c = 1$. Notwithstanding the fact that we have not contributed parameters, models are shown to have a performance in most TREC experiments better than BM25 (TREC-10 included). In the following we discuss the results shown in Tables 7.13–7.9.

1. There is no convincing evidence or argument in favour of either normalization $B$ or $L$. The results of TREC-7 (Table 7.8) are confirmed on TREC-8 (Table 7.9) and similarly, the relative performance of the models in TREC-1, TREC-2 and TREC-3 (see Tables 7.2,7.3,7.4) shows similar trends. In TREC-1, TREC-2 and TREC-3, $L2$ is in general superior to $B2$ independently of the basic model used, while in TREC-7, TREC-8 and TREC-10 (see Tables 7.8,7.9,7.13), $B2$ is in general superior to $L2$ independently of the basic model used. The notable exception is the Poisson model $P$: $L1$ and $L2$ performs in general better than $B2$.

Disks 1 and 2 of TREC 1, topics 51-100. Relevant documents: 16386

| Models | MAP | Pr@5 | Pr@10 | Pr@30 | Pr@100 | Pr@R | Rel Ret |
|---|---|---|---|---|---|---|---|
| $I(F)B1$ | 0.1989 | 0.6200 | 0.5660 | 0.4973 | 0.3886 | 0.2813 | 7128 |
| $I(F)L1$ | 0.1933 | 0.5760 | 0.5760 | 0.4853 | 0.3814 | 0.2751 | 6993 |
| $I(F)B2$ | **0.2103** | 0.6400 | 0.5740 | **0.5333** | 0.4038 | 0.2878 | **7396** |
| $I(F)L2$ | 0.2068 | 0.6200 | 0.5700 | 0.5127 | 0.3978 | 0.2843 | 7300 |
| $I(n)B1$ | 0.1911 | 0.6040 | 0.5740 | 0.5027 | 0.3798 | 0.2675 | 6928 |
| $I(n)L1$ | 0.1968 | 0.5920 | 0.5600 | 0.5013 | 0.3908 | 0.2787 | 7034 |
| $I(n)B2$ | 0.2003 | 0.6280 | 0.5900 | 0.5200 | 0.3964 | 0.2781 | 7123 |
| $I(n)L2$ | 0.2077 | 0.6200 | 0.5800 | 0.5193 | 0.4030 | 0.2863 | 7267 |
| $I(n_e)B1$ | 0.1985 | 0.6240 | 0.5660 | 0.4987 | 0.3882 | 0.2795 | 7109 |
| $I(n_e)L1$ | 0.1946 | 0.5800 | 0.5420 | 0.4907 | 0.3856 | 0.2764 | 7006 |
| $I(n_e)B2$ | 0.2098 | **0.6440** | 0.5860 | 0.5327 | 0.4054 | 0.2865 | 7395 |
| $I(n_e)L2$ | 0.2073 | 0.6200 | 0.5720 | 0.5153 | 0.4004 | 0.2852 | 7307 |
| $GB1$ | 0.1984 | 0.6120 | 0.5820 | 0.5093 | 0.3934 | 0.2782 | 7144 |
| $GL1$ | 0.1968 | 0.5920 | 0.5560 | 0.4953 | 0.3878 | 0.2771 | 7093 |
| $GB2$ | 0.2041 | 0.6320 | **0.5980** | 0.5193 | 0.3974 | 0.2816 | 7274 |
| $GL2$ | 0.2047 | 0.6280 | 0.5660 | 0.5107 | 0.3952 | 0.2856 | 7232 |
| $B_EB1$ | 0.1984 | 0.6120 | 0.5820 | 0.5093 | 0.3934 | 0.2782 | 7144 |
| $B_EL1$ | 0.1968 | 0.5920 | 0.5560 | 0.4953 | 0.3878 | 0.2771 | 7093 |
| $B_EB2$ | 0.2042 | 0.6320 | **0.5980** | 0.5193 | 0.3974 | 0.2816 | 7276 |
| $B_EL2$ | 0.2047 | 0.6280 | 0.5660 | 0.5107 | 0.3952 | 0.2856 | 7232 |
| $PB1$ | 0.1696 | 0.5360 | 0.5020 | 0.4587 | 0.3536 | 0.2517 | 6404 |
| $PL1$ | 0.1741 | 0.5360 | 0.5300 | 0.4593 | 0.3562 | 0.2572 | 6442 |
| $PB2$ | 0.2003 | 0.6000 | 0.5900 | 0.5127 | 0.3970 | 0.2755 | 7094 |
| $PL2$ | 0.2065 | 0.6360 | 0.5780 | 0.5087 | 0.4056 | 0.2861 | 7124 |
| $DB1$ | 0.1695 | 0.5360 | 0.5000 | 0.4587 | 0.3536 | 0.2513 | 6404 |
| $DL1$ | 0.1741 | 0.5360 | 0.5300 | 0.4587 | 0.3562 | 0.2572 | 6442 |
| $DB2$ | 0.2003 | 0.6000 | 0.5900 | 0.5127 | 0.3970 | 0.2755 | 7094 |
| $DL2$ | 0.2065 | 0.6360 | 0.5780 | 0.5087 | 0.4056 | 0.2861 | 7124 |
| $BM25$ | 0.2091 | 0.6240 | 0.5740 | 0.5260 | **0.4080** | **0.2882** | 7307 |

Table 7.2: Results from TREC–1 with the long queries. The best precision values are in bold. See Section 1.8 and Table 1.2 for an explanation of the model names.

Disks 1 and 2 of TREC 2, topics 101-150. Relevant documents: 11645

| Models | MAP | Pr@5 | Pr@10 | Pr@30 | Pr@100 | Pr@R | Rel Ret |
|---|---|---|---|---|---|---|---|
| $I(F)B1$ | 0.2320 | 0.5640 | 0.5180 | 0.4800 | 0.4090 | 0.3069 | 6356 |
| $I(F)L1$ | 0.2333 | 0.5720 | 0.5420 | 0.4853 | 0.4026 | 0.3116 | 6322 |
| $I(F)B2$ | 0.2413 | 0.5640 | 0.5440 | 0.4960 | 0.4134 | 0.3142 | 6464 |
| $I(F)L2$ | **0.2456** | 0.5880 | 0.5540 | 0.5087 | 0.4160 | 0.3208 | 6497 |
| $I(n)B1$ | 0.2225 | 0.5480 | 0.5160 | 0.4780 | 0.4028 | 0.3006 | 6261 |
| $I(n)L1$ | 0.2364 | 0.5680 | 0.5440 | 0.5047 | 0.4130 | 0.3148 | 6380 |
| $I(n)B2$ | 0.2262 | 0.5600 | 0.5200 | 0.4907 | 0.4086 | 0.3037 | 6258 |
| $I(n)L2$ | 0.2439 | 0.5560 | 0.5420 | **0.5147** | 0.4224 | 0.3187 | 6472 |
| $I(n_e)B1$ | 0.2325 | 0.5560 | 0.5260 | 0.4873 | 0.4110 | 0.3093 | 6410 |
| $I(n_e)L1$ | 0.2348 | 0.5720 | 0.5460 | 0.4920 | 0.4050 | 0.3137 | 6349 |
| $I(n_e)B2$ | 0.2406 | 0.5600 | 0.5420 | 0.4993 | 0.4154 | 0.3155 | 6483 |
| $I(n_e)L2$ | **0.2456** | **0.5960** | 0.5540 | 0.5087 | 0.4176 | 0.3219 | 6503 |
| $GB1$ | 0.2329 | 0.5440 | 0.5280 | 0.4833 | 0.4112 | 0.3094 | 6392 |
| $GL1$ | 0.2379 | 0.5800 | 0.5540 | 0.4980 | 0.4074 | 0.3178 | 6392 |
| $GB2$ | 0.2336 | 0.5400 | 0.5220 | 0.4947 | 0.4106 | 0.3089 | 6320 |
| $GL2$ | 0.2417 | 0.5800 | 0.5440 | 0.5120 | 0.4142 | 0.3177 | 6391 |
| $B_EB1$ | 0.2329 | 0.5440 | 0.5280 | 0.4833 | 0.4112 | 0.3094 | 6392 |
| $B_EL1$ | 0.2379 | 0.5800 | 0.5540 | 0.4980 | 0.4074 | 0.3179 | 6392 |
| $B_EB2$ | 0.2336 | 0.5400 | 0.5220 | 0.4947 | 0.4106 | 0.3089 | 6321 |
| $B_EL2$ | 0.2418 | 0.5800 | 0.5440 | 0.5120 | 0.4144 | 0.3181 | 6391 |
| $PB1$ | 0.1951 | 0.5280 | 0.5060 | 0.4667 | 0.3772 | 0.2780 | 5769 |
| $PL1$ | 0.2089 | 0.5640 | 0.5260 | 0.4700 | 0.3836 | 0.2892 | 5924 |
| $PB2$ | 0.2223 | 0.5760 | 0.5420 | 0.4940 | 0.4144 | 0.3039 | 6232 |
| $PL2$ | 0.2383 | 0.5880 | 0.5540 | 0.5000 | 0.4194 | 0.3223 | 6402 |
| $DB1$ | 0.1951 | 0.5280 | 0.5060 | 0.4660 | 0.3772 | 0.2776 | 5769 |
| $DL1$ | 0.2089 | 0.5640 | 0.5260 | 0.4693 | 0.3836 | 0.2892 | 5924 |
| $DB2$ | 0.2223 | 0.5760 | 0.5420 | 0.4940 | 0.4144 | 0.3039 | 6232 |
| $DL2$ | 0.2383 | 0.5880 | 0.5540 | 0.5000 | 0.4196 | 0.3223 | 6403 |
| $BM25$ | **0.2455** | 0.5720 | **0.5560** | 0.5087 | **0.4252** | **0.3230** | **6523** |

**Table 7.3: Results from TREC–2 with the long queries. The best precision values are in bold. See Section 1.8 and Table 1.2 for an explanation of the model names.**

Disks 1 and 2 of TREC 3, topics 151-200. Relevant documents: 9805

| Models | MAP | Pr@5 | Pr@10 | Pr@30 | Pr@100 | Pr@R | Rel Ret |
|---|---|---|---|---|---|---|---|
| $I(F)B1$ | 0.2565 | 0.6960 | 0.6520 | 0.5320 | 0.3776 | 0.3217 | 5437 |
| $I(F)L1$ | 0.2675 | 0.6960 | 0.6560 | 0.5367 | 0.3832 | 0.3336 | 5460 |
| $I(F)B2$ | 0.2644 | 0.7160 | 0.6620 | 0.5380 | 0.3846 | 0.3254 | 5516 |
| $I(F)L2$ | 0.2765 | 0.7440 | 0.6660 | 0.5540 | 0.3902 | **0.3390** | 5524 |
| $I(n)B1$ | 0.2439 | 0.6800 | 0.6400 | 0.5193 | 0.3694 | 0.3100 | 5320 |
| $I(n)L1$ | 0.2669 | 0.7080 | 0.6740 | 0.5367 | 0.3870 | 0.3329 | 5535 |
| $I(n)B2$ | 0.2480 | 0.7000 | 0.6540 | 0.5307 | 0.3714 | 0.3114 | 5315 |
| $I(n)L2$ | 0.2716 | 0.7280 | 0.6720 | 0.5500 | 0.3926 | 0.3325 | 5524 |
| $I(n_e)B1$ | 0.2569 | 0.7000 | 0.6540 | 0.5313 | 0.3820 | 0.3223 | 5454 |
| $I(n_e)L1$ | 0.2682 | 0.6880 | 0.6580 | 0.5420 | 0.3826 | 0.3348 | 5483 |
| $I(n_e)B2$ | 0.2637 | 0.7080 | 0.6680 | 0.5400 | 0.3848 | 0.3258 | 5514 |
| $I(n_e)L2$ | **0.2767** | 0.7320 | 0.6720 | 0.5533 | 0.3906 | 0.3379 | 5543 |
| $GB1$ | 0.2548 | 0.6880 | 0.6580 | 0.5227 | 0.3746 | 0.3182 | 5436 |
| $GL1$ | 0.2681 | 0.6960 | 0.6800 | 0.5393 | 0.3842 | 0.3343 | 5495 |
| $GB2$ | 0.2527 | 0.7040 | 0.6520 | 0.5260 | 0.3750 | 0.3165 | 5373 |
| $GL2$ | 0.2682 | 0.7120 | 0.6680 | 0.5447 | 0.3818 | 0.3303 | 5446 |
| $B_EB1$ | 0.2548 | 0.6920 | 0.6580 | 0.5220 | 0.3746 | 0.3182 | 5436 |
| $B_EL1$ | 0.2681 | 0.6960 | 0.6780 | 0.5393 | 0.3840 | 0.3343 | 5495 |
| $B_EB2$ | 0.2527 | 0.7040 | 0.6520 | 0.5260 | 0.3750 | 0.3165 | 5373 |
| $B_EL2$ | 0.2683 | 0.7120 | 0.6680 | 0.5447 | 0.3820 | 0.3303 | 5446 |
| $PB1$ | 0.2107 | 0.5800 | 0.5400 | 0.4667 | 0.3330 | 0.2821 | 4990 |
| $PL1$ | 0.2314 | 0.6280 | 0.5800 | 0.4873 | 0.3466 | 0.3056 | 5092 |
| $PB2$ | 0.2459 | 0.7120 | 0.6660 | 0.5267 | 0.3744 | 0.3093 | 5336 |
| $PL2$ | 0.2705 | **0.7520** | 0.6780 | 0.5573 | 0.3934 | 0.3274 | 5490 |
| $DB1$ | 0.2107 | 0.5800 | 0.5400 | 0.4667 | 0.3330 | 0.2821 | 4990 |
| $DL1$ | 0.2314 | 0.6280 | 0.5800 | 0.4873 | 0.3466 | 0.3056 | 5092 |
| $DB2$ | 0.2459 | 0.7120 | 0.6660 | 0.5273 | 0.3744 | 0.3093 | 5336 |
| $DL2$ | 0.2706 | **0.7520** | 0.6780 | 0.5573 | 0.3934 | 0.3274 | 5490 |
| $BM25$ | 0.2754 | 0.7320 | 0.6840 | 0.5587 | 0.3960 | 0.3352 | 5586 |

Table 7.4: Results from TREC–3 with the long queries. The best precision values are in bold. See Section 1.8 and Table 1.2 for an explanation of the model names.

Disks 4 and 5 of TREC 6, topics 301-350. Relevant documents: 4611

| Models | MAP | Pr@5 | Pr@10 | Pr@30 | Pr@100 | Pr@R | Rel Ret |
|---|---|---|---|---|---|---|---|
| $I(F)B1$ | 0.2457 | 0.5160 | 0.4580 | 0.3427 | 0.2162 | 0.2885 | 2667 |
| $I(F)L1$ | 0.2557 | 0.5400 | 0.4420 | 0.3293 | 0.2074 | 0.2979 | 2640 |
| $I(F)B2$ | 0.2482 | 0.5240 | 0.4840 | 0.3367 | 0.2092 | 0.2863 | 2651 |
| $I(F)L2$ | 0.2597 | 0.5400 | 0.4600 | 0.3267 | 0.2058 | 0.2962 | 2595 |
| | | | | | | | |
| $I(n)B1$ | 0.2381 | 0.5280 | 0.4620 | 0.3413 | 0.2144 | 0.2794 | 2607 |
| $I(n)L1$ | 0.2560 | 0.5520 | 0.4480 | 0.3327 | 0.2090 | 0.3017 | 2654 |
| $I(n)B2$ | 0.2362 | 0.5440 | 0.4640 | 0.3327 | 0.2062 | 0.2730 | 2546 |
| $I(n)L2$ | 0.2544 | **0.5760** | 0.4840 | 0.3333 | 0.2126 | 0.2887 | 2594 |
| | | | | | | | |
| $I(n_e)B1$ | 0.2479 | 0.5280 | 0.4640 | **0.3487** | **0.2182** | 0.2940 | **2689** |
| $I(n_e)L1$ | 0.2557 | 0.5560 | 0.4700 | 0.3427 | 0.2164 | 0.2950 | 2654 |
| $I(n_e)B2$ | 0.2488 | 0.5480 | **0.4860** | 0.3393 | 0.2112 | 0.2855 | 2638 |
| $I(n_e)L2$ | **0.2600** | 0.5480 | 0.4620 | 0.3313 | 0.2086 | 0.2931 | 2595 |
| | | | | | | | |
| $GB1$ | 0.2458 | 0.5480 | 0.4700 | 0.3473 | 0.2124 | 0.2883 | 2653 |
| $GL1$ | 0.2567 | 0.5400 | 0.4620 | 0.3367 | 0.2116 | **0.3051** | 2623 |
| $GB2$ | 0.2414 | 0.5320 | 0.4720 | 0.3333 | 0.2058 | 0.2797 | 2566 |
| $GL2$ | 0.2548 | 0.5400 | 0.4560 | 0.3253 | 0.2074 | 0.2879 | 2538 |
| | | | | | | | |
| $B_EB1$ | 0.2452 | 0.5480 | 0.4680 | 0.3467 | 0.2120 | 0.2878 | 2652 |
| $B_EL1$ | 0.2562 | 0.5400 | 0.4620 | 0.3353 | 0.2114 | 0.3045 | 2622 |
| $B_EB2$ | 0.2410 | 0.5320 | 0.4720 | 0.3327 | 0.2058 | 0.2791 | 2565 |
| $B_EL2$ | 0.2546 | 0.5400 | 0.4560 | 0.3253 | 0.2072 | 0.2879 | 2537 |
| | | | | | | | |
| $PB1$ | 0.2032 | 0.4600 | 0.4140 | 0.3100 | 0.1878 | 0.2445 | 2307 |
| $PL1$ | 0.2243 | 0.4760 | 0.4260 | 0.3247 | 0.2000 | 0.2642 | 2452 |
| $PB2$ | 0.2183 | 0.5040 | 0.4440 | 0.3113 | 0.1870 | 0.2509 | 2373 |
| $PL2$ | 0.2424 | 0.5320 | 0.4560 | 0.3300 | 0.2010 | 0.2778 | 2497 |
| | | | | | | | |
| $DB1$ | 0.2027 | 0.4600 | 0.4120 | 0.3100 | 0.1878 | 0.2440 | 2306 |
| $DL1$ | 0.2238 | 0.4760 | 0.4260 | 0.3240 | 0.1998 | 0.2636 | 2451 |
| $DB2$ | 0.2178 | 0.5040 | 0.4440 | 0.3107 | 0.1868 | 0.2503 | 2372 |
| $DL2$ | 0.2421 | 0.5320 | 0.4560 | 0.3300 | 0.2008 | 0.2778 | 2496 |
| | | | | | | | |
| $BM25$ | 0.2440 | 0.5600 | 0.4700 | 0.3233 | 0.2032 | 0.2834 | 2511 |

**Table 7.5: Results from TREC–6 with the long queries. The best precision values are in bold. See Section 1.8 and Table 1.2 for an explanation of the model names.**

Disks 4 and 5 without CR collection, topics 301-350 of TREC 6. Rel. doc.: 4290

| Models | MAP | Pr@5 | Pr@10 | Pr@30 | Pr@100 | Pr@R | Rel Ret |
|---|---|---|---|---|---|---|---|
| $I(n)B1$ | 0.2550 | 0.5240 | 0.4600 | **0.3420** | 0.2130 | 0.2906 | 2535 |
| $I(n)L1$ | 0.2689 | 0.5320 | 0.4540 | 0.3380 | 0.2112 | 0.3089 | 2568 |
| $I(n)B2$ | 0.2581 | 0.5560 | 0.4680 | 0.3320 | 0.2036 | 0.2866 | 2470 |
| $I(n)L2$ | 0.2705 | 0.5560 | **0.4840** | 0.3267 | 0.2088 | 0.3004 | 2510 |
| $I(n_e)B1$ | 0.2648 | 0.5400 | 0.4480 | 0.3393 | **0.2176** | 0.3025 | **2615** |
| $I(n_e)L1$ | 0.2711 | 0.5320 | 0.4500 | 0.3320 | 0.2058 | 0.3154 | 2545 |
| $I(n_e)B2$ | 0.2662 | **0.5680** | 0.4680 | 0.3373 | 0.2100 | 0.2991 | 2566 |
| $I(n_e)L2$ | **0.2751** | 0.5440 | 0.4620 | 0.3213 | 0.2044 | 0.3129 | 2493 |
| $GB1$ | 0.2615 | 0.5400 | 0.4500 | 0.3407 | 0.2118 | 0.2997 | 2576 |
| $GL1$ | 0.2714 | 0.5400 | 0.4540 | 0.3327 | 0.2070 | **0.3169** | 2527 |
| $GB2$ | 0.2605 | 0.5560 | 0.4740 | 0.3340 | 0.2038 | 0.2893 | 2502 |
| $GL2$ | 0.2707 | 0.5440 | 0.4540 | 0.3247 | 0.2028 | 0.3018 | 2444 |
| $PB1$ | 0.2170 | 0.4640 | 0.4060 | 0.3073 | 0.1842 | 0.2566 | 2271 |
| $PL1$ | 0.2373 | 0.4600 | 0.4220 | 0.3187 | 0.1960 | 0.2750 | 2373 |
| $PB2$ | 0.2338 | 0.5160 | 0.4400 | 0.3073 | 0.1868 | 0.2653 | 2318 |
| $PL2$ | 0.2569 | 0.5160 | 0.4480 | 0.3213 | 0.1972 | 0.2882 | 2417 |
| $BM25$ | 0.2584 | 0.5200 | 0.4560 | 0.3167 | 0.1978 | 0.2943 | 2420 |

Table 7.6: Results from TREC–6 with the long queries and removing long documents. The best precision values are in bold. See Section 1.8 and Table 1.2 for an explanation of the model names.

| TREC | MAP | Pr@5 | Pr@10 | Pr@30 | Pr@100 | Pr@R | Rel Ret |
|---|---|---|---|---|---|---|---|
| 1 | $I(n_e)B2$ | $I(n_e)B2$ | GB2 | $I(n_e)B2$ | BM25 | BM25 | $I(n_e)B2$ |
| 2 | $I(n_e)L2$ | $I(n_e)L2$ | BM25 | I(n)L2 | BM25 | BM25 | BM25 |
| 3 | $I(n_e)L2$ | PL2 | BM25 | BM25 | BM25 | $I(n_e)L2$ | BM25 |
| 6 | $I(n_e)L2$ | I(n)L2 | $I(n_e)B2$ | $I(n_e)B1$ | $I(n_e)B1$ | GL1 | $I(n_e)B1$ |
| 7 | $I(n_e)B2$ | $I(n_e)B2$ | $I(n_e)B2$ | $I(n_e)B2$ | GB1 | $I(n_e)B2$ | $I(n_e)B2$ |
| 8 | $I(n_e)B2$ | PL2 | $I(n_e)B2$ | $I(n_e)B2$ | $I(n_e)B2$ | $I(n_e)B2$ | $I(n_e)B2$ |

Table 7.7: Best performing models for each test collection and for different precision measures. The basic probability models $I(F)$, $D$ and $B_E$ are not considered here, as they do not differ significantly from their alternative approximations $I(n_e)$, $P$ and $G$ respectively. See Section 1.8 and Table 1.2 for an explanation of the model names.

Disks 4 and 5 of TREC 7, topics 351-400. Relevant documents: 4674

| Models | MAP | Pr@5 | Pr@10 | Pr@30 | Pr@100 | Pr@R | Rel Ret |
|---|---|---|---|---|---|---|---|
| $I(F)B1$ | 0.2352 | 0.5720 | 0.4960 | 0.3700 | 0.2370 | 0.2785 | 2876 |
| $I(F)L1$ | 0.2180 | 0.5320 | 0.4780 | 0.3553 | 0.2170 | 0.2586 | 2777 |
| $I(F)B2$ | **0.2484** | **0.5800** | **0.5200** | **0.3813** | 0.2374 | 0.2869 | **2883** |
| $I(F)L2$ | 0.2312 | 0.5400 | 0.5000 | 0.3647 | 0.2158 | 0.2711 | 2796 |
| $I(n)B1$ | 0.2191 | 0.5240 | 0.4720 | 0.3413 | 0.2116 | 0.2625 | 2531 |
| $I(n)L1$ | 0.2225 | 0.5520 | 0.4920 | 0.3620 | 0.2230 | 0.2659 | 2828 |
| $I(n)B2$ | 0.2337 | 0.5520 | 0.4840 | 0.3467 | 0.2164 | 0.2700 | 2540 |
| $I(n)L2$ | 0.2360 | 0.5400 | 0.4960 | 0.3687 | 0.2278 | 0.2763 | 2845 |
| $I(n_e)B1$ | 0.2352 | 0.5680 | 0.4960 | 0.3700 | 0.2382 | 0.2778 | 2861 |
| $I(n_e)L1$ | 0.2184 | 0.5440 | 0.4760 | 0.3553 | 0.2176 | 0.2601 | 2782 |
| $I(n_e)B2$ | 0.2482 | **0.5800** | 0.5100 | **0.3813** | 0.2386 | **0.2874** | 2881 |
| $I(n_e)L2$ | 0.2320 | 0.5400 | 0.4980 | 0.3613 | 0.2174 | 0.2717 | 2810 |
| $GB1$ | 0.2364 | 0.5720 | 0.5000 | 0.3760 | **0.2390** | 0.2787 | 2859 |
| $GL1$ | 0.2196 | 0.5360 | 0.4720 | 0.3527 | 0.2166 | 0.2640 | 2770 |
| $GB2$ | 0.2463 | 0.5720 | 0.5100 | 0.3753 | 0.2350 | 0.2847 | 2858 |
| $GL2$ | 0.2315 | 0.5520 | 0.4880 | 0.3587 | 0.2174 | 0.2713 | 2780 |
| $B_EB1$ | 0.2361 | 0.5720 | 0.5000 | 0.3760 | **0.2390** | 0.2787 | 2859 |
| $B_EL1$ | 0.2196 | 0.5360 | 0.4720 | 0.3527 | 0.2166 | 0.2640 | 2770 |
| $B_EB2$ | 0.2462 | 0.5720 | 0.5100 | 0.3753 | 0.2350 | 0.2847 | 2858 |
| $B_EL2$ | 0.2315 | 0.5520 | 0.4880 | 0.3580 | 0.2174 | 0.2713 | 2780 |
| $PB1$ | 0.1914 | 0.4840 | 0.4300 | 0.3407 | 0.2126 | 0.2434 | 2526 |
| $PL1$ | 0.1944 | 0.4640 | 0.4480 | 0.3440 | 0.2092 | 0.2465 | 2584 |
| $PB2$ | 0.2194 | 0.5200 | 0.5020 | 0.3533 | 0.2208 | 0.2624 | 2669 |
| $PL2$ | 0.2212 | 0.5120 | 0.4880 | 0.3607 | 0.2194 | 0.2634 | 2743 |
| $DB1$ | 0.1914 | 0.4840 | 0.4300 | 0.3407 | 0.2126 | 0.2434 | 2526 |
| $DL1$ | 0.1944 | 0.4640 | 0.4480 | 0.3440 | 0.2092 | 0.2465 | 2584 |
| $DB2$ | 0.2194 | 0.5200 | 0.5020 | 0.3533 | 0.2206 | 0.2624 | 2669 |
| $DL2$ | 0.2212 | 0.5120 | 0.4880 | 0.3607 | 0.2194 | 0.2634 | 2743 |
| $BM25$ | 0.2274 | 0.5320 | 0.4880 | 0.3540 | 0.2152 | 0.2643 | 2676 |

Table 7.8: Results from TREC–7 with the long queries. The best precision values are in bold. See Section 1.8 and Table 1.2 for an explanation of the model names.

Disks 4 and 5 of TREC 8, topics 401-450. Relevant documents: 4728

| Models | MAP | Pr@5 | Pr@10 | Pr@30 | Pr@100 | Pr@R | Rel Ret |
|---|---|---|---|---|---|---|---|
| $I(F)B1$ | 0.2734 | 0.5400 | 0.4820 | 0.3820 | 0.2496 | 0.3135 | 3135 |
| $I(F)L1$ | 0.2645 | 0.5280 | 0.4860 | 0.3700 | 0.2416 | 0.3103 | 3067 |
| $I(F)B2$ | 0.2833 | 0.5520 | 0.5060 | **0.3967** | 0.2528 | 0.3280 | **3189** |
| $I(F)L2$ | 0.2767 | 0.5240 | 0.4860 | 0.3840 | 0.2448 | 0.3179 | 3095 |
| $I(n)L1$ | 0.2681 | 0.5120 | 0.5000 | 0.3787 | 0.2444 | 0.3164 | 3046 |
| $I(n)B1$ | 0.2664 | 0.5240 | 0.4740 | 0.3880 | 0.2524 | 0.3221 | 3000 |
| $I(n)B2$ | 0.2763 | 0.5520 | 0.4980 | 0.3900 | 0.2528 | 0.3235 | 3038 |
| $I(n)L2$ | 0.2792 | 0.5360 | 0.5040 | 0.3927 | 0.2492 | 0.3233 | 3073 |
| $I(n_e)B1$ | 0.2735 | 0.5320 | 0.4960 | 0.3807 | 0.2504 | 0.3286 | 3142 |
| $I(n_e)L1$ | 0.2664 | 0.5240 | 0.4840 | 0.3707 | 0.2420 | 0.3114 | 3061 |
| $I(n_e)B2$ | **0.2841** | 0.5520 | **0.5080** | **0.3967** | **0.2532** | **0.3295** | 3178 |
| $I(n_e)L2$ | 0.2769 | 0.5200 | 0.4940 | 0.3887 | 0.2452 | 0.3171 | 3067 |
| $GB1$ | 0.2757 | 0.5360 | 0.4800 | 0.3880 | 0.2494 | 0.3292 | 3142 |
| $GL1$ | 0.2667 | 0.5120 | 0.4840 | 0.3727 | 0.2416 | 0.3146 | 3031 |
| $GB2$ | 0.2826 | 0.5440 | 0.5040 | 0.3960 | 0.2514 | 0.3290 | 3153 |
| $GL2$ | 0.2757 | 0.5280 | 0.4860 | 0.3887 | 0.2438 | 0.3183 | 3032 |
| $B_EB1$ | 0.2757 | 0.5400 | 0.4800 | 0.3880 | 0.2494 | 0.3292 | 3142 |
| $B_EL1$ | 0.2669 | 0.5120 | 0.4860 | 0.3727 | 0.2416 | 0.3146 | 3031 |
| $B_EB2$ | 0.2827 | 0.5440 | 0.5040 | 0.3960 | 0.2514 | 0.3290 | 3153 |
| $B_EL2$ | 0.2758 | 0.5280 | 0.4880 | 0.3887 | 0.2438 | 0.3183 | 3032 |
| $PB1$ | 0.2379 | 0.5240 | 0.4800 | 0.3520 | 0.2246 | 0.2905 | 2838 |
| $PL1$ | 0.2350 | 0.5120 | 0.4700 | 0.3553 | 0.2232 | 0.2898 | 2829 |
| $PB2$ | 0.2559 | 0.5560 | 0.4980 | 0.3847 | 0.2360 | 0.3060 | 2948 |
| $PL2$ | 0.2562 | **0.5680** | 0.4880 | 0.3780 | 0.2374 | 0.3044 | 2923 |
| $DB1$ | 0.2379 | 0.5240 | 0.4800 | 0.3520 | 0.2246 | 0.2905 | 2839 |
| $DL1$ | 0.2350 | 0.5120 | 0.4700 | 0.3553 | 0.2232 | 0.2898 | 2829 |
| $DB2$ | 0.2559 | 0.5560 | 0.4980 | 0.3840 | 0.2358 | 0.3060 | 2948 |
| $DL2$ | 0.2562 | 0.5680 | 0.4880 | 0.3780 | 0.2374 | 0.3044 | 2923 |
| $BM25$ | 0.2716 | 0.5400 | 0.4980 | 0.3827 | 0.2464 | 0.3181 | 3083 |

Table 7.9: Results from TREC–8 with the long queries. The best precision values are in bold. See Section 1.8 and Table 1.2 for an explanation of the model names.

It is interesting to observe that results of TREC-6 (Table 7.5) (whose test bed uses the additional collection CR containing long documents) are significantly different from all other TREC experiments. This allows us to conjecture but not to assert that the statistics of the collection (e.g. number of unique terms, mean and variance of document length) may have more effect on the relative performance of models than the content of the submitted topics. We tried a small experiment which seemed to begin to corroborate this hypothesis. We used the topics of TREC-6 on the collection used in TREC-7 and TREC-8 (without indexing the collection CR). In order to compare the two Tables 7.5 and 7.6 we considered the means of different precision values and of the number of retrieved documents in Table 7.6 and computed the variation rates with respect to the values of Table 7.5 and then normalized to the mean values. Results show that the normalization $B$ increases average precision and more significantly the early precision, that is the precision at the beginning of the ranking, while $L$ slightly increases the precision for high values of recall (R-precision included). Model $G$ showed the most sensitivity to the effect of the normalization process.

2. The Poisson model $PL2$ gave a good performance for precision early in the ranking (precision at 5 documents retrieved). For the average precision, Poisson performance is good in TREC-1, TREC-2 and TREC-3 (see Tables 7.2,7.3,7.4), less satisfactory in TREC-6 and TREC-7 (see Tables 7.5,7.8), unsatisfactory in TREC-8 (Table 7.9) (but in TREC-8, $PL2$ has the best performance for precision at 5 documents retrieved). By contrast, the normalization $B2$ seems to work poorly with $P$.

3. Model $G$ with both normalizations $B2$ and $L2$ gave a good performance in all TREC experiments. $G$'s performance depends on the choice of the normalization $B2$ (better in TREC-7 and TREC-8, see Tables 7.8 and 7.9) and $L2$ (better in TREC-1, TREC-2, TREC-3, TREC-6 and TREC-10, see Tables 7.2, 7.3, 7.4, 7.5 and 7.13). Surprisingly, our experiments with TREC-10 show that $B_EL2$ is the model which best combines with the query expansion technique. Indeed $B_EL2$

with query expansion was the best over all performing run at TREC-10.

4. The model $I(n_e)$ works well with both normalizations $B2$ and $L2$. We observe also that, the $I(n_e)$ performance depends on the choice of the normalization, $B2$ better in TREC-1, TREC-7, TREC-8 and TREC-10 (see Tables 7.2, 7.8, 7.9, and 7.13) or $L2$ which is better in TREC-2, TREC-3 and TREC-6 (see Tables 7.3, 7.4, and 7.5).

5. The model $I(n)$ gives results similarly to $I(n_e)$ but always performs less well than $I(n_e)$.

6. By comparing the results from the models which are approximations or limiting forms of one theoretical basic model, we may observe that they are indistinguishable. We do not need to distinguish between the models $P$ and $D$ for the binomial basic model nor between the models $G$ and $B_E$ for the Bose-Einstein basic model. Similarly, we may observe that $I(F)$ and $I(n_e)$ do not differ significantly in the experiments. Since $I(F)$ can be considered as an approximation of $I(n_e)$, the experiments show that we may reduce the seven basic models ($P$, $D$ $G$, $B_E$, $I(n_e)$, $I(F)$ and $I(n)$) to four: $P$, $G$, $I(n_e)$ and $I(n)$.

7. The term-frequency normalization $H2$ of formula 6.10 seems to be superior to the term-frequency normalization $H1$ of formula 6.9. Indeed, given any model $X \in \{P, G, I(n), I(n_e)\}$ and any normalization $Y \in \{L, B\}$ the model $XY2$ performs better that its analogous $XY1$. There are some partial exceptions especially in the experiment of TREC–6 for high values of recall (Pr@30, Pr@100, Pr@R and for the number of relevant retrieved) as shown in Tables 7.7 and 7.5.

## 7.4 Experiments with short queries

Our results show that all the models are robust with respect to different data sets. Unlike the experiments with long queries, we have used a parameter-based version of the term-frequency normalization **H2**, that is assuming a fixed value $c = 7$. Though this is not the best matching value for each collection, this value was shown to lie within a large

interval of optimal values as shown in Sections 6.2 and 6.2.1 (see Figures 6.2 and 6.3 on page 129 and page 130).

We used two collections the wT2G collection of TREC-8 of 2 GBytes and the collection WT10g of 10 GBytes of TREC-9 and TREC-10. We have already presented in Section 7.3 the collection of TREC-8. The WT10g collection is made up of 1.69 million pages selected from the WEB.

wT10G contains 666 million tokens, 3.09 million unique words and 273.74 million pointers (term–frequency and document pairs). We compressed the inverted file achieving a number 12.60 of bits per pointer for an overall size 411.3 MB of inverted file.

We used only the titles of the TREC-8, TREC-9 and TREC-10 queries. After the application of the stop list the average query–length was 2.6. words.

From the analysis of the results from long queries we could reduce the number of experiments to be presented here. We proved that limiting forms behave similarly and that **H1** was not performing as good as **H2**. The performance of the Zipfian technique **Z** was shown to lie in between **H1** and **H2**. Therefore we have compared only 10 term-weighting models of divergence from randomness $B_E B2$, $B_E L2$, $I(n)B2$, $I(n)L2$, $I(n_e)B2$, $I(n_e)L2$, $I(n_e)B3$ and $DL2$ (=$PL2$) with the language model based on the Dirichlet Priors and the $BM25$.

Notice that $I(n_e)B3$ uses the Dirichlet Priors as term-frequency normalization (see Section 6.4 on page 137).

We submitted at TREC-10 four runs as shown in Table 7.13 to compare retrieval with or without query expansion. The indexing and stemming techniques were different from those used in the previous experiments.

Because of a different IR system used to participate to the conference and because of the size of the collection (10 Gbytes for about 1,600,000 WEB documents), and as we had very limited storage capabilities, we reduced the size of the inverted files and we performed some document and word pruning. Specifically, we indexed with single terms only, ignoring punctuation and case. The whole text was indexed except for HTML tags, which were removed from documents. *Pure single keyword indexing was performed, and link information was not used.* We did some document pruning. We removed 2,897 documents with more than 10,000 words and 57,031 documents with

| TREC 8. | | | | | | | |
|---|---|---|---|---|---|---|---|
| Models | MAP | MAP@10 | Pr @5 | Pr @10 | Pr @20 | Pr@R | RelRet |
| $BB2$ | 0.2616 | 0.3619 | 0.4920 | 0.4580 | **0.4160** | 0.3032 | 2882 |
| $BL2$ | 0.2587 | 0.3548 | 0.4880 | 0.4540 | 0.4040 | 0.3027 | 2834 |
| $I(n)B2$ | **0.2629** | 0.3675 | 0.5080 | **0.4700** | 0.4150 | 0.3046 | **2910** |
| $I(n)L2$ | 0.2606 | 0.3548 | 0.4920 | 0.4560 | 0.4110 | **0.3056** | 2890 |
| $I(n_e)B2$ | 0.2616 | **0.3639** | 0.4920 | 0.4600 | 0.4140 | 0.3019 | 2859 |
| $I(n_e)L2$ | 0.2588 | 0.3518 | 0.4840 | 0.4500 | 0.4060 | 0.3043 | 2823 |
| $PL2$ | 0.2477 | 0.3587 | 0.4880 | 0.4580 | 0.3970 | 0.2967 | 2866 |
| $I(n_e)B3\mu = 1700$ | 0.2509 | 0.3467 | 0.4800 | 0.4520 | 0.3940 | 0.2878 | 2845 |
| $BM25$ | 0.2361 | 0.3509 | 0.4720 | 0.4500 | 0.3960 | 0.2910 | 2776 |
| $LM\mu 300$ | 0.2548 | 0.3461 | **0.5120** | 0.4440 | 0.4000 | 0.3017 | 2862 |

Table 7.10: Baselines for short queries of TREC-8

| TREC 9. | | | | | | | |
|---|---|---|---|---|---|---|---|
| Models | MAP | MAP@10 | Pr @5 | Pr @10 | Pr @20 | Pr@R | RelRet |
| $BB2$ | 0.2029 | 0.2448 | 0.3040 | 0.2620 | 0.2120 | 0.2353 | 1479 |
| $BL2$ | **0.2085** | 0.2459 | 0.3080 | 0.2560 | 0.2080 | **0.2410** | 1547 |
| $I(n)B2$ | 0.1975 | 0.2450 | **0.3160** | **0.2640** | 0.2120 | 0.2320 | 1437 |
| $I(n)L2$ | 0.2067 | **0.2465** | 0.3120 | 0.2560 | 0.2060 | 0.2376 | 1521 |
| $I(n_e)B2$ | 0.1984 | 0.2393 | 0.3040 | 0.2620 | 0.2110 | 0.2331 | 1480 |
| $I(n_e)L2$ | **0.2085** | 0.2456 | 0.3040 | 0.2580 | 0.2100 | 0.2400 | **1553** |
| $PB2$ | 0.1858 | 0.2076 | 0.2640 | 0.2400 | 0.1940 | 0.2201 | 1395 |
| $PL2$ | 0.1939 | 0.2288 | 0.2960 | 0.2580 | 0.2100 | 0.2300 | 1484 |
| $I(n_e)B3\mu = 1600$ | 0.1962 | 0.2382 | 0.3040 | **0.2640** | **0.2180** | 0.2346 | 1456 |
| $BM25$ | 0.1786 | 0.2183 | 0.2880 | 0.2340 | 0.1950 | 0.2131 | 1327 |
| $LM\mu 1300$ | 0.1990 | 0.2210 | 0.3000 | 0.2520 | 0.2070 | 0.2384 | 1529 |

Table 7.11: Baselines for short queries of TREC-9

| TREC 10. | | | | | | | |
|---|---|---|---|---|---|---|---|
| Models | MAP | MAP@10 | Pr @5 | Pr @10 | Pr @20 | Pr@R | RelRet |
| $BB2$ | 0.2105 | 0.3011 | 0.4280 | 0.3720 | 0.3170 | 0.2461 | 2413 |
| $BL2$ | 0.2017 | 0.2870 | 0.4040 | 0.3620 | 0.3090 | 0.2356 | 2348 |
| $I(n)B2$ | 0.2105 | 0.2975 | 0.4240 | 0.3720 | 0.3170 | 0.2454 | 2404 |
| $I(n)L2$ | 0.2041 | 0.2852 | 0.4200 | 0.3560 | 0.3120 | 0.2393 | 2409 |
| $I(n_e)B2$ | 0.2105 | 0.2979 | 0.4200 | 0.3720 | 0.3170 | **0.2473** | 2415 |
| $I(n_e)L2$ | 0.2023 | 0.2870 | 0.4040 | 0.3640 | 0.3100 | 0.2386 | 2353 |
| $PB2$ | 0.1995 | 0.2690 | 0.3800 | 0.3460 | 0.2950 | 0.2340 | 2391 |
| $PL2$ | 0.2065 | 0.2909 | 0.4120 | **0.3740** | 0.3230 | 0.2366 | 2448 |
| $I(n_e)B3\mu = 1200$ | **0.2132** | **0.2983** | **0.4400** | 0.3680 | 0.3240 | 0.2451 | 2458 |
| $BM25$ | 0.1866 | 0.2680 | 0.3800 | 0.3480 | 0.3080 | 0.2285 | 2318 |
| $LM\mu 1200$ | 0.2126 | 0.2837 | 0.4160 | 0.3620 | **0.3250** | 0.2437 | 2443 |

Table 7.12: Baselines for short queries of TREC-10

| Method | Run | MAP | Pr@10 | Pr@20 | Pr@30 |
|---|---|---|---|---|---|
| Model performance without query expansion | | | | | |
| $B_E L2$ | | 0.1788 | 0.3180 | 0.2730 | 0.2413 |
| $I(n)L2$ | | 0.1725 | 0.3180 | 0.2740 | 0.2353 |
| $I(n_e)L2$ | official | 0.1790 | 0.3240 | 0.2720 | 0.2440 |
| $B_E B2$ | | 0.1881 | 0.3280 | 0.2980 | 0.2487 |
| $I(n)B2$ | official | 0.1900 | 0.3360 | 0.2880 | 0.2580 |
| $I(n_e)B2$ | | 0.1902 | 0.3340 | 0.2860 | 0.2580 |
| Model performance with query expansion | | | | | |
| $B_E L2$ | official | 0.2225 | 0.3440 | 0.2860 | 0.2513 |
| $I(n)L2$ | | 0.1973 | 0.3200 | 0.2730 | 0.2380 |
| $I(n_e)L2$ | official | 0.1962 | 0.3280 | 0.2760 | 0.2507 |
| $B_E B2$ | | 0.2152 | 0.3400 | 0.2870 | 0.2527 |
| $I(n)B2$ | | 0.2052 | 0.3380 | 0.2970 | 0.2680 |
| $I(n_e)B2$ | | 0.2041 | 0.3360 | 0.2990 | 0.2660 |

Table 7.13: Comparison of models with TREC-10 data without using Porter's stemming algorithm.

less than 10 words. Also, we removed 86,146 documents containing more than 50% of unrecognized English words. In all, we removed 118,087 documents. Words contained in less than 11 documents, that were apparently exclusively misspelled words, were not included for the indexing. Words containing more than three consecutive equal characters or longer than 20 characters were also deleted. In this way, the number of distinct words in the collection was only 293,484. We used a very limited stop list and did not perform word stemming at all.

### 7.4.1 Results from experiments with short queries

As observed in the previous section we assumed the value $c = 7$ for the normalization **H2**. As shown in Figures 6.2 and 6.3 this setting is not the best matching value for each collection, but this value was shown to be within a large interval of best matching values (see Sections 6.2 and 6.2.1).

In the following we discuss the results shown in Tables 7.10, 7.11, 7.12 and 7.13.

1. Similarly to the results from experiments with long queries, there is no convincing evidence or argument in favour of either normalization $B$ or $L$. it seems that $B$ combines well with the inverse document-frequency based models $I(n)$ and $I(n_e)$, though in TREC-9 $I(n_e)$ was the best performing model. We are not in the position to draw any conclusion for the other models. The exception is the binomial model, for which Laplace's law normalization $L$ provides good results.

2. The Poisson model $PL2$ gave a good performance in TREC-10. Also results from TREC-11 evaluation, where models of divergence from randomness were used, is reported that $PL2$ gave better results than other models in topic distillation task [81]. By contrast, the normalization $B2$ seems to work poorly with $P$.

3. Model $B_E$ with both normalizations $B2$ and $L2$ gave a good performance. It was the best performing model together with $I(n_e)L2$ in TREC-9.

4. Again, the model $I(n)$ gives results similarly to $I(n_e)$.

5. The term-frequency normalization **H3** of formula 6.4 gave the best run for TREC-10 but not with other two set of queries. Also, we are not in the position to draw a conclusion on which method between **H2** and **H3** is the best one.

6. The $BM25$ worked poorly in comparison to the divergence from randomness models and language model (MAP= 0.2361, 0.1786 and 0.1866 against 0.2629, 0.2085 and 0.2132 of the best runs). We will see that the $BM25$ reduces its gap from the other models by using our query expansion technique (see Chapter 8).

7. The language model $LM$ is as effective as the models of divergence from randomness (MAP= 0.2548, 0.1990 and 0.2126 against 0.2629, 0.2085 and 0.2132 of the best runs). However its best performing parameter $\mu$ is not stable (it varies from 300 in TREC-8 to 1300 in TREC-9). This value seems also to depend on the length of the query (see Chapter 8 on query expansion). In contrast, the best performing value of $\mu$ is stable when Dirichlet's priors are used as term-frequency normalizing factors **H3** in the models of divergence from randomness over both the collections and the query–lengths (see Chapter 8 on query expansion).

8. In Table 7.13 there are the results from TREC-10 with the collection indexed without stemming. The first normalization $B2$ is superior to $L2$ with the exception of model $B_E$ in combination with the query expansion.

## 7.5 Conclusions

We have created a framework for generating non-parametric Information Retrieval models. We constructed a weighting formula which is a combination of three different probabilities. The first and basic probability models were obtained from urn models with random drawings. We computed a second probability, the probability of relevance of a term in its "elite set". This provided a normalization factor on the weighting formula. Finally, a probability related to the length of a document was constructed to resize the cardinality of the term-frequency in the document. Four hypotheses about the distribution of document length were tested.

We used the basic probability models to derive for IR, a Bernoulli model, the *tf-idf* model $I(n)$, the *tf-itf* model $I(F)$ and the model $I(n_e)$ which is a combination of the Poisson and the *idf* models. Two workable approximations of Bernoulli's model were introduced: the Poisson model $P$ and the information theoretic approximation model $D$. These two approximation models performed equally under all normalizations.

The other basic model is Bose-Einstein. Two approximations of the Bose-Einstein model were also introduced: the geometric models $G$ and $B_E$. These two approximation models performed equally under all normalizations.

All models were compared with the $BM25$ formula, which is frequently used by many participants of TREC and the language model based on Dirichlet's priors. $I(n_e)B2$ and $I(n_e)L2$ were often shown to be superior at many recall levels and in average precision. Experiments showed that the model $I(n_e)$ and $I(F)$ perform in a similar way. $I(n_e)$ was shown to perform often better than the standard idf model $I(n)$ under all normalizations.

$B2$, $L2$ and $B3$ are shown to be universal normalization factors, in the sense that the normalization works independently of models and independently of variation in document length. $L2$ is less sensitive to the variation of document length. On the other hand, when the variation is moderate $B2$ seems to perform better. The normalization factor $B2$, containing both the document frequency and the term-frequency, derives formally

from Bernoulli process and from the standard axioms of utility theory.

Our models are all formally derived. Parameter-free models work well with long queries, while a stable value for the parameter $c$ was found to hold for short queries.

Finally we have shown that $BM25$ can be formally derived from our framework together with its parameter values.

# Chapter 8

# Query expansion

## 8.1 Introduction

In contrast to the inherent difficulty of representing complex concepts such as for example our information need, which we would like to express by a long and articulate statement, the average length of a query submitted to search engines is typically short.

Although users often formulate very short queries, automatic query expansion is highly effective for many information retrieval tasks. However, automatic query expansion may be detrimental in some situations. If early precision is critical, or if the number of relevant documents is a few, then automatic query expansion may be not rewarding or may even harm the effectiveness of retrieval. In cases where only early precision is required, the employment of query expansion may induce the system to include irrelevant documents high in the ranking. As observed by Harman [47], automatic query expansion can make a gain in recall that is countered by a loss in precision. Notwithstanding these considerations, we show that query expansion performed with our methodology brings in a substantial increment of the mean average precision (MAP). However, MAP increment is not uniform over all queries. Indeed, the average precision drops for approximately one-third of queries. The same proportion was also reported in [80] with different query expansion techniques. The decision has to be taken when performing query expansion as to whether the increase of the mean average precision is more valuable than the loss produced in average precision for a significant number of queries. The decision depends on the type of application, but if the utility function measuring the effectiveness of the

system is MAP or even the early precision, such as the exact precision at 10 documents retrieved, then query expansion seems to be of benefit.

The literature on automatic query expansion and its strictly related subjects, such as relevance feedback, is huge [64, 90, 27, 86, 95, 48, 26, 45, 120, 130, 69, 131, 70]. The basic and most effective strategy for performing query expansion is *local feedback*, also known as *pseudofeedback*. The term local feedback was introduced by Attar and Fraenkel [8] to denote the process of formulating a new improved search based on clustering terms from the documents returned in a previous search. Clustering terms can be computationally expensive because of the size of term-by-term matrices which have to be built with a global statistical analysis. The local feedback technique is able to select a set of terms from the topmost retrieved documents in a first ranking pass. After this phase, the selected terms are added to the original query with a weight. Rocchio's methodology[92] is generally used to compute the weights of the terms in the expanded query. The term-weights for selection and the actual term-weights used for the second ranking pass may not necessarily coincide [88].

A non-probabilistic approach to query expansion is taken by Bruza and Song with the Hyperspace Analogue to Language [15, 16] whereby the strength of an information flow is computed between pairs of queries, conceived as logical concepts, and terms. Hyperspace Analogue to Language is claimed to be as effective as probabilistically motivated expansion model.

In this Chapter we follow Rocchio's approach to define the query expansion model. We introduce a general methodology of query expansion following the leading idea of divergence from randomness as introduced in Section 3. Our approach based on the divergence from randomness is able to explain how Kullback-Leibler divergence (see Section 2.2.3) is connected to the binomial distribution and why it performs similarly to the binomial in the case of query expansion task. Our approach can be seen as a generalization of the approach used by Carpineto and Romano in [19, 17] which applied the Kullback-Leibler divergence to the unexpanded version of $BM25$ [19]. Our framework precisely relates different query expansion formulae, such as the binomial formula, the Poisson, the $\chi$-square and the Bose-Einstein statistics. Results show that this methodology is effective for all probabilistic models of $IR$ from the $BM25$ and the language model

based on the Dirichlet priors to the divergence-based probabilistic models. Finally, we show how to perform query expansion with a parameter-free Rocchio formula.

## 8.2 Term-weighting in the expanded query

For the moment we assume that an arbitrary $IR$ model is used. The model computes a weight $w(t|d)$ for each pair of term $t$ of the vocabulary and document $d$ in the collection. We also assume that the weight $w(q|d)$ of a query $q$ given a document $d$ satisfies two conditions:

1. Let $q = (t_1, \ldots, t_k)$ and let $t_1, \ldots, t_k$ be independent. We assume that the weight is *additive*, namely $w(q|d) = \sum_{i=1}^{k} w(t_i|d)$
2. All tokens in the query are independent, even if they are tokens of the same word. Therefore, the weight $w(q|d)$ of the query given a document is:

$$w(q|d) = \sum_{t \in q} tf_q \cdot w(t|d)$$

where $tf_q$ is the number of occurrences of $t$ in the query $q$.

The document score given a query is thus made up of two components:

- The *term-weight $tf_q$ in the query* . In absence of other evidence, $tf_q$ is taken to be the raw frequency of the term in the query.
- The *term-weight $w(t|d)$ in the document* , called more briefly the *term-weight.*

Query expansion consists of enlarging the set $q$ of initial terms with a superset $q^* \supset q$. A weight $tf_{q^*}$ of the term in the query is associated with each term $t \in q^*$. The final document score $w^*(q^*|d)$ is the new weight of the query given the document:

$$w^*(q^*|d) = \sum_{t \in q^*} (tf_q + \alpha tf_{q^*}) \cdot w(t|d) \tag{8.1}$$

With Equation 8.1, we assume that the process of query expansion does not affect the original term-weight $w(t|d)$ in the document, but it only modifies the component of the term-weight in the query. The value $tf_{q^*}$, to be added to the original $tf_q$ (possibly $= 0$ when the term is new) is computed on the basis of a first pass retrieval. $tf_{q^*}$ will be

in general a real number drawn from the term-frequency observed in a new "Elite" set, that is the set $E_q$ of the topmost retrieved documents.

To introduce the underlying idea of our approach, we assume that $tf_{q^*}$ is a monotonically decreasing function of a probability $p(t|E_q)$. $p(t|E_q)$ is defined to be the probability that the tokens of the term $t$ in the set $E_q$ occur accidentally. Again, the notion of *accidental* occurrence of a term is here explained by a suitable urn model. The balls are word tokens, the successes are all the balls of the same colour drawn from the urn. Once again, we assume that human beings put the word tokens in sequence diverging as much as possible from the way this urn model would instead generate an arbitrary text, that is randomly. The divergence from randomness assumption is equivalent to asserting that only non–informative words possess a distribution fitted by, for example, a binomial process, or by its approximations such as, for example, by a Poisson. It turns out that non-informative words are also non-discriminant, in the sense that the frequency of the term in an arbitrary piece of text is exactly that obtained by chance, that is that predicted by the binomial distribution. Our divergence from randomness idea is similar to that which has conceived the 2-Poisson model of $IR$[52], that in turn contributed to the formulation of the $BM25$ formula.

We have already seen in Chapter 4 that Formula 4.2 on page 81 captures the divergence between the information content of a term and the probability of its frequency following a suitable urn model. Therefore, we can display the *fundamental* equation for query expansion:

$$w^*(q^*|d) = \sum_{t \in q^*} (tf_q - \alpha \cdot \log p(t|E_q)) \cdot w(t|d) \tag{8.2}$$

## 8.3 Query expansion

Let us assume that the weight $w(q|d)$ has produced a first pass ranking of documents. The topmost documents in the ranking are probably relevant and therefore we may regard them as constituting a second "Elite set $E_q$ of documents", that is the set of documents which best describe the content of the query $q$.

We employ the same models which are used to define the retrieval functions $Inf_1$ of

Chapter 3. Like the information content $Inf_1$ of the term in a document, we define the information content $Inf(t|E_q)$ of the term in the elite set $E_q$ of the query. $E_q$ is the set of $r$ documents with the highest weights $Inf(t|E_q)$.

$$Inf(t|E_q) = -\log P(t|E_q)$$

The probability $P(t|E_q)$ is computed by either the binomial or the Bose-Einstein statistics.

Let us show the computation of $P(t|E_q)$ by an example. Let "What is a prime factor?" be a query. After the use of the stop list the query reduces to "prime factor". We produce a first pass document ranking and we use $r = 3$ topmost documents to derive the information content of all terms contained in these documents. The parameter $r$ is set to the same value as in [3]. In general, $r$ is set to much higher value than 3 by most query expansion techniques. For example, the query models of the language model based based on the Markov chains use 50 documents [70] for the training.

Then, we filter the terms according to the condition that $t$ belongs to at least 2 documents of $E_q$. This is a simple application of the hypothesis that a common term from the top-ranked documents tends to co–occur with all query terms within this top-ranked document set [130, 131]. However, we do not use ad hoc co–occurrence metrics for selection, but we interpret this hypothesis as a simple Boolean condition to be satisfied. Our constraint is to avoid the noise which may be generated by very frequent terms appearing in only one single non–relevant document of the Elite set $E_q$. Indeed, the occurrence of a highly informative term in two distinct documents out of 3 topmost retrieved documents would make it quite improbable to belong to both non-relevant documents, especially if the exact precision at 3 is close or greater than 50%, as in actual situation. For example, if the precision at 3 is exactly 0.5 then the prior probability that an arbitrary term, co-occurring into 2 different documents, belongs to at least one relevant document is 87.5%, $(1 - 0.5^3)$. This probability grows to 93.7% if the precision at 3 is 0.6. However, this probability should be much higher, since we have not assumed in our computation the fact that a highly informative term in general is a rare term in the collection, and therefore its probability of occurring in a non relevant document is very low.

After selection, the terms are ordered according to the weights computed by means

of the binomial. The first $\tau = 10$ terms of the ranking are chosen to expand the original query.

Turning back to our example, the term "prime" is used $F_{E_q} = 55$ times out of $TotFr_{E_q} = 1535$ of words used in the first 3 retrieved documents. Its relative frequency in the collection is $p = 6.4 \cdot 10^{-5}$. The probability of obtaining the term-frequency $F_{E_q}$ by chance can be obtained by the binomial law.

$$B(55, 1535, p) = \binom{1535}{55} p^{55} q^{1480} = 1.4 \cdot 10^{-129}$$

where $p = 6.4 \cdot 10^{-5}$ and $q = 1 - p$.

The information content of the term $t$ in the set $E_q$ of the topmost documents is obtained by the logistic function:

$$Inf(t|E_q) = Inf_{E_q}(t) = -\log_2 B(F_{E_q}, TotFr_{E_q}, p) \tag{8.3}$$

For the term "prime" of our example we have

$$Inf_{E_q}(t) = -\log_2 B(55, 1535, 6.4 \cdot 10^{-5}) = 428.04$$

In a similar way we may compute the information content of all terms contained in the first $r = 3$ retrieved documents, under the condition that they must belong to most of these $r$ documents. We regard the most informative terms as good candidates for query expansion. In our example, the first $\tau = 10$ stemmed terms with the highest information content are shown in Table 8.1.

Once the information content of the terms related to the query is computed, we have to face the problem of exploiting these scores to obtain the new weighted expanded query and thus the final document ranking. One approach [19], is simply to add the first $r$ informative terms to the original query $q$, obtaining thus a new query $q^*$. Then, the information content values are normalized to a value $nInf_{E_q}(t)$ less or equal to 1. The information content values are normalized by their maximum

$$M = \arg_{t \in Q} \max Inf_{E_q}(t)$$

or by a maximal value $M$. Finally, Rocchio's formula is derived, that is the new weighted query is a linear combination of the original query vector and the normalized information

| term | $tf_q$ | $F_{E_q}$ | $p$ | $Inf_{E_q}$ | $nInf_{E_q}$ | $tf_q + 0.5 \cdot nInf_{E_q}$ |
|---|---|---|---|---|---|---|
| prime | 1 | 55 | $6.4 \cdot 10^{-5}$ | 428.04 | 1.0000 | 1.5000 |
| number | 0 | 99 | $1.4 \cdot 10^{-3}$ | 412.48 | 0.9636 | 0.4818 |
| factor | 1 | 49 | $1.83 \cdot 10^{-4}$ | 299.67 | 0.7001 | 1.3500 |
| integ | 0 | 30 | $4.36 \cdot 10^{-5}$ | 225.19 | 0.5261 | 0.2630 |
| primal | 0 | 8 | $3.17 \cdot 10^{-6}$ | 76.77 | 0.1794 | 0.0896 |
| multipl | 0 | 15 | $1.78 \cdot 10^{-4}$ | 68.74 | 0.1606 | 0.0802 |
| test | 0 | 21 | $6.24 \cdot 10^{-4}$ | 68.28 | 0.1595 | 0.0797 |
| divid | 0 | 11 | $6.28 \cdot 10^{-5}$ | 62.53 | 0.1461 | 0.0730 |
| common | 0 | 15 | $2.65 \cdot 10^{-4}$ | 60.34 | 0.1410 | 0.0704 |
| composit | 0 | 9 | $2.62 \cdot 10^{-5}$ | 60.26 | 0.1408 | 0.0703 |

Table 8.1: The highest informative terms for the query 502 (Prime factor?) of TREC-10 data. The last column shows the weights of the terms in the new expanded query.

content term-vector. If $\overrightarrow{tf_q}$ is the original query and

$$nInf_{E_q} = \frac{\overrightarrow{Inf}_{E_q}}{M}$$

is the normalized information content term-vector, then the new weighted query $\overrightarrow{q^*}$ is:

$$\overrightarrow{q^*} = \overrightarrow{tf_q} + \alpha \cdot \overrightarrow{nInf}_{E_q} \tag{8.4}$$

$$\alpha \leq 1$$

With this approach the term-weighting formulae are not modified. Indeed the information content is used to weight a new query constituting thus an independent component of the system. In addition, an independent query expansion component may be easily combined with any arbitrary term-weighting formula, unexpanded $BM25$ or language model included.

The final term-weighting is thus provided by the following formula:

$$w^*(t|d) = \left(tf_q + \alpha \cdot \frac{Inf(t|E_q)}{M}\right) \cdot w(t|d) \tag{8.5}$$

In the following we exploit approximations of function 8.3 with the limiting forms described in Section 2.2. Among the approximations we use the Poisson process, the $\chi^2$

statistics, the asymmetric Kullback-Leibler divergence function, the information theoretic divergence function $D$. All of these limiting forms are equivalent in terms of performance with the exception of the $\chi^2$. They are also equivalent in theoretical terms up to an approximation error and to a proportional factor. This proportional factor is independent of the term, and thus all expansion methods are equivalent to the binomial for term-ranking up to an approximation error.

In addition, we also use the Bose Einstein statistics and show that Bose Einstein statistics performs in a similar way of the binomial.

## 8.4 Rocchio's method

A simple and commonly used method of query expansion is due to Rocchio [92]. Actually, the Rocchio method was designed to process the relevance feedback and provides a measure for the selection of query expansion terms as follows:

$$tf_q^* = \alpha \cdot tf_q + \beta \sum_{d_k \in E_q^r} \frac{w(t|d_k)}{|E_q^r|} - \gamma \sum_{d_k \in E_q - E_q^r} \frac{w(t|d_k)}{|E_q - E_q^r|} \quad \text{[Rocchio]} \tag{8.6}$$

where $tf_q$ is the original weight of the term in the query, and $E_q$ is the set of the retrieved documents, and $w(t|d_k)$ is the term-weight within the $k$-th retrieved and relevant or non-relevant document, as assigned in Section 8.2, and $E_q^r$ is the set of relevant and retrieved documents, and $\alpha$, $\beta$ and $\gamma$ are parameters. This formula can be used both for selecting terms and weighting terms in the new expanded query.

In local or blind feedback, that is assuming $E_q = E_q^r$ we use a simplified version of the Rocchio formula:

$$tf_q^* = \alpha \cdot tf_q + \beta \sum_{d_k \in E_q} \frac{w(t|d_k)}{|E_q|} \quad \text{[Rocchio]} \tag{8.7}$$

Notice that we may set $\alpha + \beta + \gamma = 1$ for the scores would be equal up to a proportional factor and thus the ranking would be the same. Rocchio's method contains three parameters to be estimated, that is $\beta$, the number of topmost documents to be processed, and the number of terms to be added to the query.

Notice that Formula 8.5 is more general from the standard definition of Rocchio's formula. Rocchio's formula computes the mean of within–document term–weights $w(t|d_k)$

in the set of pseudo relevant documents and this value is averaged with the original within–query term–weight $tf_q$ according to the priors $\alpha + \beta = 1$. Then, this averaged mean is matched against the within–document term–weights $w(t|d)$.

Formula 8.5 computes the informative content of the terms in the set of pseudo relevant documents and this value is averaged with the original term-within query–weight $tf_q$ according to the priors $\alpha$ and $\beta$. Then, this value is matched against the within–document term–weights $w(t|d)$ obtaining the final second pass ranking.

## 8.5 The Binomial Law for query expansion

Let $\mathcal{P}_D = (p_D, q_D)$ be the relative term-frequency of a term $t$ in the collection $D$, that is

$$p_D = \frac{F}{TotFr_D} \qquad\qquad q_D = 1 - p_D$$

The probability $p_D$ is the *a priori* probability of occurrence of $t$ in $D$.

Let $E$ be a subset of $D$ and let

$$p_E = \frac{F_E}{TotFr_E}$$

be the frequency of the term in $E$. We regard each occurrence of a word in $E$ as a trial of an *experiment.* A successful outcome for the term $t$ is when $t$ occurs. Then the text of $E$ becomes a sequence of trials. The *a priori* probability of obtaining $F_E$ successes over $TotFr_E$ trials (the total number of occurrences of words in $E$), with prior $\mathcal{P}_D = (p_D, q_D)$ can be modeled by Bernoulli process:

$$B(F_E, TotFr_E, p_D) = \binom{TotFr_E}{F_E} p_D^{F_E} q_D^{TotFr_E - F_E} \tag{8.8}$$

If $E$ is large and randomly chosen, then $E$ can be considered as a *sample* of the entire collection $D$. In this case, for any term $t$ the frequency $P_E$ should be close to its prior $p_D$. To see this one can use the Chernoff bounds: a deviation $\varepsilon$, with $0 < \varepsilon < 1$ from the average number of successes $TotFr_E \cdot p_D$ in $TotFr_E$ experiments in a Bernoulli process is analyzed by observing the tail probability

$$P\left(\left|\frac{P_E}{P_D} - 1\right| > \varepsilon\right) = \sum_{\left|\frac{p_E}{p_D} - 1\right| > \varepsilon} B(F_E, TotFr_E, p_D) \leq 2e^{-\varepsilon^2 p_D \cdot TotFr_E} \tag{8.9}$$

The tail in the inequality of Formula 8.9 (see also Formula 3.7 ) becomes definitely small as soon as the size of $E$ becomes large. However, when we expand queries, $E$ is neither large nor chosen by a random process. $E$ is rather chosen according to the document ranking, that is when $E$ is the Elite set $E_q$ of the query.

In the case of query expansion, when $E = E_q$, the two frequencies $p_{E_q}$ and $p_D$ should diverge and not converge as in the limiting process of sample selection. More generally, notwithstanding $E_q$ is obtained by a query driven document selection, if a term $t$ still shows a distribution $\mathcal{P}_{E_q}$ not much dissimilar to the prior $\mathcal{P}_{\mathcal{D}}$, then we assume that the term $t$ is not a discriminant term of $E$ from the entire collection $D$ and thus it is not a good descriptor of the content of the query. Only those terms $t$, for which the probability distribution $\mathcal{P}_{E_q}$ diverges from $\mathcal{P}_{\mathcal{D}}$, should have a significant weight in the expanded query. In different words, if the distribution $\mathcal{P}_{E_q}$ generates a large tail, then the term is a discriminant, and equivalently, the probability of frequency $p_{E_q}$ given by Bernoulli process of Formula 8.8 should be very small.

The probability $B(F_{E_q}, TotFr_{E_q}, p_D)$ defined by a *Bernoulli process* is thus inversely related to the information content of the term in the set $E_q$. Once again, as in Chapter 3, we assume that the information content $Inf(t|E_q)$ is inversely proportional to $B(F_{E_q}, TotFr_{E_q}, p_D)$.

$$Inf(t|E_q) = -\log_2 B(F_{E_q}, TotFr_{E_q}, p_D)$$

## 8.6 The hypergeometric model of query expansion

This section is a straightforward application of Section 2.3 on page 55.

A different model of query expansion can be defined by using the hypergeometric distribution. The hypergeometric distribution is generated by a process of sampling from an urn (Type I) without replacement of the extracted balls. There is a population $D$ of $TotFr_D$ tokens having a number $F$ of tokens of the same word $t$. A sample $E$ of $D$ is given. For query expansion $E$ is the set of topmost documents in a first pass document ranking. A number $F_E$ of tokens of the same word $t$ is observed in the sample $E$. The hypergeometric distribution defines the probability $P(F_E|D, E)$ of observing exactly $F_E$ tokens in the sample assuming that the sample was chosen at random. Since $E$ is the set

of topmost documents, the hypergeometric distribution provides a measure of divergence from randomness of the sample with respect to a given word.

In the hypergeometric distribution we do not replace the balls drawn from the urn as we do for the binomial model. In the process of drawing balls from the urn without replacement, the probability of observing a given word frequency from a population is computed by counting the number of possible exchangeable combinations of tokens having the given word frequency out of all possible combinations. The probability of the given word frequency is Formula 2.22 of Section 2.3. However, as shown in Section 2.3, the limiting form of the hypergeometric distribution for query expansion is still the binomial model of query expansion defined in Section 8.5. From Equation 2.23 on page 56 we obtain:

$$-\log_2 P(F_E|D,E) = -\log_2 \binom{TotFr_E}{F_E} p_D^{F_E} q_D^{TotFr_E - F_E} \left(1 - \frac{TotFr_E}{TotFr_D}\right)^{-TotFr_E}$$

$$(8.10) \qquad Inf(t|E_q) = -\log_2 B(F_{E_q}, TotFr_{E_q}, p_D) - TotFr_E \cdot \log_2 \left(1 - \frac{TotFr_E}{TotFr_D}\right)$$

Since the second expression of the sum 8.10 does not depend on the term but on the size of the collection and the sample, the term-weights within the query are those from the binomial model up to a constant.

From this fact we can say that for large populations and small samples we may regard all tokens of $E_q$ as independent trials with fixed probability of success. This is the same remark made by Feller[37, page 59], who observed that for large populations there is no practical difference between sampling with or without replacement.

### 8.6.1 Approximations of the Binomial

We use some of the approximations of the Bernoulli process used in [5] and described in Section 2.2.

**The approximation of Bernoulli's process *via* the divergence function**

Formula ( 8.8) can be rewritten as Formula 2.15 of Section 2.2:

$$(8.11) \qquad B(F_{E_q}, TotFr_{E_q}, p_D) = \frac{2^{-TotFr_{E_q} D(p_{E_q}, p_D)}}{(2\pi TotFr_{E_q}(1-p_{E_q}))^{\frac{1}{2}}}(1 + O(\frac{1}{TotFr_{E_q}}))$$

where $D(p_{E_q}, p_D) = p_{E_q} \cdot \log_2 \frac{p_{E_q}}{p_D} + (1-f) \cdot \log_2 \frac{(1-p_{E_q})}{(1-p_D)}$, $p_{E_q}$ and $p_D$ are the frequencies of the term in the subset $E_q$ and in the collection $D$ respectively, as introduced in Section 8.5. The information content is then obtained from Formula 8.3 and by Formula 2.16:

$$Inf_{E_q}(t) = TotFr_{E_q} D(p_{E_q}, p_D) + \frac{1}{2} \log_2(2\pi TotFr_{E_q}(1-p_{E_q})) \quad [Bi] \tag{8.12}$$

The approximation error is $+O(\frac{1}{TotFr_{E_q}})$.

**The Kullback-Leibler divergence approximation**

From Formula 8.12, if $c = \frac{1}{2}\log_2 2\pi$ then:

$$Inf_{E_q}(t) = TotFr_{E_q} D(p_{E_q}, p_D) + \frac{\log_2 TotFr_{E_q} + c}{2} + O((\frac{1}{TotFr_{E_q}})^2)$$

Since $\frac{\log_2 TotFr_{E_q}+c}{2}$ is independent of the term $t$, then its contribution in the sum is a constant. Therefore, the information content can be supposed to be proportional to:

$$Inf_{E_q}(t) \sim D(p_{E_q}, p_D)$$

Moreover the contribution $(1-p_{E_q}) \log_2 \left(\frac{1-p_{E_q}}{1-p_D}\right)$ in $D(p_{E_q}, p_D)$ is very small and negative, because we may in general assume that $p_{E_q} > p_D$. Thus, as obtained in Section 2.2.3 we derive a further approximation of the Bernoulli process:

$$Inf(t|E_q) \sim p_{E_q} \cdot \log_2 \frac{p_{E_q}}{p_D} \quad [KL] \tag{8.13}$$

which is the asymmetric Kullback-Leibler divergence.

**The $\chi^2$ divergence weighting formula**

Then the divergence $D$ of Formula ( 8.13), From Equation 2.21, is approximated as follows:

$$D(p_{E_q}, p_D) \sim \frac{\log_2 e}{2} \cdot \chi^2(\mathcal{F}, \mathcal{P}) \quad [\chi] \tag{8.14}$$

The error of the approximation of $D(p_{E_q}, p_D)$ is $O((f_i - p_i)^3)$. This error must be added to the error of the approximation of the binomial with $D$. Hence the error of using $\chi^2$ as

an approximation of the binomial is larger, than errors produced by the Formulae 8.12 and 8.13, which is is confirmed by our experimental results.

$$Inf(t|E_q) \propto \frac{\log_2 e}{2} \cdot \chi^2(\mathcal{F}, \mathcal{P}) \quad [\chi] \tag{8.15}$$

## 8.7 Query expansion with the Bose-Einstein distribution

The computation of the information content of a term in the term-weighting formula does not differ from the computation of the information content of a term in the expanded query. In fact, the mean is $\lambda_1 = \frac{F_t}{N}$ in the case of the Bernoulli model $D$ of IR, while in the Poisson model of the query expansion the mean is $\lambda_2 = TotFr_{E_q} \cdot \frac{F_{E_q \cdot}}{TotFr_D}$ .

This analogy suggests us to use the other urn model for IR to obtain alternative methods of expansion for the query , that is the Bose-Einstein statistics. We have seen that one possible approximation of the Bose-Einstein statistics is given by the *geometric distribution G*. The probability

$$p = \frac{1}{1+\lambda}$$

generating the geometric distribution has the same parameter $\lambda = \frac{F_t}{N}$ as the Poisson process. The urn model based on $B_E$ of Formula 2.31 can be thus used for measuring the information content of terms in the query expansion process giving us:

$$Inf_{E_q}(t) = -\log_2\left(\frac{1}{1+\lambda_{E_q}}\right) - F_{E_q} \cdot \log_2\left(\frac{\lambda_{E_q}}{1+\lambda_{E_q}}\right)$$

$$\lambda_{E_q} = \begin{cases} \frac{F_{E_q}}{N} & [Bo1] \\ TotFr_{E_q} \cdot \frac{F_{E_q \cdot}}{TotFr_D} & [Bo2] \end{cases} \tag{8.16}$$

## 8.8 Normalized term-frequency in the expanded query

We have used Formula 8.4 to obtain a *virtual* term-frequency within the query for each of the 10 terms with the highest information content, that were extracted from the first three retrieved documents. An alternative way of obtaining the term-frequency in the new query can be computed as follows. One possible upper bound of the information content described by Formula 8.12 can be obtained by observing that the divergence is

maximum when $F_{E_q} = F_t$. In such a case:

$$Inf_{E_q}(t) \leq F_{E_q} \cdot \log_2 \frac{TotFr_D}{TotFr_{E_q}}$$

$$M = \arg_{t \in T} maxF_{E_q} \cdot \log_2 \frac{TotFr_D}{TotFr_{E_q}} = (\arg_{t \in T} maxF_{E_q}) \cdot \log_2 \frac{TotFr_D}{TotFr_{E_q}}$$

The expanded query becomes:

$$\vec{q^*} = \vec{tf_q} + \frac{\vec{Inf}_{E_q}}{M} \quad [BM] \tag{8.17}$$

## 8.9 Experiments with query expansion

For the sake of space we could not report all possible experiments with all 56 models introduced so far by our theoretical framework. We have compared only 10 term-weighting models $B_EB2$, $B_EL2$, $I(n)B2$, $I(n)L2$, $I(n_e)B2$, $I(n_e)B3$, $I(n_e)L2$, $DL2$ (=$PL2$), the language model based on the Dirichlet Priors and the BM25 formula with the 6 information content formulae for query expansion: 8.12, 8.13, 8.15, [Bo1] and [Bo2] of 8.16 and 8.17. We used two different collections and 3 sets of TREC queries, that is TREC 8, TREC 9 and TREC 10 data.

We used only the titles of the queries. The parameters of the query expansion are three: the parameter $\alpha$ for combining the information content with the term-frequency $tf_q$ in the query, the number of retrieved documents to be used for term extraction and finally the optimal number of extracted terms which are added to the original query. For sake of space we show in the results just the best value for $\alpha$ and a variant for the normalization which eradicates the parameter $\alpha$. The number of documents processed for each query is also set to 3 and the number $\tau$ of terms added to the query is 10. The same choice, $\tau = 10$ was taken at TREC-10 for the official runs of Okapi.

These parameter values are optimal for short queries for both the collection of 2 GB of TREC 8 and the collection wT10g of 10 GBytes used for TREC 9 and TREC 10 queries.

## 8.10 Results from query expansion

For the language model with Dirichlet priors we tried different values for the parameter $\mu$. The best performing value with the expanded queries is $\mu = 300$ for all TREC

collections. In the case of the original queries, $\mu = 300$ was still the best performing value for TREC 8, whilst $\mu = 1200$ was for the WT10g collection.

Beside the Mean Average Precision we used the Mean Average Precision at 10 documents retrieved (MAP@10). MAP@10 is defined as MAP except that the average precision is computed only for the first 10 retrieved documents in the ranking and is normalized by the minimum between 10 and the number of relevant documents existing in the collection for the query. In the case that two rankings have the same Pr@10, MAP@10 provides further information on the quality of the ranking, since it considers the position of the relevant documents and also the recall for very specific queries, that is in the cases where the number of relevant documents is less than 10 (and in such cases MAP@10 can be greater then Pr@10, see the behaviour of the model $I(n)L2$ in TREC 9).

The results are summarised in Tables 8.2, 8.3, 8.4.

The runs which gave the best Mean Average Precision (MAP) are:

TREC 8. $I(n)B2$ with $Bo2$ expansion (0.2904) and $BB2$ with $Bo2$ expansion (0.2880).

TREC 9. $B_EL2$ with $Bo2$ expansion (0.2256) and $I(n_e)L2$ with $Bo2$ expansion (0.2254).

TREC 10. $I(n_e)B2$ with $BM$ expansion (0.2528) and $LM(\mu = 300)$ with $Bo2$ expansion (0.2513).

The best runs with Precision at 10 are:

TREC 8. $I(n)B2$ with $BM$ expansion (0.4880) and $I(n)L2$ with $KL$ expansion (0.4860).

TREC 9. $B_EL2$ with $Bi$ expansion (0.2820) and $I(n_e)L2$ with $Bi$ expansion (0.2800).

TREC 10. $BM25$ with $Bi$ expansion (0.4280) and $I(n_e)B2$ with $KL$ expansion (0.4160).

1. All expansion methods, except the divergence $\mathcal{X}^2$, work similarly and they do not differ from one another significantly.

2. The parameter $\mu$ of the language model for the Dirichlet priors must be tuned when the original query is expanded (greater value with shorter queries).

   Dirichlet priors can also be used to define an alternative term-frequency normalization function for all models of randomness (see Section 6.4). Unlike the language model, this usage of the Dirichlet priors, as used in the model $I(n_e)B3$ has shown the same best performing value for the parameter $\mu$ for both non–expanded and expanded query.

3. Expansion methods based on Bose-Einstein statistics are good choices to improve the Mean Average Precision. Dirichlet priors of the language model show the biggest increase +17.5% of MAP in comparison to an average of +12% (+30% of increment in TREC 10 with Bo2 in comparison to an average of about 17%). Also the increment of MAP for $BM25$ is greater (+14.1%) than that observed for other models, especially if the Bose-Einstein statistics expansion is performed.

4. Kullback-Leibler and the binomial expansions are most effective for achieving a good early precision.

5. The parameter–free expansion method $BM$, based on the binomial performs similarly to both the parameter-based version of the binomial and to the Kullback-Leibler approximation.

## 8.11 Conclusions

We derived general models for query expansions using the binomial and the Bose-Einstein statistics. The binomial was approximated through the information theoretic divergence $D$ which lead to several further approximations, the Kullback-Leibler asymmetric divergence and the $\chi$-square divergence. The Bose-Einstein statistics produced two different formulae to be used as distribution means. The query expansion method required three parameters: a) the number $r$ of documents for learning the new query terms, b) the

| | | Expansion methods | | | | | |
|---|---|---|---|---|---|---|---|
| | Baseline | B | KL | $\mathcal{X}^2$ | Bo1 | Bo2 | BM |
| | | | MAP | | | | |
| TREC 8 | 0.2547 | 0.2786 | 0.2787 | 0.2671 | 0.2805 | 0.2801 | 0.2768 |
| TREC 9 | 0.1981 | 0.2126 | 0.2128 | 0.2061 | 0.2135 | 0.2133 | 0.2129 |
| TREC 10 | 0.2052 | 0.2403 | 0.2385 | 0.2296 | 0.2417 | 0.2416 | 0.2387 |
| Average | 0.2193 | 0.2438 | 0.2433 | 0.2343 | 0.2452 | 0.2450 | 0.2428 |
| | | | MAP @10 | | | | |
| TREC 8 | 0.3548 | 0.3819 | 0.3833 | 0.3694 | 0.3815 | 0.3843 | 0.3813 |
| TREC 9 | 0.2356 | 0.2600 | 0.2604 | 0.2476 | 0.2595 | 0.2545 | 0.2600 |
| TREC 10 | 0.2873 | 0.3282 | 0.3256 | 0.3127 | 0.3255 | 0.3272 | 0.3242 |
| Average | 0.2925 | 0.3233 | 0.3231 | 0.3099 | 0.3222 | 0.3220 | 0.3218 |
| | | | Prec at 10 | | | | |
| TREC 8 | 0.4553 | 0.4683 | 0.4688 | 0.4605 | 0.4642 | 0.4678 | 0.4663 |
| TREC 9 | 0.2555 | 0.2622 | 0.2637 | 0.2503 | 0.2615 | 0.2582 | 0.2642 |
| TREC 10 | 0.3628 | 0.3917 | 0.3925 | 0.3775 | 0.3867 | 0.3850 | 0.3880 |
| Average | 0.3579 | 0.3741 | 0.3750 | 0.3628 | 0.3708 | 0.3703 | 0.3728 |

Table 8.2: Precision obtained by different expansion methods averaged over all models and TREC collections.

| | | Expansion methods | | | | | |
|---|---|---|---|---|---|---|---|
| | Baseline | B | KL | $\mathcal{X}^2$ | Bo1 | Bo2 | BM |
| | | | MAP | | | | |
| TREC 8 | 0.2547 | 9.4% | 9.4% | 4.9% | 10.1% | 10.0% | 8.7% |
| TREC 9 | 0.1981 | 7.3% | 7.4% | 4.1% | 7.8% | 7.7% | 7.5% |
| TREC 10 | 0.2052 | 17.1% | 16.2% | 11.9% | 17.8% | 17.8% | 16.3% |
| Average | 0.2193 | 11.2% | 10.9% | 6.8% | 11.8% | 11.7% | 10.7% |
| | | | MAP @10 | | | | |
| TREC 8 | 0.3548 | 7.6% | 8.0% | 4.1% | 7.5% | 8.3% | 7.5% |
| TREC 9 | 0.2356 | 10.4% | 10.6% | 5.1% | 10.2% | 8.0% | 10.4% |
| TREC 10 | 0.2873 | 14.2% | 13.3% | 8.8% | 13.3% | 13.9% | 12.8% |
| Average | 0.2925 | 10.5% | 10.4% | 5.9% | 10.1% | 10.1% | 10.0% |
| | | | Prec at 10 | | | | |
| TREC 8 | 0.4553 | 2.9% | 3.0% | 1.1% | 1.9% | 2.7% | 2.4% |
| TREC 9 | 0.2555 | 2.6% | 3.2% | -2.0% | 2.3% | 1.0% | 3.4% |
| TREC 10 | 0.3628 | 7.9% | 8.2% | 4.0% | 6.6% | 6.1% | 6.9% |
| Average | 0.3579 | 4.5% | 4.8% | 1.4% | 3.6% | 3.5% | 4.2% |

Table 8.3: Increment of precision obtained by different expansion methods averaged over all models and TREC collections.

| Models | MAP | | | | Prec@10 | | | |
|---|---|---|---|---|---|---|---|---|
| | Unex. | Exp. Met. | MAP | % | Unex. | Exp. Met. | Prec @10 | % |
| | TREC 8 | | | | | | | |
| $BB2$ | 0.262 | Bo2 | 0.288 | 10.1% | 0.458 | KL | 0.482 | 5.2% |
| $BL2$ | 0.259 | Bo1 | 0.282 | 9.2% | 0.454 | Bo2 | 0.474 | 4.4% |
| $I(n)B2$ | **0.263** | Bo2 | **0.290** | 10.5% | **0.470** | BM | **0.488** | 3.8% |
| $I(n)L2$ | 0.261 | Bo1 | 0.288 | 10.4% | 0.456 | KL | 0.486 | 6.6% |
| $I(n_e)B2$ | 0.262 | Bo2 | 0.288 | 9.9% | 0.460 | KL | 0.480 | 4.3% |
| $I(n_e)B3$ ($\mu = 1600$) | 0.251 | Bo2 | 0.269 | 7.2% | 0.454 | Bo2 | 0.444 | -2.2% |
| $I(n_e)L2$ | 0.259 | Bo2 | 0.282 | 8.9% | 0.450 | Bo2 | 0.472 | 4.9% |
| $PL2$ | 0.248 | Bo1 | 0.280 | **12.8%** | 0.458 | BM | 0.476 | 3.9% |
| $BM25$ | 0.236 | Bo1 | 0.266 | 12.6% | 0.450 | Bo1 | 0.466 | 3.6% |
| $LM$ ($\mu = 300$) | 0.255 | Bo1 | 0.287 | 12.4% | 0.444 | KL | 0.486 | **9.5%** |
| | TREC 9 | | | | | | | |
| $BB2$ | 0.203 | Bo1 | 0.210 | 3.5% | 0.262 | KL | 0.262 | 0.0% |
| $BL2$ | 0.208 | Bo2 | **0.226** | 8.3% | 0.256 | B | **0.282** | 10.2% |
| $I(n)B2$ | 0.198 | KL | 0.216 | 9.5% | **0.264** | BM | 0.278 | 5.3% |
| $I(n)L2$ | 0.207 | Bo1 | 0.223 | 8.0% | 0.256 | BM | 0.276 | 7.8% |
| $I(n_e)B2$ | 0.198 | Bo1 | 0.220 | 10.7% | 0.262 | BM | 0.276 | 5.3% |
| $I(n_e)B3$ ($\mu = 1600$) | 0.196 | Bo2 | 0.220 | **12.1%** | **0.264** | KL | 0.270 | 2.3% |
| $I(n_e)L2$ | **0.209** | Bo2 | 0.225 | 8.1% | 0.258 | B | 0.280 | 8.5% |
| $PL2$ | 0.194 | Bo1 | 0.206 | 6.4% | 0.258 | BM | 0.246 | -4.7% |
| $BM25$ | 0.179 | Bo1 | 0.188 | 5.3% | 0.234 | KL | 0.248 | 6.0% |
| $LM$ ($\mu = 300$) | 0.192 | Bo1 | 0.211 | 10.2% | 0.234 | Bo1 | 0.262 | **12.0%** |
| $LM$ ($\mu = 1200$) | 0.199 | BM | 0.206 | 3.9% | 0.254 | BM | 0.236 | -7.1% |
| | TREC 10 | | | | | | | |
| $BB2$ | 0.211 | B | 0.251 | 19.2% | 0.372 | KL | 0.412 | 10.8% |
| $BL2$ | 0.202 | B | 0.234 | 15.9% | 0.362 | B | 0.384 | 6.1% |
| $I(n)B2$ | 0.211 | Bo1 | 0.249 | 18.1% | 0.372 | KL | 0.390 | 4.8% |
| $I(n)L2$ | 0.204 | Bo2 | 0.251 | 22.9% | 0.356 | KL | 0.402 | 12.9% |
| $I(n_e)B2$ | 0.211 | KL | **0.253** | 20.1% | 0.372 | KL | 0.416 | 11.8% |
| $I(n_e)B3$ ($\mu = 1600$) | 0.212 | Bo1 | 0.246 | 16.0% | 0.360 | B | 0.378 | 5.0% |
| $I(n_e)L2$ | 0.202 | Bo2 | 0.232 | 14.7% | 0.364 | KL | 0.386 | 6.0% |
| $PL2$ | 0.207 | Bo2 | 0.239 | 15.5% | **0.374** | B | 0.380 | 1.6% |
| $BM25$ | 0.187 | Bo2 | 0.232 | 24.4% | 0.348 | B | **0.428** | **23.0%** |
| $LM$ ($\mu = 300$) | 0.193 | Bo2 | 0.251 | **30.0%** | 0.352 | B | 0.408 | 15.9% |
| $LM$ ($\mu = 1200$) | **0.213** | Bo1 | 0.240 | 12.8% | 0.362 | Bo1 | 0.382 | 5.5% |
| | TREC 8, TREC 9 and TREC 10 | | | | | | | |
| Average | 0.219 | | 0.246 | 12.0% | 0.358 | | 0.379 | 5.7% |

Table 8.4: Best expansion methods for each model and TREC collection. The best values for each TREC data are in bold.

number $\tau$ of terms to add to the query, c) the parameter $\alpha$ combining the original query terms and the term-weights in the expanded query. For short queries (with an average length of 2.4), for two different collections and different set of queries, $\tau = 10$ and $r = 3$ were shown to be optimal. The best performing choice for the parameter $\alpha$ was 0.5. A query dependent value $M$ was instead computed as a substitution for $\alpha$ and it provided similar performance. All models, with the exception of the $\chi$-square method, performed similarly, though a slight preference goes to the Bose-Einstein statistics for improving the MAP measure, and to the binomial based methods for improving early precision (MAP@10 and Pr@10).

# Chapter 9

# Conclusions

This chapter reviews the contributions of this work and discusses a number of directions for future investigation.

## 9.1 Summary of the results from the experiments

We ran experiments with very long queries and with short queries with and without query expansion. The experiments were conceived to evaluate the following features of the theoretical framework:

1. *The robustness of models with respect to the term independence assumption.*

   The additivity property of the term-weighting function may cause a deterioration of the effectiveness with the long queries. The experiments with long queries were dedicated to test the effects of additivity on performance. From the experiments we demonstrated that our framework generates many different models which are very robust and do not suffer from the *term independence assumption.*

2. *The consistency of the three components of the theoretical framework in the construction of the basic models of divergence from randomness.*

   From the cross comparison of all models we drew conclusions about the effects which the single basic models and the normalization processes may have on per-

formance. The three components were shown to be each highly effective, and the results largely corroborated the theory underpinning each component of the framework.

3. *The competitiveness of the models of IR based on divergence from randomness with respect to the commonly used IR models, such as the $BM25$ formula and the language model.*

   We showed that our models achieved often the best performance. In particular, our framework produced the best run at the TREC-10 evaluation conference.

4. *The effectiveness of the parameter-free models of divergence from randomness.*

   The parameter-free models showed the best performance with long query. The standard length used in the term-frequency normalization component was set to a larger value than the average length for short or moderately long queries.

   The parameter of the standard document length was theoretically motivated and the best match value we can choose lies within a stable, that is independent of the collection, and quite large interval.

5. *The effectiveness of the query expansion technique based on the basic models of divergence from randomness.*

   The idea of divergence from randomness was applied to offer a solution to the problem of expanding a query. We showed that the performance of any IR model, language model and $BM25$ included, was largely improved using only a few documents and a few newly added terms.

## 9.2 Research Contributions and Future Research

1. The theory of eliteness of Harter was revisited. This theory was related to automatic indexing and consisted in fitting the empirical data to a 2-Poisson distribution. In this dissertation eliteness was explained by the notion of informative content of a term within a document. The informative content measures the divergence of the term-frequency distribution from randomness. In order to relax

the *term independence assumption*, the informative content has been normalized with a probabilistic process, which is related to the aftereffect of sampling from the set of the elite documents of the query. In statistics the apparent aftereffect is illustrated by the theory of accidents. If accidents, like rare terms, do occur frequently, then there must be a cause explaining the proneness to have accidents. The apparent aftereffect of sampling is measured by a conditional probability. We have tested two main possible explanations of the aftereffect of sampling, that is Laplace's law of succession and a ratio of two Bernoulli processes. The apparent aftereffect measure is used to compute the risk involved in the decision of accepting a term as a descriptor of the observed document. The risk function is then used to compute the portion of informative content gained with the term. The gain provides the weighting score of our functions.

2. We used the urn models to introduce the models of divergence from randomness. Urn models were not used as a metaphor to exemplify the inference process in Information Retrieval but they were used systematically to derive the models of divergence from randomness. We have first made clear what rules we would have applied for deriving our models. Weighting formulas were not displayed and motivated *a posteriori* on the basis of the evaluation results. We gave great attention to the basic definitions of space of events, possible outcomes, experiments, trials in the context of Information Retrieval. We did not rush to experiment novel ways to combine the observables of IR following arbitrary heuristic reasoning.

We first investigated how to represent the terms and the documents, how to connect these entities with the notion of sampling in the context of IR. A new theory of IR was found. Our theory has a unifying view of the processes involved in IR. For example the query expansion was reduced to the same process described by the basic models of IR. We have demonstrated a tight connection of our models with a new paradigm of IR modelling, the language modelling. Indeed we have shown that any language model can be used, with a more steady parameter, as a second normalization component of our models. This is a further item of evidence that our framework is sound and robust.

3. The effectiveness of the models based on divergence from randomness is very high in comparison with both $BM25$ and language model. For short queries the performance of the models of divergence from randomness is definitely better than the BM25 model, which since 1994 has been used as a standard baseline for the comparison of the models.

4. We even derived the $BM25$ formula from a parameter-free model of divergence from randomness. The empirical values of the three parameters of the $BM25$ were formally derived with an extreme and surprising precision. This is a further item of evidence of the generality and soundness of our theoretical framework.

5. We introduced several limiting forms as workable models for IR. Possibly the missing investigation on the use of some distributions for IR was also due to the difficulty to manage cumbersome formulae from the implementation point of view.

6. We unified the query expansion problem and the query-document matching problem within a single approach. We offered the same solution to both the problems. The divergence from randomness idea is so powerful, that, unlike other query expansion techniques, our method needs only a few documents to achieve best performance.

7. With the query expansion component we provided four independent components. Our framework is very general and flexible. We can modify each component by choosing alternative techniques which capture the semantics of the component. Therefore our framework opens different and independent research directions for future investigation: new probability distributions for capturing the informative content, new techniques for obtaining the information gain and the term-frequency normalizations.

8. We formulated the term-frequency normalization problem following a formal approach. The term-frequency normalization (called the length normalization) has been viewed before only as an empirical problem of penalizing the term-weights in long documents.

   The length normalization has not been a central topic and it has not been studied

independently of the specific matching function.

One exception comes from the language modelling that we successfully deployed in our framework, using a more stable parameter value, which is independent of the collection. We believe that term-frequency normalization is a central issue not only in the context of our framework but also in other probabilistic approaches to IR, and more research will be done in this direction.

9. The notion of gain was introduced as an attenuation of the *term independence assumption.* We devised two different models of gain. We still would like to test other models of aftereffect, such as those used in the theory of accidents [37, 129, 67, 65], and other potential models for term-frequency normalization such as one using bivariate discrete distribution [67].

10. We have also revisited the literature of Information Retrieval under the light of our probabilistic approach. For example we showed how to use De Finetti's theorem to connect with a single thread the standard probabilistic models, our divergence-based models and language models. De Finetti's theorem suggests that Dirichlet's priors based on the multinomial distribution is effective for IR, but other distributions are equally possible.

In conclusion, our theory offers a unifying theoretical framework able to explain the inductive problem of IR and to construct many different and highly effective models of Information Retrieval. The future research is enhancing and finding out novel instances of the components of the theoretical framework.

# Appendix A

# Evaluation

## A.1 Evaluation measures

A theory on evaluation of IR systems is mainly developed in van Rijsbergen's book [119]. The effectiveness of an IR system is evaluated by the standard measures of *recall* and *precision* as follows:

$$Recall = \frac{\mid Rel \cap Ret \mid}{\mid Rel \mid}$$

$$Precision = \frac{\mid Rel \cap Ret \mid}{\mid Ret \mid}$$

where $Ret = \{d | d \text{ is retrieved}\}$ and $Rel = \{d | d \text{ is relevant}\}$, so that $\mid Rel \cap Ret \mid$ is the number of relevant and retrieved documents, $\mid Ret \mid$ is the number of retrieved documents, and $\mid Rel \mid$ is the number of relevant documents.

The definition of recall and precision is based on the counting measure $\mid \cdot \mid$, because of the binary relationship of relevance. We may generalize recall and precision with an arbitrary measure $\mathbf{m}$ used at the place of the counting function $\mid \cdot \mid$[4]. We call them *multi–valued recall* (M-recall) and *multi–valued precision* (M-precision), which are based on non-binary values of relevance.

Let $\mathbf{m}$ be a discrete measure on the set of documents $D$ on $n \geq 2$ positive real values.

Let $\mathbf{w}$ be the score function on the set of documents. Let $\leq_{\mathbf{m}}$ be the decreasing

ordering induced by relevance $\mathbf{m}$ on the set of documents,

$$d_i \leq_{\mathbf{m}} d_j \Leftrightarrow \mathbf{m}(d_i) \geq \mathbf{m}(d_{o(j)})$$

and $\leq_{\mathbf{w}}$be the decreasing ordering induced by the score $\mathbf{w}$ on the set of documents,

$$d_i \leq_{\mathbf{w}} d_j \Leftrightarrow \mathbf{w}(d_i) \geq \mathbf{w}(d_{r(j)})$$

We want to compare the ordering $\leq_{\mathbf{w}}$ against the ordering $\leq_{\mathbf{m}}$.

We define the *M-recall* measure as

$$M - \text{recall} = \frac{\mathbf{m}(Ret)}{\mathbf{m}(D)} = \frac{\mathbf{m}(Ret)}{\mathbf{m}(Rel)} = \frac{\mathbf{m}(Ret \cap Rel)}{\mathbf{m}(Rel)}$$

where $Ret = \{d | d \text{ is retrieved }\}$ and $Rel = \{d | \mathbf{m}(d) > 0\}$. It is easy to note that $\mathbf{m}(Ret) = \mathbf{m}(Ret \cap Rel)$ and $\mathbf{m}(D) = \mathbf{m}(Rel)$.

Let $0 < x \leq 1$. We now define two positive integers $k_x$ ed $m_x$ such that:

1. Let $d_1 \leq_{\mathbf{w}} \ldots, \leq_{\mathbf{w}} d_N$. $k_x$ is the lowest $k$ satisfying the condition

$$\sum_{i=1}^{k} \mathbf{m}(d_i) \geq x \cdot \mathbf{m}(D)$$

2. Let $d_1 \leq_{\mathbf{m}} \ldots, \leq_{\mathbf{m}} d_N$. $m_x$ is the lowest $m$ so that

$$\sum_{i=1}^{m} \mathbf{m}(d_i) \geq x \cdot \mathbf{m}(D)$$

$m_x$ ($k_x$) is the minimum number of documents according to the decreasing ordering of $\leq_{\mathbf{w}}$ ($\leq_{\mathbf{m}}$) which is sufficient to retrieve a set of documents whose measure of relevance is at least $x \cdot \mathbf{m}(D)$. Note that $k_x \leq m_x$ and $k_x = m_x$ for all $x$ if and only if the two rankings are equivalent (that is are equal up to permutations of documents which preserves their values by $\mathbf{m}$).

Let us define the *M-precision at x of M-recall*:

$$p_x = \frac{k_x}{m_x}$$

**Example 5** *Consider the two rankings:*

| *Rank* | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $\leq_{\mathbf{m}}$ | 1 | 1 | 1 | 1 | 1 | 1 | $\frac{1}{2}$ | $\frac{1}{2}$ | $\frac{1}{2}$ | $\frac{1}{2}$ | $\frac{1}{2}$ | $\frac{1}{2}$ | $\frac{1}{2}$ | $\frac{1}{2}$ | 0 | 0 | 0 | 0 | 0 | 0 |
| $\leq_{\mathbf{w}}$ | 0 | 1 | 1 | 1 | 0 | $\frac{1}{2}$ | 0 | 0 | 1 | $\frac{1}{2}$ | 0 | $\frac{1}{2}$ | 0 | $\frac{1}{2}$ | $\frac{1}{2}$ | 1 | 1 | $\frac{1}{2}$ | $\frac{1}{2}$ | $\frac{1}{2}$ |

*At the recall values* 0.1, 0.3, 0.5 *and* 1 *we have* $\mathrm{m}(D) = 10$ *and* $x \cdot \mathrm{m}(D) = 1, 3, 5, 10$ *respectively. So,* $m_{0.1} = 2$, $m_{0.3} = 4$, $m_{0.5} = 9$ *and* $m_1 = 20$ *while* $k_{0.1} = 1$, $k_{0.3} = 3$, $k_{0.5} = 5$ *and* $k_1 = 14$. *We get the precision values* $p_{0.1} = 0.5$ ,$p_{0.3} = 0.75$, $p_{0.5} = 0.55$ *e* $p_1 = 0.7$

*On the other hand, if the relevance values were binary as*

| *Rank* | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $\leq_{\mathrm{m}}$ | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| $\leq_{\mathrm{w}}$ | 0 | 1 | 1 | 1 | 0 | 1 | 0 | 0 | 1 | 1 | 0 | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |

*we would have* $\mathrm{m}(D) = 14$ *and* $x \cdot \mathrm{m}(D) = 1.4, 5.2, 7, 14$ *respectively. Then,* $m_{0.1} = 3$ ,$m_{0.3} = 10$, $m_{0.5} = 12$ *and* $m_1 = 20$, *while* $k_{0.1} = 2$, $k_{0.3} = 6$, $k_{0.5} = 7$ *e* $k_1 = 14$, *with precision* $p_{0.1} = 0.66$, $p_{0.3} = 0.6$, $p_{0.5} = 0.58$ *and* $p_1 = 0.7$ *respectively.*

The definition of recall and precision in the binary case derives easily. Observe that

$$k_x \overline{\mathrm{k}}_x = \sum_{i=1}^{k_x} \mathrm{m}(d_{o(i)}) \geq x \cdot \mathrm{m}(D)$$

where $\overline{\mathrm{k}}_x$ is the mean value. Similarly,

$$m_x \overline{\mathrm{m}}_x = \sum_{i=1}^{m_x} \mathrm{m}(d_{o(i)}) \geq x \cdot \mathrm{m}(D)$$

where $\overline{\mathrm{m}}_x$ is the mean. In the binary case it is always $\overline{\mathrm{k}}_x = 1$, $m_x$ is the number of retrieved documents and $k_x$ is the number of retrieved and relevant documents, that is the *exact precision* at recall $x$ is

$$p_x = \frac{k_x}{m_x} = \overline{\mathrm{m}}_{m_x} = \frac{|Rel \cap Ret|}{|Ret|}$$

Similarly, if $i$ is the $i$-th document in the ranking, then the *exact precision* at the $i$-th retrieved document is:

$$Pr@i = \frac{|Rel \cap Ret|}{|Ret|} = \frac{|Rel \cap Ret|}{i}$$

Once precision at recall $r$ or at the first $n$ retrieved documents is defined we can plot the precision curve at different values of the number of retrieved documents. If we used exact precision at given recall values we would obtain a certain number of precision values for measuring the effectiveness of the system, for example precision at $0.1, 0.2, \ldots 1.0$. In order to have 11 points of precision values and define the precision at the point 0.0, we may use the *interpolated precision*. For each level of recall $r$, one may consider the maximum exact precision at any retrieved document with any recall value after $r$:

$$Pr_r = \arg_i \max\{Pr@i| \text{ with recall at the } i-\text{th retrieved document} \geq r\}$$

The interpolated precision is obviously nondecreasing and is defined in the recall point 0.0, i.e. the maximum precision $Pr@i$ for every $i$.

All previous precision measures are functions of the recall or the retrieved documents. In order to rank different systems, it is important to have a unique value as a useful indicator of the system performance.

A single-value measure is given by the *non-interpolated average precision* which was proposed by Chris Buckley and was first used in TREC-2 [49]). The non-interpolated average measure of precision is defined as follows: for each $i$-th retrieved relevant document the exact precision $p_i$ is first computed

$$p_i = \frac{i}{n}$$

where $n$ is the document position in the ranking, then the average precision is obtained

$$MAP = \frac{1}{R}\sum_i p_i$$

where $R$ is the number of relevant documents in the collection. The evaluation with the TREC collections considers the first 1000 documents of the ranking, therefore the Mean Average Precision (MAP) is the mean average precision non-interpolated with the first 1000 retrieved documents.

We may generalize the definition of MAP considering the first $n$ retrieved documents. The normalization of the sum of the precision values $p_i$ is done with respect to the minimum between the number $R$ of relevant documents and $n$:

$$MAP@n = \frac{1}{\min\{n, R\}}\sum_i p_i$$

The exact precision at $n$ retrieved documents is denoted by $Pr@n$. We also consider the R-precision, which is the exact precision at $R$ retrieved documents, where $R$ is the number of relevant documents of the query. R-precision is denoted by $Pr@R$.

Finally, for a set of queries, the performace value is obtained by the mean over all queries of the precision measure on single queries.

# Appendix B

# Functions and probability distributions

## B.1 Functions and distributions

The *Gamma function* is defined by

$$\Gamma(x) = \int_0^\infty t^{x-1}e^{-t}dt \quad (\text{ where } x > 0) \tag{B.1}$$

The following relations hold

$$\Gamma(a+1) = a\Gamma(a) \qquad (\text{ with } a > 0\ ) \tag{B.2}$$

$$\Gamma(n) = (n-1)! \qquad (\text{ with } n \text{ integer }) \tag{B.3}$$

$$\Gamma(x) \sim \sqrt{2\pi}e^{-x}x^{x-0.5} \quad (\text{[37, Problem II.12.22] }) \tag{B.4}$$

A random variable $X$ has a *Gamma distribution* with parameters $\alpha$ and $\beta$ ($\alpha > 0$ and $\beta > 0$) if $X$ has the probability density function defined by

$$\frac{\beta^\alpha}{\Gamma(\alpha)}x^{\alpha-1}e^{-\beta x} \quad (\text{ where } x > 0) \tag{B.5}$$

The *Beta distribution* of a random variable $Y$ with parameters $\alpha$ and $\beta$ is defined by the probability density function:

$$f_Y(y,\alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}y^{\alpha-1}(1-y)^{\beta-1} \quad (0 < y < 1) \tag{B.6}$$

Therefore

$$\int_0^1 y^{\alpha-1}(1-y)^{\beta-1}dy = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)} \tag{B.7}$$

Mandelbrot's probability density function of rank-word frequency is:

$$p(r) = (B-1)V^{B-1}(r+V)^{-B} \tag{B.8}$$

where $V$ is the size of the vocabulary. The following distribution is the Zipf distribution and was claimed to be experimentally an excellent approximation of the Equation B.8 by Mandelbrot:

$$p(r) = P \cdot r^{-B} \quad \text{where } P^{-1} = \sum_{r=1}^{\infty}(r+V)^{-B} \tag{B.9}$$

The *Yule distribution* is:

$$p(r) = C \cdot \frac{\Gamma(r)\Gamma(\rho+1)}{\Gamma(r+\rho+1)} \quad r,\rho > 0 \tag{B.10}$$

The *Multinomial distribution* function of a random variable $Y$ is defined by the probability density function:

$$f_Y(y) = \frac{n!}{n_1!\cdots n_k!}y^{n_1}\cdots y_k^{n_k} \quad (\textstyle\sum_{i=1}^{k} y_i = 1) \quad (\sum_{i=1}^{k} n_i = n) \tag{B.11}$$

# Bibliography


[1] AALBERSBERG, I. J. A document retrieval model based on term frequency ranks. In *Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval* (1994), Springer-Verlag New York, Inc., pp. 163–172.

[2] ALLAN, J., CALLAN, J. P., CROFT, W. B., BALLESTEROS, L., BROGLIO, J., XU, J., AND SHU, H. INQUERY at TREC-5. In *In Proceedings of the 5th Text REtrieval Conference (TREC-5), NIST Special Publication 500-238* (Gaithersburg, MD, 1996), pp. 119–132.

[3] AMATI, G., CARPINETO, C., AND ROMANO, G. FUB at TREC 10 web track: a probabilistic framework for topic relevance term weighting. In *In Proceedings of the 10th Text Retrieval Conference TREC 2001* (Gaithersburg, MD, 2002), E. Voorhees and D. Harman, Eds., NIST Special Pubblication 500-250, pp. 182–191.

[4] AMATI, G., AND CRESTANI, F. Probabilistic learning by uncertainty sampling with non-binary relevance. In *Soft Computing in Information Retrieval: techniques and applications*, F. Crestani and G. Pasi, Eds. Physica Verlag, Heidelberg, Germany, 2000, pp. 299–313.

[5] AMATI, G., AND RIJSBERGEN, C. J. V. Probabilistic models of information retrieval based on measuring the divergence from randomness. *ACM Transactions on Information Systems (TOIS) 20*, 4 (2002), 357–389.

[6] AMATI, G., AND VAN RIJSBERGEN, C. J. Term frequency normalization via Pareto distributions. *Lecture Notes in Computer Science 2291* (2002), 183–192.

[7] ARNOLD, B. C. *Pareto distributions.* International Co-operative Publishing House, Fairland, Md., 1983.

[8] ATTAR, R., AND FRAENKEL, A. S. Local feedback in full-text retrieval systems. *Journal of the ACM (JACM) 24*, 3 (1977), 397–417.

[9] BAILEY, P., CRASWELL, N., AND HAWKING, D. Engineering a multi-purpose test collection for Web retrieval experiments. *Information Processing and Management to appear* (2002).

[10] BAR-HILLEL, Y., AND CARNAP, R. Semantic information. *British Journal of the Philosophy of Science 4* (1953), 147–157.

[11] BAR-HILLEL, Y., AND CARNAP, R. *An outline of the Theory of Semantic Information.* Addison–Wesley, Reading, Mass., 1964, pp. 221–274.

[12] BLAIR, D. C. *Language and Representation in Information Retrieval.* Elsevier, Amsterdam, The Netherlands, 1990.

[13] BOOKSTEIN, A., AND KRAFT, D. Operations research applied to document indexing and retrieval decisions. *Journal of the ACM (JACM) 24*, 3 (1977), 418–427.

[14] BOOKSTEIN, A., AND SWANSON, D. Probabilistic models for automatic indexing. *Journal of the American Society for Information Science 25* (1974), 312–318.

[15] BRUZA, P., MCARTHUR, R., AND DENNIS, S. Interactive internet search: keyword, directory and query reformulation mechanisms compared. In *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval* (2000), ACM Press, pp. 280–287.

[16] BRUZA, P. D., AND SONG, D. Inferring query models by computing information flow. In *Proceedings of the Eleventh International Conference on Information and Knowledge Management* (2002), ACM Press, pp. 260–269.

[17] CAI, D., VAN RIJSBERGEN, C. J., AND JOSE, J. M. Automatic query expansion based on divergence. In *Proceedings of the Tenth International Conference on Information and Knowledge Management (CIKM-01)* (New York, Nov. 5–10 2001), H. Paques, L. Liu, and D. Grossman, Eds., ACM Press, pp. 419–426.

[18] CARNAP, R. *Logical Foundations of probability.* Routledge and Kegan Paul Ltd, London, UK, 1950.

[19] CARPINETO, C., DE MORI, R., ROMANO, G., AND BIGI, B. An information theoretic approach to automatic query expansion. *ACM Transactions on Information Systems 19*, 1 (2001), 1–27.

[20] CHAMPERNOWNE, D. The theory of income distribution. *Econometrica 5* (1937), 379–381.

[21] COOPER, W. A definition of relevance for information retrieval. *Information Storage and Retrieval 7* (1971), 19–37.

[22] COOPER, W., AND MARON, M. Foundations of probabilistic and utility-theoretic indexing. *Journal of the ACM (JACM) 25*, 1 (1978), 67–80.

[23] COX, R. T. *The algebra of probable inference.* The Johns Hopkins Press, Baltimore, Md, 1961.

[24] CRESTANI, F., LALMAS, M., VAN RIJSBERGEN, C. J., AND CAMPBELL, I. Is this document relevant? probably: a survey of probabilistic models in information retrieval. *ACM Computing Surveys (CSUR) 30*, 4 (1998), 528–552.

[25] CRESTANI, F., AND VAN RIJSBERGEN, C. Information retrieval by logical imaging. *Journal of Documentation 51*, 1 (1995), 1–15.

[26] CROFT, W. Relevance feedback and inference networks. In *Proceedings of the 16th Annual International ACM SIGIR Conference* (1993), pp. 2–11.

[27] CROFT, W., AND HARPER, D. Using probabilistic models of document retrieval without relevance information. *Journal of Documentation 35* (1979), 285–295.

[28] CROFT, W. B., CALLAN, J. P., AND BROGLIO, J. TREC-2 routing and ad-hoc retrieval evaluation using the inquery system. In *Proceedings of the TREC Conference* (Gaithersburg, MD, USA, 1993), NIST Special publication 500-215.

[29] CROVELLA, M. E., TAQQU, M. S., AND BESTAVROS, A. Heavy–tailed probability distributions in the world wide web. In *A practical guide to heavy tails*, R. Adler, R. Feldman, and M. Taqqu, Eds. Birkhauser, Boston, Basel and Berlin, 1998.

[30] DAMERAU, F. An experiment in automatic indexing. *American Documentation 16* (1965), 283–289.

[31] DEGROOT, M. H. *Probability and Statistics*, 2nd ed. Addison-Wesley, 1989.

[32] DEMPSTER, A. P., LAIRD, N. M., AND RUBIN, D. B. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society series B 39*, 1 (1977), 1–38.

[33] EDMUNDSON, H. A statistician's view of linguistic models and language-data processing. In *Natural language and the computer*, P. L. Gavin, Ed. McGraw-Hill, New York, 1963, pp. 151–179.

[34] EDMUNDSON, H., AND WYLLYS, R. E. Automated abstracting and indexing–survey and recommendations. *Communications of the ACM 4*, 5 (May 1961), 226–234. Reprinted in Readings in Information Retrieval, pp. 390-412. Edited by H. Sharp. New York, NY: Scarecrow; 1964.

[35] ESTOUP, J. *Gammes Stenographiques.* 4th edition, Paris, 1916.

[36] FANO, R. M. *Transmission of Information: A Statistical Theory of Communications.* MIT Press, Cambridge, Mass. and Wiley, New York (published jointly), 1961.

[37] FELLER, W. *An introduction to probability theory and its applications. Vol. I*, third ed. John Wiley & Sons Inc., New York, 1968.

[38] FELLER, W. *An Introduction to Probability Theory and Its Applications*, second ed., vol. II. John Wiley & Sons, New York, 1971.

[39] FISHER, R. On the probable error of a coefficient of correlation from a small sample. *Metron 1* (1921), 3–32.

[40] FUHR, N. Models for retrieval with probabilistic indexing. *Information Processing and Management 25*, 1 (1989), 55–72.

[41] FUHR, N. Probabilistic models in information retrieval. *The Computer Journal 35*, 3 (1992), 243–255.

[42] FUHR, N., AND HÜTHER, H. Optimum probability estimation from empirical distributions. *Information Processing & Management 25*, 5 (1989), 493–507.

[43] GOLDMAN, S. *Information theory.* Prentice–Hall, Englewood Cliffs, N.J., 1953.

[44] GOOD, I. J. *The Estimation of Probabilities: an Essay on Modern Bayesian Methods*, vol. 30. The M.I.T. Press, Cambridge, Massachusetts, 1968.

[45] HAINES, D., AND CROFT, W. B. Relevance feedback and inference networks. In *Proceedings of the 16th annual international ACM SIGIR conference on Research and development in information retrieval* (1993), ACM Press, pp. 2–11.

[46] HARMAN, D. Overview of the First Text REtrieval Conference (TREC-1). In *Proceeding of the TREC Conference* (Gaithersburg, MD, USA, 1992), NIST Special publication 500-207, pp. 1–20.

[47] HARMAN, D. Relevance feedback and other query modification techniques. In *Information Retrieval: Data Structures & Algorithms* (1992), W. B. Frakes and R. A. Baeza-Yates, Eds., Prentice-Hall, pp. 241–263.

[48] HARMAN, D. Relevance feedback revisited. In *Proceedings of ACM SIGIR* (Copenhagen, Denmark, June 1992), pp. 1–10.

[49] HARMAN, D. Overview of the Second Text REtrieval Conference (TREC–2). In *Proceeding of the TREC Conference* (Gaithersburg, MD, USA, 1993), NIST Special publication 500-215, pp. 1–20.

[50] HARMAN, D. Overview of the Fifth Text REtrieval Conference (TREC–5). In *Proceeding of the TREC Conference* (Gaithersburg, MD, USA, 1996).

[51] HARPER, D., AND VAN RIJSBERGEN, C. An evolution of feedback in document retrieval using co–occurrence data. *Journal of Documentation 34*, 3 (1978), 189–216.

[52] HARTER, S. P. *A probabilistic approach to automatic keyword indexing.* PhD thesis, Graduate Library, The University of Chicago, Thesis No. T25146, 1974.

[53] HARTER, S. P. A probabilistic approach to automatic keyword indexing. part I: On the distribution of specialty words words in a technical literature. *Journal of the ASIS 26* (1975), 197–216.

[54] HARTER, S. P. A probabilistic approach to automatic keyword indexing. part II:An algorithm for probabilistic indexing. *Journal of the ASIS 26* (1975), 280–289.

[55] HAWKING, D. Overview of the trec-9 web track. In *In Proceedings of the 9th Text Retrieval Conference (TREC-9)* (Gaithersburg, MD, 2001), pp. 87–102.

[56] HAWKING, D., AND CRASWELL, N. Overview of the trec-2001 web track. In *In Proceedings of the 10th Text Retrieval Conference (TREC-10)* (Gaithersburg, MD, 2002), pp. 61–67.

[57] HEATH, D., AND SUDDERTH, W. De Finetti Theorem on Exchangeable Variables. *The American Statistician 30*, 4 (1976), 188–189.

[58] HERDAN, G. *Quantitative Linguistics.* Butterworths, London, 1964.

[59] HIEMSTRA, D. Term-Specific Smoothing for the Language Modeling Approach to Information Retrieval: The Importance of a Query Term. In *Proceedings of ACM SIGIR* (Tampere, Finland, August 12-15 2002), ACM Press, New York, NY, USA, pp. 35–41.

[60] HIEMSTRA, D., AND DE VRIES, A. Relating the new language models of information retrieval to the traditional retrieval models. Research Report TR–CTIT–00–09, Centre for Telematics and Information Technology, 2000.

[61] HILPINEN, R. On information provided by observation. In *Information and Inference*, J. Hintikka and P. Suppes, Eds., Synthese Library. D. Reidel publishing company, Dordrecht-Holland, 1970, pp. 97–122.

[62] HINTIKKA, J. The varieties of information and scientific explanation. In *Logic, Methodology and Philosphy of Science III* (Amsterdam, 1968), B. van Rootselaar and J. Staal, Eds., North-Holland, pp. 311–331.

[63] HINTIKKA, J. On semantic information. In *Information and Inference*, J. Hintikka and P. Suppes, Eds., Synthese Library. D. Reidel publishing company, Dordrecht-Holland, 1970, pp. 3–27.

[64] IDE, E. New experiments in relevance feedback. In *The SMART Retrieval System*, Salton, Ed. Prentice-Hall, 1971, pp. 337–354.

[65] IRWIN, J. O. The generalized Waring distribution applied to accident theory. *J. Roy. Statist. Soc. Ser. A 131* (1968), 205–225.

[66] JEFFREYS, H. *Theory of Probability*. Oxford University Press, Oxford, 1961. Third Edition, First Published in 1939.

[67] KOCHERLAKOTA, S., AND KOCHERLAKOTA, K. *Bivariate discrete distributions*. Marcel Dekker Inc., New York, 1992.

[68] KWOK, K. Experiments with component theory of Probabilistic Information Retrieval based on single terms as document components. *ACM Transactions on Information Systems 8*, 4 (1990), 363–386.

[69] KWOK, K. L. A new method of weighting query terms for ad-hoc retrieval. In *Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval* (1996), ACM Press, pp. 187–195.

[70] LAFFERTY, J., AND ZHAI, C. Document Language Models, Query Models, and Risk Minimization for Information Retrieval. In *Proceedings of ACM SIGIR* (New Orleans, Louisiana, USA, September 9-12 2001), ACM Press, New York, NY, USA, pp. 111–119.

[71] LI, M., AND VITANYI, P. Philosophical issues in kolmogorov complexity. *Lecture Notes in Computer Science 623* (1992), 1–15.

[72] LI, M., AND VITANYI, P. *An Introduction to Kolmogorov Complexity and Its Apllications.* Springer Verlag, New York, 1997.

[73] LUHN, H. A Statistical Approach to Mechanized Encoding and Searching of Literary Information. *ibmjrd 1* (1957), 309–317.

[74] MANDELBROT, B. On the theory of word frequencies and on related markovian models of discourse. In *Proceedings of Symposia in Applied Mathematics. Vol. XII: Structure of language and its mathematical aspects.* American Mathematical Society, Providence, R.I., 1961, pp. 190–219. Roman Jakobson, editor.

[75] MANMATHA, R., RATH, T., AND FENG, F. Modeling score distributions for combining the outputs of search engines. In *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval* (2001), ACM Press, pp. 267–275.

[76] MARGULIS, E. L. N-poisson document modelling. In *Proceedings of the 15th annual international ACM SIGIR conference on Research and development in information retrieval* (1992), ACM Press, pp. 177–189.

[77] MARON, M. Automatic indexing: an experimental inquiry. *Journal of the Association for Computing Machinery 8* (1961), 404–417.

[78] MARON, M. E., AND KUHNS, J. L. On relevance, probabilistic indexing and information retrieval. *Journal of the ACM (JACM) 7*, 3 (1960), 216–244.

[79] MCLACHLAN, G. J., AND KRISHNAN, T. *The EM Algorithm and Extensions.* Wiley, 1997.

[80] MITRA, M., SINGHAL, A., AND BUCKLEY, C. Improving automatic query expansion. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval* (1998), ACM Press, pp. 206–214.

[81] PLACHOURAS, V., OUNIS, I., AMATI, G., AND VAN RIJSBERGEN, C. J. University of Glasgow at the TREC 2002 Web Track, 2002.

[82] PONTE, J., AND CROFT, B. A Language Modeling Approach in Information Retrieval. In *The 21st ACM SIGIR Conference on Research and Development in Information Retrieval* (Melbourne, Australia, 1998), B. Croft, A. Moffat, and C. van Rijsbergen, Eds., pp. 275–281.

[83] POPPER, K. *The Logic of Scientific Discovery* (The bulk of the work was first published in Vienna in 1935, this reprint was first published by Hutchinson in 1959, new notes and footnotes in the present reprint). Routledge, London, 1995.

[84] PURI, P., AND GOLDIE, C. Poisson mixtures and quasi-infinite divisibility of distributions. *Journal of Applied Probability 16*, 1 (1979), 138–153.

[85] RENYI, A. *Foundations of probability.* Holden-Day Press, San Francisco, USA, 1969.

[86] ROBERTSON, S. On relevance weight estimation and query expansion. *Journal of Documentation 42*, 3 (1986), 288–297.

[87] ROBERTSON, S., AND WALKER, S. Some simple approximations to the 2-Poisson Model for Probabilistic Weighted Retrieval. In *Proceedings of the Seventeenth Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval* (Dublin, Ireland, June 1994), Springer-Verlag, pp. 232–241.

[88] ROBERTSON, S., WALKER, S., BEAULIEU, M., GATFORD, M., AND PAYNE, A. Okapi at trec-4. In *NIST Special Publication 500-236: The Fourth Text REtrieval Conference (TREC-4)* (1996), D. Harman, Ed., Department of Commerce, National Institute of Standards and Technology.

[89] ROBERTSON, S. E. The probability ranking principle in IR. *Journal of Documentation 33* (1977), 294–304.

[90] ROBERTSON, S. E., AND SPARCK-JONES, K. Relevance weighting of search terms. *Journal of the American Society for Information Science 27* (1976), 129–146.

[91] ROBERTSON, S. E., VAN RIJSBERGEN, C. J., AND PORTER, M. Probabilistic models of indexing and searching. In *Information retrieval Research*, S. E. Robertson, C. J. van Rijsbergen, and P. Williams, Eds. Butterworths, 1981, ch. 4, pp. 35–56.

[92] ROCCHIO, J. Relevance feedback in information retrieval. In *The SMART Retrieval System*, Salton, Ed. Prentice-Hall, 1971, pp. 313–323.

[93] SALTON, G. *The SMART Retrieval System.* Prentice Hall, New Jersey, 1971.

[94] SALTON, G., AND BUCKLEY, C. Term-weight approaches in automatic text retrieval. *Information Processing and Management 24*, 5 (1988), 513–523.

[95] SALTON, G., AND BUCKLEY, C. Improving Retrieval Performance by Relevance Feedback. *Journal of the American Society for Information Science 41*, 4 (1990), 182–188.

[96] SALTON, G., AND MCGILL, M. *Introduction to modern Information Retrieval.* McGraw–Hill, New York, 1983.

[97] SALTON, G., AND MCGILL, M. *The SMART retrieval system - experiments in automatic document retrieval.* Prentice Hall Inc., Englewood Cliffs, USA, 1983.

[98] SHANNON, C. A mathematical theory of communication. *Bell System Technical Journal 27* (July and October 1948), 379–423 and 623–656.

[99] SHANNON, C., AND WEAVER, W. *The Mathematical Theory of Communication.* University of Illinois Press, Urbana, Illinois, 1949.

[100] SICHEL, H. S. Parameter estimation for a word frequency distribution based on occupancy theory. *Comm. Statist. A—Theory Methods 15*, 3 (1986), 935–949.

[101] SICHEL, H. S. Word frequency distributions and type-token characteristics. *Math. Sci. 11*, 1 (1986), 45–72.

[102] SIMON, H. A. On a class of skew distribution functions. *Biometrika 42* (1955), 425–440.

[103] SINGHAL, A., BUCKLEY, C., AND MITRA, M. Pivoted document length normalization. In *Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval* (1996), ACM Press, pp. 21–29.

[104] SINGHAL, A., SALTON, G., AND BUCKLEY, C. Lenght normalisation in degraded text collections. Research Report 14853-7501, Department of Computer Science, Cornell University, Ithaca, NY, USA, 1995.

[105] SINGHAL, A., SALTON, G., MITRA, M., AND BUCKLEY, C. Document length normalization. *Information Processing and Management 32*, 5 (1996), 619–633.

[106] SOLOMONOFF, R. A formal theory of inductive inference. Part I. *Information and Control 7*, 1 (Mar. 1964), 1–22.

[107] SOLOMONOFF, R. A formal theory of inductive inference. Part II. *Information and Control 7*, 2 (June 1964), 224–254.

[108] SOLOMONOFF, R. Complexity-based induction systems: comparisons and convergence theorems. *IEEE Transactions on Information Theory 24* (1978), 422–432.

[109] SPARCK JONES, K. A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation 28(1)* (1972), 11–21.

[110] SRINIVASAN, P. On generalizing the two-poisson model. *Journal of The American Society for Information Science 41*, 1 (1990), 61–66.

[111] STEEL, R. G., TORRIE, J. H., AND DICKEY, D. A. *Principles and Procedures of Statistics. A Biometrical Approach*, 3rd ed. MacGraw–Hill, 1997.

[112] STEINHAUS, H. The problem of estimation. *The Annals of Mathematical Statistics 28* (1957), 633–648.

[113] STONE, D., AND RUBINOFF, B. Statistical generation of a technical vocabulary. *American Documentation 19*, 4 (1968), 411–412.

[114] TITTERINGTON, D. M., SMITH, A. F. M., AND MAKOV, U. E. *Statistical analysis of finite mixture distributions.* John Wiley & Sons Ltd., Chichester, 1985.

[115] TURTLE, H., AND CROFT, W. A comparison of text retrieval models. *The Computer Journal 35*, 3 (June 1992), 279–290.

[116] TURTLE, H., AND CROFT, W. B. Evaluation of an inference network-based retrieval model. *ACM Transactions on Information Systems (TOIS) 9*, 3 (1991), 187–222.

[117] TURTLE, H. R. *Inference Networks for Document Retrieval.* Department of Computer and Information Science, University of Massachusetts, 1991.

[118] VAN RIJSBERGEN, C. A theoretical basis for the use of co-occurrence data in information retrieval. *Journal of Documentation 33* (1977), 106–119.

[119] VAN RIJSBERGEN, C. *Information Retrieval, second edition.* Butterworths, London, 1979.

[120] VOORHEES, E. M. Query expansion using lexical-semantic relations. In *Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval* (1994), Springer-Verlag New York, Inc., pp. 61–69.

[121] VOORHEES, E. M. Overview of TREC 2001. In *In Proceedings of the 10th Text Retrieval Conference (TREC-10)* (Gaithersburg, MD, 2001), NIST Special Publication 500-250, pp. 1–15.

[122] VOORHEES, E. M., AND HARMAN, D. Overview of the Eighth Text REtrieval Conference. In *In Proceedings of the 8th Text Retrieval Conference (TREC-8)* (Gaithersburg, MD, 1999), NIST Special Pubblication 500-246, pp. 1–24.

[123] VOORHEES, E. M., AND HARMAN, D. Overview of TREC 2000. In *In Proceedings of the 9th Text Retrieval Conference (TREC-9)* (Gaithersburg, MD, 2000), NIST Special Pubblication 500-249, pp. 1–14.

[124] WILLIS, D. G. Computational complexity and probability constructions. *Journal of the ACM (JACM) 17*, 2 (1970), 241–259.

[125] WILLIS, J. *Age and area.* Cambridge University Press, London and New York, 1922.

[126] WITTEN, I. H., MOFFAT, A., AND BELL, T. C. *Managing Gigabytes*, second ed. Morgan Kaufmann Publishers, San Francisco, California, 1999.

[127] WONG, S., AND YAO, Y. A probabilistic inference for information retrieval. *Information Systems 16* (1991), 301–321.

[128] WONG, S., AND YAO, Y. On modeling information retrieval with probabilistic inference. *ACM Transactions on Information Systems 16* (1995), 38–68.

[129] XEKALAKI, E. The bivariate generalized Waring distribution and its application to accident theory. *J. Roy. Statist. Soc. Ser. A 147*, 3 (1984), 488–498.

[130] XU, J., AND CROFT, W. Query expansion using local and global document analysis. In *Proceedings of ACM SIGIR* (Zurich, Switzerland, Aug. 1996), pp. 4–11.

[131] XU, J., AND CROFT, W. B. Improving the effectiveness of information retrieval with local context analysis. *ACM Transactions on Information Systems (TOIS) 18*, 1 (2000), 79–112.

[132] ZHAI, C., AND LAFFERTY, J. A Study of Smoothing Methods for Language Models Applied to Ad Hoc Information Retrieval. In *Proceedings of ACM SIGIR* (New Orleans, Louisiana, USA, September 9-12 2001), ACM Press, New York, NY, USA, pp. 334–342.

[133] ZHAI, C., AND LAFFERTY, J. Two-Stage Language Models for Information Retrieval. In *Proceedings of ACM SIGIR* (Tampere, Finland, August 12-15 2002), ACM Press, New York, NY, USA, pp. 49–56.

[134] ZIPF, G. *Human behavior and the principle of least effort.* Addison-Wesley Press, Reading, Massachusetts, 1949.