Modèles probabilistes pour les fréquences de mots et la recherche d'information

Stéphane Clinchant

UNIVERSITÉ DE GRENOBLE

THÈSE

Pour obtenir le grade de

DOCTEUR DE L'UNIVERSITÉ DE GRENOBLE

Spécialité : **Informatique**

Arrêté ministérial : 7 août 2006

Présentée par

Stéphane Clinchant

Thèse dirigée par **Eric Gaussier**
et codirigée par **Boris Chidlovskii**

préparée au sein **Laboratoire d'Informatique de Grenoble**
et de **l'Ecole Doctorale Mathématiques, Sciences et technologies de l'information, Informatique**

Probabilistic Models of Word Frequencies and Information Retrieval

Thèse soutenue publiquement le ,
devant le jury composé de :

Hervé Martin
Professeur LIG, Président
Mohand Boughanem
Professeur IRIT, Rapporteur
Jean-Cédric Chappelier
EPFL, Rapporteur
François Yvon
Professeur LIMSI, Rapporteur
Giambattista Amati
Fondation Ugo Bordoni, Examinateur
Stephen Robertson
Microsoft Research Cambridge, Examinateur
Eric Gaussier
Professeur LIG, Directeur de thèse
Boris Chidlovskii
Xerox Research Center Europe, Co-Directeur de thèse

Résumé

Cette thèse propose de relier des observations empiriques sur les fréquences de mots dans des collections textuelles aux modèles probabilistes de Recherche d'Information (RI). Concernant les modèles statistiques de fréquences de mots, nous portons notre attention sur l'étude du phénomène de rafale (a rich get richer phenomenon). Nous établissons une propriété sur les distributions de probabilité caractérisant leur capacité à modéliser ce phénomène et nous montrons ensuite que la distribution Beta Négative Binomiale est un bon modèle statistique pour les fréquences des mots.

Nous portons ensuite notre attention sur les modèles probabilistes de RI et leur propriétés fondamentales. Nous introduisons une nouvelle famille de modèle probabiliste, fondé sur la notion d'information de Shannon qui permet d'établir un lien conséquent entre les propriétés importantes des modèles de RI et le phénomène de rafale. Ces nouveaux modèles obtiennent des résultats comparables aux modèles de réference et les surpassent avec la boucle de rétro pertinence.

Enfin, les meilleurs performances de nos modèles pour la rétro-pertinence nous ont conduit à étudier empiriquement et théoriquement les modèles de rétro-pertinence. Nous proposons un cadre théorique qui permet d'expliquer en partie leurs caractéristiques empiriques et leur performances. Ceci permet, entre autres, de mettre en avant les propriétés importantes des modèles de retro-pertinence et de montrer que certains modèles de référence sont déficients.

Abstract

The present study deals with word frequencies distributions and their relation to probabilistic Information Retrieval (IR) models. We examine the burstiness phenomenon (a rich get richer phenomenon) of word frequencies in textual collections. We propose to model this phenomenon as a property of probability distributions and we show that the Beta Negative Binomial distribution is a good statistical model for words frequencies.

We then focus on probabilistic IR models and their fundamental properties. We then introduce a novel family of probabilistic models, based on Shannon information. These new models bridge the gap between significant properties of IR models and the burstiness phenomenon of word frequencies. These new models yield comparable performances to state of the art IR models and outperform them when Pseudo Relevance Feedback is used.

Lastly, the better performances of our models for Pseudo Relevance Feedback (PRF) lead us to study empirically and theoretically PRF models. We propose a theoretical framework which explain well the empirical behaviour and performance of PRF models. Overall, this analysis highlights interesting properties for pseudo relevance feedback and shows that some state-of-the-art model are inadequate.

Résumé de la Thèse

Introduction

Si la recherche d'information (RI) sur le web est dominée par des systèmes apprenant des fonctions d'ordonnancement à partir de log de données, la RI *ad hoc* est largement dominée par des modèles probabilistes avec peu de paramètres à régler, comme Okapi, les modèles de langues et les modèles DFR (*Divergence from Randomness*). Ces derniers sont fondés sur plusieurs distributions de probabilité et hypothèses qui facilitent leur déploiement en pratique. Si ces modèles semblent bien fondés d'un point de vue RI, les distributions de probabilités sous-jacentes s'accordent mal avec les données empiriques collectées dans les collections textuelles.

Il y a eu beaucoup d'études empiriques sur les distributions de fréquences de mots, dont les modèles de recherche d'information pourraient bénéficier. Quelle connaissance sur les lois régissant les fréquences de mots devrait être appliquée au problème de recherche d'information ? On pourrait penser qu'un 'bon' modèle statistique de fréquences de mots devrait conduire à un 'bon' modèle de RI. Il s'avère pourtant que ce n'est pas le cas ainsi que le suggère l'état de l'art. C'est pourquoi nous nous demandons quelles sont les propriètés des fréquences de mots qui pourraient être utiles en RI et s' il serait possible de concevoir un modèle probabiliste à la fois efficace, performant en RI et motivé par des études statistiques sur le comportement des mots.

Nous nous interessons plus particulièrement à un phénomène important, observé par Church et Gale [12] et d'autres, qui est celui du comportement en rafale, ou crépitement (en anglais *burstiness*) des mots. Ce phénomène décrit le fait que les mots, dans un document, tendent à apparaître par paquets. En d'autres termes, une fois que l'on a observé une occurrence d'un mot dans un document, il est bien plus probable d'observer de nouvelles occurrences de ce mot.

Pour résumer, nous nous posons donc les questions suivantes:

1. Comment le phénomène de rafale peut etre modélisé dans un cadre probabiliste ?
2. Pouvons nous trouver de 'meilleurs' modèles probabilistes ?
3. Comment utiliser ces nouvelles distributions pour la RI ?

Résumé des Chapitres

Dans une premier temps, nous examinons dans le chapitre 2 les modèles proposés pour modéliser les fréquences de mots, telles que le modèle 2-Poisson, Négative Binomiale et Dirichlet Multinomial. Nous discuterons du phénomène de rafale et d'adaptation pour les fréquences de mots et des contributions importantes de Katz et Church [45, 13]. Même si le phénomène de rafale a été abordé dans differentes études et avec différentes distributions, notre approche se distingue par la volonté de caractériser les distributions

qui peuvent naturellement prendre en compte ce phénomène. C'est pourquoi nous proposons une définition formelle de loi de probabilité qui sont 'en rafale', par extension avec le phénomène que l'on veut modéliser. Cette définition est en fait équivalente à la log convexité de la fonction de survie $P(X > x)$ de la loi de probabilité considérée.

Ainsi, nous pouvons caractériser les distributions classiques de fréquences de mots et montrer que la plupart sont inadéquates au regard du phénomène de rafale. Nous avons alors étudié deux distributions afin de modéliser les fréquences de mots: la loi Beta Negative Binomiale et la distribution Log-Logistique. Nous avons reconsidéré la distribution Negative Binomiale, dont le comportement en rafale depend de ses paramètres, pour obtenir une distribution de probabilité qui soit toujours en rafale. Nous montrons ensuite comment et dans quels cas la distribution Log-Logistique peut être vue comme une approximation continue de la distribution Beta Negative Binomiale.

Nous verifions ensuite l'adéquation de ces modèles aux données de fréquences à travers plusieurs expériences et nous validons ainsi ces distributions. Ceci nous ensuite au problème de l'application de ces lois de probabilités aux problèmes de RI.

Pour cette raison, nous passons en revue les modèles références de RI dans le chapitre 3. Nous rappelons les hypothèses principales des modèles BM25, des modèles de langues et de modèles 'Divergence from Randomness' (DFR). Nous examinons ensuites les propriètés fondamentales des modèles de RI dans le chapitre 4, comme les effets et conditions sur la croissance et concavité des fonctions de pondérations et l'effet IDF entre autres. Nous montrons aussi que le premier principe de normalisation des modèles DFR est une des conséquences du fait que les lois de probabilités sous-jacente ne sont pas en rafale. Plus généralement, nous discutons de la relation entre la propriété de concavité des modèles de RI et la propriété de rafale des lois de probabilités des modèles sous-jacents. Tous les modèles de référence en RI sont des fonctions concaves avec les fréquence des mots mais toutes les distributions de probabilités utilisées ne sont pas en rafale. On pourrait considérer que le phenomène de rafale et la concavité des modèles de RI comme deux versants différents du même problème, à savoir comment traiter et ne pas surévaluer ou sous evaluer les fortes fréquences des mots.

Par conséquent, nous pensons que les modèles probabilistes de RI actuels ne sont pas compatibles avec les distributions Beta Negative Binomiale et Log-Logistique et nous introduisons donc une nouvelle famille de modèles probabilistes pour la RI, fondée sur la notion d'information de Shannon. Lorsque la loi de probabilité sous-jacente est capable de modéliser le phénomène de rafale, alors le modèle devient naturellement valide au sens des propriètés fondamentales des modèles de RI.

Nous donnons l'exemple de deux modèles dans cette famille. Le premier modèle repose sur une distribution log-logistique et le deuxième modèle sur une loi que nous avons appelé Loi de Puissance Lissé (Smoothed Power Law). Ces deux modèles sont évalués sur plusieurs collections de documents et offrent des performances similaires voire identique aux modèles de références. Nous étendons ces modèles d'information au cadre de retro-pertinence (Pseudo Relevance Feedback). Avec cette extension, les modèles que nous avons proposé surpassent les modèles référence de rétro-pertinence sur plusieurs collections.

Le bon comportement de nos modèles pour la retro-pertinence nous a amener à examiner en détail les caractéristiques qui les distingue des modèles classiques. Nous nous sommes basés sur l'étude des propriètés fondamentales des modèles de RI pour l'étendre au modèle de retro-pertinence. Nous dressons donc une liste de contraintes classiques avant d'introduire une nouvelle contrainte pour les modèles de rétro-pertinence, contrainte liée à la fréquence documentaire (DF) des mots dans l'ensemble de rétro-pertinence. Nous analysons ensuite, d'un point de vue théorique, différents modèles de rétro-pertinence par rapport à ces contraintes. Cette analyse montre que plusieurs modèles références ne satisfont pas plusieurs contraintes au contraire des modèles d'information. Les contraintes

que nous mentionnons sont validées empiriquement sur plusieurs collection afin de vérifier leur bien fondé. Au final, nous avons établi un panorama des modèles de retro-pertinence avec une théorie qui permet d'expliquer les résultats experimentaux de ces modèles.

Conclusion

Nous avons étudier des modèles probabilistes pour les fréquences de mots et pour la recherche d'information. Puis, nous avons essayé de relier ces modèles dans le but d'obtenir à la fois un 'bon' modèle statistique des fréquences de mots et un 'bon' modèle de recherche d'information.

Nous avons proposé de modéliser le phénomène de rafale comme une propriété des distributions de probabilités caractérisant ainsi leur capacité à modéliser ce phénomène. Ceci nous a amené à considérer de nouvelles distributions comme la Beta Négative Binomiale. Nous montrons que cette distribution est un relativement bon modèle statistique pour les fréquences des mots et explique mieux les données que la plupart des distributions de probabilités utilisées en recherche d'information.

Nous avons ensuite analysé les modèles de RI afin de mieux comprendre leur propriètés fondamentales. Nous introduisons une nouvelle famille de modèles probabilistes pour la recherche d'information, fondé sur la notion d'information de Shannon et qui permet d'établir un lien conséquent entre les propriètés importantes des modèles de Recherche d'Information et le phénomène de rafale. Par exemple, nous montrons une relation directe entre entre la concavité des modèles de RI et le comportement en rafale des distributions modelisant les fréquences des mots. Nos expériences montrent que ces nouveaux modèles obtiennent des résultats comparables aux modèles de réferences et les surpassent avec la boucle de rétro pertinence.

Enfin, les meilleurs performances de nos modèles pour la rétro-pertinence nous ont conduit à étudier empiriquement et théoriquement les modèles de rétro-pertinence. Nous proposons un cadre théorique qui permet ainsi d'expliquer en partie leurs caractéristiques empiriques et leur performances. Ceci permet, entre autres, de mettre en avant les propriètés importantes des modèles de retro-pertinence et de montrer que certains modèles de référence sont déficients.

Ces nouveaux modèles fondés sur l'information sont performants, intuitifs, aisés à mettre en oeuvre et ont été extensivement analysés d'un point de vue théorique et pratique.

Contents

List of Figures

List of Tables

Notations

Table 1: Notations

Notation	Description
w	A term, or term index
d	A document, or document index
N	Number of documents in the collection
M	Number of indexing terms in the collection
x_{wd}	Number of occurrences of w in document d
x_{wq}, q_w	Number of occurrences of w in query q
t_{wd}, or $t(w,d)$	Normalized version of x_{wd}
l_d	Length of document d
$avgl$	Average document length
F_w	Number of occurrences of w in collection: $F_w = \sum_d x_{wd}$
N_w	Number of documents containing w: $N_w = \sum_d I(x_{wd} > 0)$
IDF(w)	$-\log(N_w/N)$
L	Length of collection $= \sum_w F_w$
X_w	Univariate discrete random variable for the frequencies of w
T_w	Univariate continuous random variable for normalized frequencies of w
X^d	Multivariate Random variable modelling a document.
$P(X > x)$	Survival function
PRF Notation	
n	# of docs retained for PRF
$\mathbf{F}$	Set of documents retained for PRF: $\mathbf{F} = (d_1, \ldots, d_n)$
tc	*TermCount*: # of terms in $\mathbf{F}$ added to query
$TF(w)$	$=\sum_{d \in F} x_{wd}$
$DF(w)$	$=\sum_{d \in F} I(x_{wd} > 0)$

Chapter 1

Introduction

The beginning of Information Retrieval can be dated to Luhn's works in the 50's. Luhn, a computer scientist working at IBM, had to deal with new problems raised by libraries and documentation centers. Since then, information access techniques were developed in order to face the information society advent. According to Manning et. al [54], Information Retrieval (IR) can be defined as:

Information Retrieval. *Information Retrieval is finding material (usually documents) of an unstructured nature (usually text) that satisfies an information need from within large collections (usually stored on computers).*

Due to the large amount of information available on computers, users need efficient methods to access and search various source of information. Information Retrieval organizes and models unstructured information as opposed to database systems. It enables users to access a large collection of documents/information in diverse ways. As one of the first media digitized was texts written in natural language, IR emerged naturally as a sub-domain of Natural Language Processing (NLP).

The typical ad-hoc IR scenario confronts a user, with his information need expressed in a given query language to a document representation given as an index. A function matches the query to the document representation in order to return a ranked list of objects to the user. An information retrieval system, as shown in figure 1.1, consists in 3 elements:

- A query model,
- A document model,
- A function, called Retrieval Status Value (RSV), matching queries and documents. The bigger the function values are, the better documents are supposed to answer the query.

The very first models in IR regarded words as first order logic predicates. From this point of view, a document d was considered relevant if it entailed the query q according to laws of logic.

$$RSV(q,d) = \begin{cases} 1 \text{ if } d \Rightarrow q \\ 0 \text{ otherwise} \end{cases} \tag{1.1}$$

Later, vectorial models represented queries and documents in Euclidean spaces. Each dimension of the Euclidean spaces corresponds to a given word, or indexing term. Then, the similarity between a document d and a query q can be calculated by the angle between these two vectors:

$$RSV(q,d) = \cos(\vec{q},\vec{d}) \tag{1.2}$$

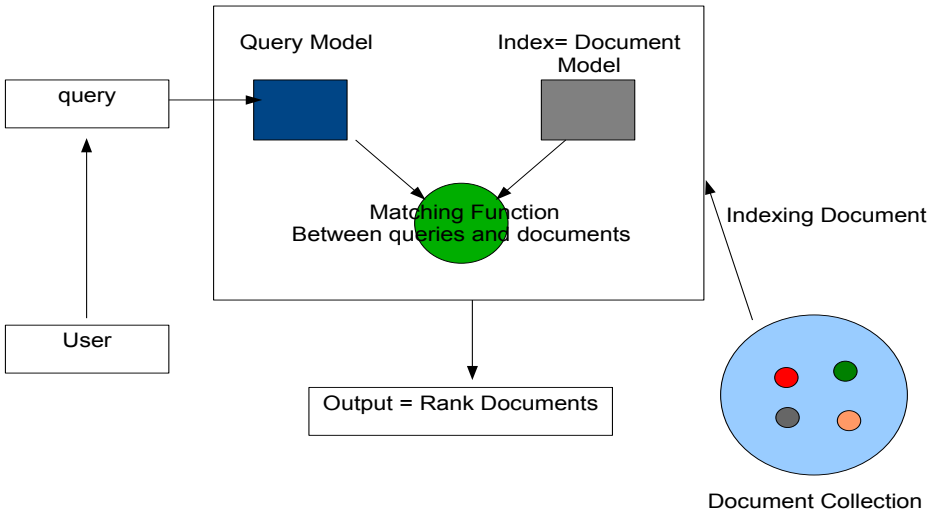


Figure 1.1: Information Retrieval System Architecture. An Information Retrieval system is composed by a query model, ie a query language/formalism, a component to index documents and a function matching queries and documents. The output of an IR system is in general a ranked list of documents.

Lastly, probabilistic models of IR considers queries and documents as the result of random processes. Many IR problems are tackled with a probabilistic framework. One way or another, all probabilistic IR models make an assumption which can be formulated as follows:

Hypothesis. *Words and their frequencies, in a document or a collection of documents, can be considered as random variables. Thus, it is possible to observe and study word frequencies as random events.*

Hence, probabilistic models rely on the choice of probability distributions to model documents and queries. Probabilistic models can be specified by 3 elements: a probability distribution P_{doc} modeling documents, a probability distribution P_{query} modeling queries and a function H matching these distributions.

Let $\mathcal{D}$ be the random variable modeling document d, $\mathcal{Q}$ the random variable modeling query q, then an IR model can be defined as follows:

$$\begin{aligned} \mathcal{D} &\sim P_{doc}(.|\lambda) \\ \mathcal{Q} &\sim P_{query}(.|\theta) \\ RSV(q,d) &= H(P_{query}(\mathcal{Q}=q|\theta), P_{doc}(\mathcal{D}=d|\lambda)) \end{aligned}$$

For example, if Multinomial distributions are chosen for both queries and documents and if H is the Kullback-Leibler divergence, then the resulting model is equivalent to the KL retrieval model defined in [46].

This PhD thesis adopts this probabilistic approach to Information Retrieval. We now proceed to a short discussion on the research questions investigated.

1.1 Research Questions

A general question raised by this retrieval framework deals with the choice of particular word frequency distributions. There has been many statistical studies on word frequencies whose probabilistic IR models could benefit from. What knowledge on word frequencies can be transferred to the retrieval tasks ? If we were to think in terms of probability, we would tend to believe that a 'good' statistical model of word frequencies should yield a valid and effective IR models. It is however not always the case in IR models, one of the reason being the significant role of the H function previously mentionned. This is why we wonder which are the properties of word frequencies that could be useful in IR and whether it would be possible to design an effective IR model whose underlying word frequencies distributions are valid or well motivated from statistical studies. Overall, these questions implicitly address the feature representation errors in probabilistic IR models.

Modeling word frequencies in documents is not a specific problem to IR. Many natural language processing tasks do require a probability model for word frequencies. In addition, word frequencies can also be studied solely from a statistical perspective, with the goal of finding a *good* model of word frequencies where the notion of *good* model can be defined in terms of mean squared error, χ^2 statistic or any statistical measure.

Word frequency distributions can be studied from different perspectives. The very first models of word frequency were typically interested in modeling the frequency spectrum or grouped frequency distribution. With one or several documents, these models first collect statistics on the number $V(m)$ of different words that appear exactly m times and fit a probabilistic model to these observed counts. For example, the number $V(1)$ is the number of words that appear only 1 time (hapax legomena). If the word *probability* and *retrieval* both appear k times, then observing these two words in a text is considered as the same statistical event: the observation of a word that appear k times. So, these models group words by frequency in a text. This is typically what addresses the Yule-Simon, the Waring-Herdan-Muller model and to some extent the Zipf Law [4].

However, this is not the kind of model we are interested in the present study. We adopt the approach used in many IR or NLP tasks [37, 13] where each different word is modeled independently. So, for each different word w, the distribution of the occurrences of w in a corpus of documents is the object under study.

While this approach is common to many IR and NLP tasks, there does not seem to be a consensus on the distributions to use in order to model word occurrences. Probabilistic IR models typically rely on a mixture of 2-Poisson distribution (Okapi [72]), Multinomial distributions (Language Models [46]) or Poisson and Geometric distributions in the Divergence From Randomness framework [2]. But, Church [13], among others, emphasized one peculiarities of word frequency : burstiness. Actually, burstiness was originally defined by Katz by the following statement:

> *burstiness*, i.e. multiple occurrences of a content word or phrase in a single text document, which is contrasted with the fact that most other documents contain no instances of this word or phrase at all"

Burstiness has then implications on the distributions to use and several studies highlighted that common distributions used in IR may not be appropriate to model correctly word frequency data. In a nutshell, the common distributions used in IR are criticized for their limited variance while several alternative distributions such as the Negative Binomial [13] or the Dirichlet Compound Multinomial [53] can account for more variance.

A naive question could be the following: how come that state of the art model do not account for burstiness ? Maybe one could think that burstiness is not important in IR tasks and that it is not necessary to account for this phenomenon. Burstiness make large

frequencies not such a rare event. Distributions with more variance can generally better estimate the probability of a large number of occurrence, ie a large deviation from the mean frequency. Hence, models accounting for burstiness are not so much 'surprised' to observe large frequencies.

How large frequencies are managed in IR models ? It turns out that IR models have found a different way to address burstiness. All IR models are concave functions with term frequency. Concavity in term frequency prevents IR models from assigning a too large score to a document because of one large frequency in a document. Hence, IR models are not so much 'surprised' to observe large frequencies.

Intuitively, burstiness and the IR model concavity in term frequency seem to be two sides of the same coin.

This PhD thesis fits in with the probabilistic approach to Information Retrieval and draw inspiration, at its beginnings, from Church seminal paper on Poisson mixtures [13]. First of all, we were primarily interested in finding better probabilistic models of words frequency that address the burstiness phenomenon. Above all, we tackle this problem from a different perspective compared to related works: we propose to characterize burstiness as a *property* of probability distribution. Therefore, this property enable to distinguish bursty probability distributions from non-bursty. In addition, the Negative Binomial proposed by Church has been reconsidered and extended toward the Beta Negative Binomial and the Log-Logistic distribution, a continuous counterpart. Both of these distributions are bursty according to our definition of burstiness.

Having suggested new probability distributions, the remaining task was to apply these models in IR or NLP tasks. It turns out this was not as straightforward as initially thought. This is why we had to reexamine IR models foundations in order to better understand the different aspects involved for ranking documents. In particular, the Divergence From Randomness framework [2] caught our attention as a starting point for our analysis. As the application of the proposed distributions in this framework revealed problematic, we then introduced a new family of IR models, *information-based models*, which require and rely on bursty distributions. This family can be seen as a simplification of the Divergence From Randomness framework in order to comply with our proposed distributions.

Finally, the good performance of information-based models for Pseudo Relevance Feedback (PRF) [1] lead us to experimentally and theoretically analyze PRF models. As a result, we establish a list of axiomatic constraints for pseudo relevance feedback models aiming at capturing 'good' properties of PRF models. Our theoretical analysis provide an explanation on why the information-based models perform better than other models in PRF settings.

In a nutshell, this PhD thesis investigated the following research questions:

1. How can burstiness be modeled in probabilistic models ?
2. Can we find better probabilistic model accounting for the burstiness phenomenon of words frequencies ?
3. How could these new models be used for ad-hoc information retrieval ?

1.2 Contributions

We now proceed to a brief summary of the main contributions presented in this thesis.

1. Our first proposal is to define burstiness as a *property* of probability distributions:

[1] Pseudo Relevance Feedback aims at automatically expanding the initial query with terms found in the top retrieved documents.

Burstiness. *Let X a random variable defined on $\mathbb{R}$ with distribution P. The distribution P is bursty iff $\forall \epsilon > 0$, the function g_ϵ defined by:*

$$\epsilon > 0,\ g_\epsilon(x) = P(X \geq x + \epsilon | X \geq x)$$

is a strictly increasing function of x. A distribution which verifies this condition is said to be bursty. The same definition applies to discrete distributions except that $\epsilon \in \mathbb{N}$.

This definition directly translates the notion of adaptation: a word is bursty if it is easier to generate it again once it has been generated a certain number of times. Moreover, it enables to characterize most distributions proposed so far to model word frequencies.

Then, we propose two models of word frequencies: the *Beta Negative Binomial* distribution, a discrete model, and the *Log-Logistic* distribution a continuous one. Finally, several experiments demonstrate the appropriate behavior of these distributions to model burstiness: the Beta Negative Binomial and the Log-Logistic distributions are sound models of word frequencies: *they enjoy good theoretical properties, as bursty distributions, and they fit well word frequencies empirically.*

2. Our second main contribution is the definition of novel family of IR model: *information-based model.* We propose the family of IR models satisfying the following equation:

$$RSV(q,d) = \sum_{w \in q} -q_w \log P(T_w > t_w | \lambda_w)$$

where T_w is a random variable modelling normalized term frequencies and λ_w is a set of parameters of the probability distribution modelling word w in the collection. This ranking function corresponds to the mean information a document brings to a query or, equivalently, to the average of the document information brought by each query term. This model has interesting properties that connect the burstiness property of probability distributions to important property of IR models. We then propose two effective IR models within this family: *the log-logistic and the smoothed power law models.* Regarding performances, both the log-logistic and smooth power law models yield state of the art performance, without pseudo relevance feedback, and *significantly outperforms* state of the art models with pseudo relevance feedback.

3. We have conducted a *theoretical analysis* of PRF models. First, we estbalish a list of theoretical properties including a novel one, called the *Document Frequency* constraint. Second, we have then investigated standard PRF models with respect to these constraints. This theoretical study has revealed several important points: a) several state-of-the-art model are deficient with respect to one or several PRF theoritecal properties, b) information-based model satistfy all the PRF properties. Thus it provides an *explanation on why the information-based models perform better than other models in PRF settings.*

1.3 Outline

Chapter 2 surveys the main probabilistic models of word frequencies. Multinomial models, including topic models, are briefly reviewed before introducing the burstiness phenomenon. 2-Poisson models, Negative Binomial models, the Katz-Mixture model and Polya Urn schema are discussed in the context of burstiness. Then, we move on to a formal definition of burstiness, which relates to the log-convexity of the survival function

and which enables to characterize probability distributions as bursty or non-bursty. The Negative Binomial model is reconsidered with the Beta Negative Binomial distribution and the Log-Logistic model is proposed as a continuous counterpart. Finally, several experiments are carried in order to validate of the proposed distributions.

Having introduced the Beta Negative Binomial and Log-Logistic models, we want to tackle ad-hoc IR with these distributions. Chapter 3 examines the foundations of the main probabilistic IR models. This chapter draws up a state of the art of probabilistic IR models including the Probability Ranking Principle, Language models and Divergence from Randomness models. Among the three families, it is the Divergence From Randomness framework that will retain our attention and which will serve us as a starting point for a formal analysis of IR models hanks to retrieval heuristics constraints in chapter 4 and to the elaboration of a suitable framework for the BNB and Log-Logistic distributions. In particular, the role of the first normalization principle is shown to be directly linked to a particular retrieval constraint, the concavity in term frequency. Finally, DFR models are shown to be inappropriate when word frequencies are modeled with a Beta Negative Binomial distribution. This will suggest that the DFR framework may not be appropriate with our candidate distributions.

Chapter 5 introduces the family of information-based models for ad-hoc IR, which can be seen as a simplification of the DFR framework. Two effective IR models are proposed: the Log-Logistic and a novel probability distribution, the Smoother Power Law. These models yield state of the art performance, without pseudo relevance feedback, and significantly outperforms state of the art models with pseudo relevance feedback. We have tested these models with different term frequency normalizations and extended them with the beneficial use of the q-logarithm.

Finally, chapter 6 analyzes pseudo relevance feedback models in order to establish a list of axiomatic constraints for pseudo relevance feedback. This chapter introduces conditions PRF models should satisfy. These conditions are based on standard IR constraints, with the addition of a *Document Frequency* (DF) constraint which we have experimentally validated. We have then investigated standard PRF models wrt to these constraints. The theoretical study we conduct reveals that several standard PRF models either fail to enforce the IDF effect or the DF effect whereas the log-logistic and the smoothed power law models satisfy all the PRF properties. Our theoretical analysis thus provide an explanation on why the information-based models perform better than other models in PRF settings.

Chapter 2

Probabilistic Models of Word Frequencies

Contents

2.1 Introduction

Word frequency data are generally represented in a term-document matrix $\mathbf{X} = (x_{wd})$ where rows stand for words and columns for documents. The term-document matrix results from the preprocessing of a document collection, which is explained in the appendix for readers unfamiliar with word frequency data.

The very first models of word frequency were typically interested in modeling the grouped frequency distribution: the number of different words that appear exactly k times in a collection of documents. The grouped frequency distribution is not the object of interest in IR or Natural Language Processing tasks, where documents or the occurrences of a word w are the object under study. For most probabilistic models, the different terms w are supposed to be independent from each other so that most probabilistic models of word frequency are in general univariate distributions.

We will call X_w the random variable for the frequencies of word w. Each words w is modeled with a distribution $P(X_w|\lambda_w)$ in a collection, where λ_w is a set of parameters for word w. All words in the collection are in general modeled with the same class of distribution but with different parameters. For example, if one choose to model word frequencies with a Poisson distribution, then each word is a represented with its own Poisson distribution. In a way, these probabilistic models look at the data matrix by line. We will call X^d the random variable for a document, a multivariate distribution where each marginal is a word frequency random variable X_w.

This chapter review the main probabilistic model of word frequencies, namely the probability distributions used to model the random variables X_w. We also present state of the art probabilistic document models, which most of all rely on a Multinomial distribution to tie all words together in a multivariate distribution. First, several points concerning the peculiarities of textual data are shortly discussed:

Discrete vs Continuous Word frequencies, ie observations are discrete. So most models are discrete probability distributions. Nevertheless, document differs in length: some documents are longer than some others and a normalization of term frequencies could be used as a preprocessing step. As most normalizations transform frequencies in continuous values, continuous probability models can also be used.

High Dimensionality Textual data is high-dimensional as many documents and many different words are observed. Typical IR test collections have sizes around several hundred thousands documents and the number of different terms is even bigger. It is common in IR collections that the number of indexing terms reach a million or more different terms.

Sparsity The observations matrix $\mathbf{X}$ is very sparse. Indeed, most words do not occur in most documents, they mostly occur in the subset of documents. Table 2.1 shows the percentage of non-zero observations for two TREC collections. Furthermore, there are a lot of rare words, which occur only a few times in the collection.

Table 2.1: Sparsity

Collection	Non-Zeros Observations Percentage
TREC-7	3×10^{-4}
TREC-3	4×10^{-4}

Estimation Excepted naive models, probabilistic models of texts suffer from estimation problems. Often being intractable, the estimation of documents models is approximated by a simpler function to optimize. Moreover, the high-dimensionality of data makes the estimation even more costly in computation. Approximations are also used to speed up the computation procedure due to the huge amount of data. Most probabilistic models of texts are approximated one way or another.

In a nutshell, textual data is **sparse high-dimensional** and **discrete** which renders **models estimation** difficult.

In addition to these general features, **the phenomenon of burstiness** have been shown to affect word frequencies, as shown by Church and Gale [13]. The term "burstiness" describes the behavior of words which tend to appear in bursts, ie once they appear in a document, they are much more likely to appear again.

The burstiness phenomenon is the connecting thread of this chapter. Section 2.2 will begin with Multinomial models of word frequencies, which have been criticized wrt burstiness. Then, in section 2.3 the burstiness phenomenon is extensively discussed in order to introduce other probabilistic models. We then propose a formal definition of burstiness and suggest two distributions: the Beta Negative Binomial and the Log-Logistic in section 2.4. The last section deals with experiments validating the Beta Negative Binomial and Log-Logistic models.

2.2 Multinomial Document Models

2.2.1 Multinomial Model

The Multinomial model is a very popular model. It was first used with naive Bayes categorization models ([57]) and later in IR through the so-called language models. The Multinomial distribution is a multivariate generalization of the Binomial distribution and its density function is as follows:

$$P(X^d|\theta, l_d) = P(X^d = (x_{1d}, \ldots, x_{Md})|\theta, l_d) = \frac{l_d!}{\prod_w x_{wd}!} \prod_w \theta_w^{x_{wd}}$$

where l_d is the document length and θ encodes the proportion of each word and its statistical mean. The Multinomial model suppose the length of a document l_d (in tokens) is known beforehand and that words occurrences are independent from each other. The independence of word occurrences is expressed by the product $\prod_w$, which in probability theory means independence of events.

A document is simply seen as a bag of tokens, where words occurrences are independent from each other. This means that different occurrences of the same term are statistically independent. For example a document could be:

$$(soviet, president, US, soviet, cold, war)$$

So, the occurrence US and *soviet* are independent from each other. So are the multiple occurrences of the word *soviet*. Drawing at random from multinomial distribution amounts to drawing from a urn filled with balls of different colors as figure shown in figure 2.1.

The marginal random variables X_w are Binomial distributions whose mean and variance are:

$$\begin{aligned} E(X_w) &= l_d\theta_w \\ Var(X_w) &= l_d\theta_w(1-\theta_w) \end{aligned}$$

So the variance of the distributions are essentialy controlled by its mean which is one of the model limitations we will discuss later.

Moreover, the Multinomial distribution is very convenient because its estimation is straightforward: the maximum likelihood estimator (mle) of θ is:

$$\hat{\theta}_w = \frac{x_{wd}}{\sum_w x_{wd}} = \frac{x_{wd}}{l_d}$$

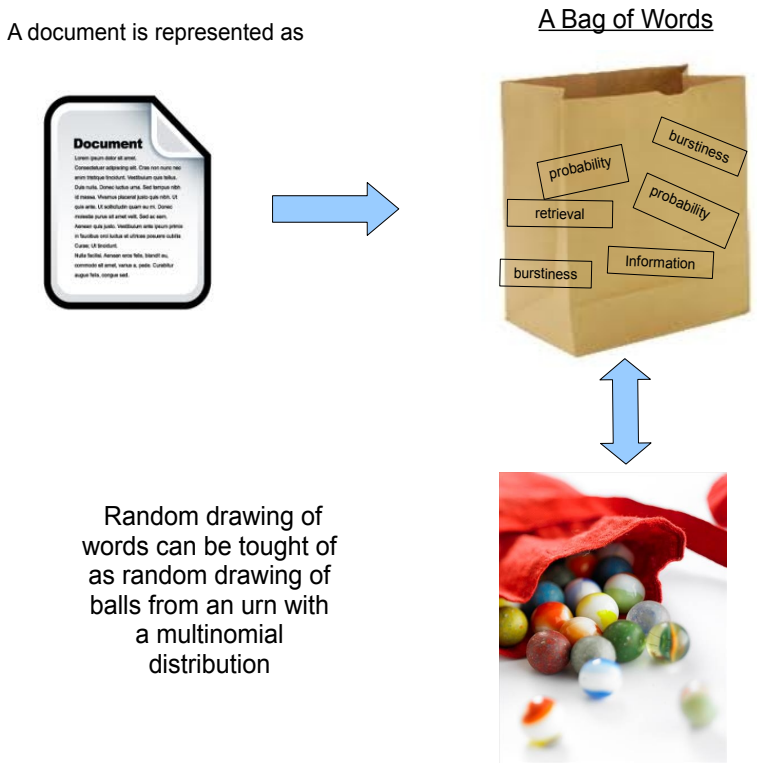


Figure 2.1: Bag of Word Analogy with a bag of balls. Different words are represented with balls of different color with possible repetitions

This is simply the proportion of a given word in a sample. However, this estimator need to be smoothed to take into account unobserved words. Well-known smoothing methods include Laplace, Jelinek-Mercer and Dirichlet smoothing [57, 85]. Overall, Multinomial models are simple but convenient. This may explain why they are so popular and why they often serve as basic units in more complex models such as topic models

2.2.2 Topic Models

Topic models build on the idea of Latent Semantic Analysis [27] in a probabilistic way. Topic models assume there are some underlying topic/themes in a collection of documents. For example, a topic could deal with politics, another with science etc. Most topic models assume a Multinomial distribution for a given topic. So, a topic is specified by a distribution of words, corresponding to the θ parameter for Multinomials. For example, words such as *election*, *president*, *poll* would have high probabilities for a topic dealing with politics. These topics are estimated from a given set of documents, thanks to the co-occurrences of words in documents as in Latent Semantic Analysis.

Then, the key idea is to model documents with a mixture of such topics. Figure 2.2 illustrates the principle of topic models. The figure represent topics by different colors for words, ie blue, red, green to indicate which words are the most likely for this topic. Recall that each topic has a probability distribution over words, so all words are possible but some are more likely. So, this figure shows that 3 different topics that will be at the basis of the document generation process. There are several ways to define mixture models which correspond to different assumptions:

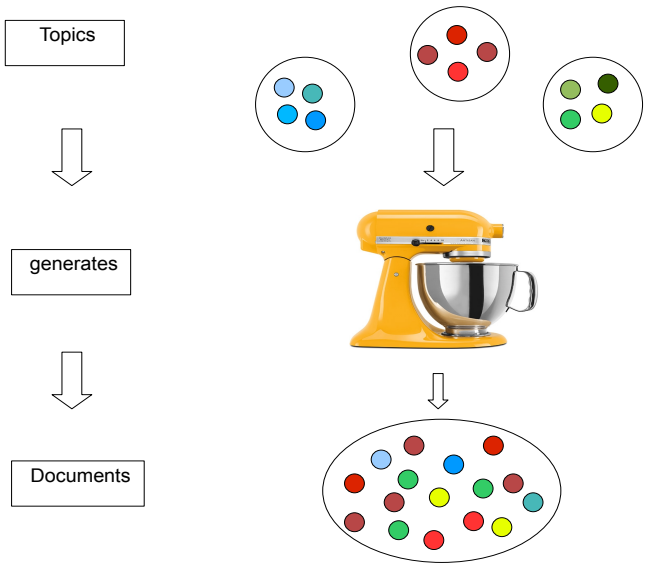


Figure 2.2: Principle of a Mixture Model. Different latent topics/themes generate documents. This mixture model here is polythematic

Monothematic A document can only deal with a single topic. A document can speak of Politics, or Science but not both at the same time. This correspond to the mixture of multinomial proposed by Nigam [64]. If the mixture model in the figure was monothematic, the resulting document would ideally only have either only blue ball, only red balls or only green balls.

Polythematic A document can deal with several topics at the same time. This assumption correspond to the probabilistic latent semantic analysis (PLSA) [40], and the latent Dirichlet allocation (LDA) models [7]. The mixture model in the figure in polythematic: the resulting documents has a combination of blue, red and green 'words'.

After this informal introduction to mixture models, we now move to a formal presentation of Nigam mixture model [64] and the polythematic mixture models.

Mixture of Multinomial

A natural extension to the multinomial model is simply to consider a mixture of Multinomials [64]. The idea underlying this model is to capture several topics or themes in a collection of documents. These different topics are modeled with K multinomials and a document is supposed to be monothematic: a document can cover only one topic.

$$P(X^d|\theta_1,\ldots,\theta_K,p_1,\ldots,p_K,l_d) = \frac{l_d!}{\prod_w x_{wd}!}\sum_k p_k \prod_w \theta_{wk}^{x_{wd}}$$

where parameter θ_k is the multinomial distribution over words for topic k and p_k the proportion of topic k in the collection of documents. This multinomial mixture is in general estimated by an EM algorithm [28] as many mixture models.

Probabilistic Latent Semantic Analysis (PLSA)

The PLSA model goes back over the assumption of monothematicity of documents [40]. In this model, a document can thus express different topics. PLSA can also be thought of as a probabilistic version of latent semantic analysis. Hoffman [40] regards the corpus as a set of document-word couples and these couples are supposed to follow a mixture of multinomials. This enables a document to use different themes to explain different words. Let d be the index of a document and w the index of a word. Then, the model is defined by:

$$P((d,w)|\alpha,\beta,p_1,\ldots,p_K) = \sum_k p_k P(d|k,\alpha)P(w|k,\beta) = \sum_k p_k \alpha_{kd}\beta_{kw}$$

with $P(d|k,\alpha)$ and $P(w|k,\beta)$ following multinomial distributions. The log-likelihood of the corpus is defined by:

$$LL = \sum_{w,d} \log P((d,w)|\alpha,\beta)$$

To generate a pair (d,w), the PLSA model chooses first a topic k with probability p_k Then, one chooses a document d with probability α_{kd} and a word w with probability β_{kw}. Conditionally to that topic k, the probabilities of the document and the word become independent. Parameters β_{kw} can be understood as the probability of word w in the topic k. Parameters of the model can be estimated by a standard EM. Let I^k_{wd} a random variable indicating which topic was used for a given word-document pair. Then, the EM equations are:

$$\begin{aligned}
\text{E-step:} \quad P(I^k_{wd} = 1|(d,w),\alpha,\beta) &= \frac{p_k\alpha_{kd}\beta_{kw}}{\sum_k p_k\alpha_{kd}\beta_{kw}} \\
\text{M-step:} \quad \alpha^{i+1}_{kd} &\propto \sum_w x_{wd} P(I^k_{wd} = 1|(d,w),\alpha,\beta) \\
\beta^{i+1}_{kw} &\propto \sum_d x_{wd} P(I^k_{wd} = 1|(d,w),\alpha,\beta) \\
p^{i+1}_k &\propto \sum_{d,w} x_{wd} P(I^k_{wd} = 1|(d,w),\alpha,\beta)
\end{aligned}$$

Hoffmann also present a tempered EM algorithm in order to boost convergence [40]. Gibbs Sampling methods can also be used to estimate the model parameters as shown in [8] Note that PLSA is not a truly generative model of documents. Theoretically, it is not possible to compute the probability of a theme given a new document in the collection. In practice, a 'fold-in' step is used to approximate this probability.

Latent Dirichlet Allocation (LDA)

LDA [7] is the generative counterpart of f PLSA: The generative process of LDA is the following:

- For each document d, draw a variable θ following a Dirichlet, where θ stands for the proportion of each topics in this document
- Do l_d times
  - Draw a topic k from a multinomial with parameter θ
  - Draw a word w from topic k (multinomial with parameter β_k)

Recall that Dirichlet distribution models multivariate data θ on the unit simplex such that $\sum_{i=1}^{n} \theta_i = 1$. The Dirichlet probability density function is given by:

$$P(\theta_1, \ldots, \theta_K|\alpha) = \frac{\Gamma(\sum_{k=1}^{K} \alpha_i)}{\prod_{k=1}^{K} \Gamma(\alpha_i)} \prod_{k=1}^{K} \theta_k^{\alpha_k - 1}$$

The probability of a word given a topic is:

$$P(w|k, \beta_k) = Multinomial(\beta_k, 1) = \beta_{kw}$$

Finally, the likelihood for a document is:

$$P(X^d|\alpha, \beta, l_d) = \int_\theta \frac{\Gamma(\sum_{k=1}^{K} \alpha_k)}{\prod_{k=1}^{K} \Gamma(\alpha_k)} \prod_{k=1}^{K} \theta_k^{\alpha_k - 1} \times \prod_{w=1}^{l_d} (\sum_k \theta_k \beta_{kw})^{x_{wd}} d\theta$$

and the data log-likelihood.

$$LL = \sum_d \log P(X^d|\alpha, \beta)$$

Estimation of these models is carried with variational methods or Gibbs Sampling algorithm [7, 8]. A comparison of the different estimation methods is proposed in [3].

2.2.3 Summary

Multinomials models are very popular, relatively easy to extend to specific cases and intuitively easy to understand thanks to their analogy to urn models. However, their main drawback is the assumption of **independence of word occurrences**. We will now examine several experiments and studies aiming at overcoming this limitation. In particular, we will discuss the burstiness phenomenon which somehow invalidates the multinomial assumption of independence.

2.3 Burstiness Phenomenon

We want to discuss here the phenomenon of burstiness adressed in many studies of word frequencies. First of all, we recall Katz's original definition of burstiness for word frequencies. Then, several arguments against the multinomial model are presented. Several probability distributions relevant to the bursty behavior of words are reviewed including:

- the 2-Poisson Model
- the Negative Binomial
- the K-Mixture
- the Pólya Urn Models

2.3.1 Definition of Burstiness

According to Church [13], Katz [45] was the first to introduce the term *burstiness*. He distinguished two notions of burstiness which act at different scales:

LA COMÈTE HARTLEY 2 AU PLUS PRÈS DE LA TERRE

Ce cliché très détaillé de Hartley 2, montre sa chevelure diffuse ainsi que sa queue de gaz. Elle a été obtenue le 13 octobre avec une lunette de 100 mm. Crédit : Nick Howes.

La comète Hartley 2 est passée au plus près de la Terre le 20 octobre 2010. Observez-la !

Il est rare de voir une comète d'aussi près : seulement 1/8 de la distance Terre-Soleil ! Son éclat reste néanmoins modéré puisque Hartley 2 est timidement visible à l'œil nu.

La raison : il s'agit d'une petite comète dont le noyau n'excède pas 1,4 km de diamètre. Ce n'est rien en comparaison de la comète géante Hale-Bopp, visible en plein Paris en 1997. Elle mesurait 40 km de large.

La Lune gênante pour observer la comète

Malheureusement, le ciel nocturne est lumineux jusqu'au 30 octobre 2010 à cause d'une Lune allant de sa phase pleine vers une phase gibbeuse. L'éclat de celle-ci masque en grande partie l'objet diffus et peu contrasté que constitue Hartley 2.

Le centre de la comète reste visible en cette période, mais sa queue s'efface en présence de la moindre lueur parasite. Elle devient même inobservable le 28, lorsqu'elle se trouve à moins de 7° d'une Lune éclairée à 68%.

Période plus favorable à partir du 30 octobre

À partir du 30 octobre 2010, cherchez la comète en milieu de nuit, juste avant que la Lune ne se lève. Vous ne perdez presque rien à attendre cette date.

Figure 2.3: A news article. The occurrences of the 'keyword' comète are highlighted

> "The notion of burstiness is fundamental for the subject matter discussed here. It will be used for characterization of two closely related but distinct phenomena:
> (a) *document-level burstiness*, i.e. multiple occurrences of a content word or phrase in a single text document, which is contrasted with the fact that most other documents contain no instances of this word or phrase at all; and
> (b) *within document burstiness* (or burstiness proper), i.e. close proximity of all or some individual instances of a content word or phrase within a document exhibiting multiple occurrences. A within-document burst always indicate an instance of a document-level burstiness, but not necessarily *vice-versa*"

In other words, Katz introduced concepts of burstiness at the document-level (case-b) and the corpus-level (case-a). At the document level, it means that there is a close agglomeration of word occurrences in a document. At the corpus level, it means that there exists few documents with a large number of occurrences for a given term and a lot of documents with few occurences. We try to illustrate this 'multiple occurrences' phenomenon in figure 2.3, which shows a news article dealing with astronomy. We borrow from [42] the figure 2.4 in order to mention that burstiness can also be observed in images with visual words.

2.3.2 Against the Multinomial Model

Church and Gale in their seminal paper [13] stressed two important points:

1. The Poisson and Binomial are inappropriate to model text due to their inability to

Figure 2.4: Illustration of the burstiness for visual keyword

model large deviations from the mean.

> "It has been our experience that the observed variance of the frequency of a word (or ngram) across documents is almost always larger than the mean, and therefore, larger than what would be expected under either the Binomial or the Poisson. The errors between the observed variance and the Poisson prediction tend to be particularly noticeable for content words in large and diverse collections."

2. They borrow the concept of *Adaptation* from speech processing:

> "We have found $Pr(k \geq 2|k \geq 1)$ to be useful for modeling adaptation [...] Under standard independence assumptions, it is extremely unlikely that lightning would strike twice (or half a dozen times) in the same document. But text is more like a contagious disease than lightning. If we see one instance of a contagious disease such as tuberculosis in a city, then we would not be surprised to find quite of few more. Similarly, if a few instances of "said" have already been observed in a document, then there will probably be some more."

The analogy between texts and diseases by Church suggests that once a word appears in a document, it is much more likely to appear again in this document. In a way, the notion of adaptation here is closer to the notion of document-level burstiness of Katz, except it does not encode the notion of proximity within the document. The behavior - the more we have, the more we'll get is likely to produce high frequency for a term but it does not directly says a word does not appear in a lot of documents, namely the definition of burstiness at the corpus level. Hence, there is not a direct alignment between the notion of adaptation and burstiness even if they are intimately related. Burstiness, according to Katz definition is *a state of affairs*, whereas, adaptation according to Church may be one *explanation* of burstiness. Roughly speaking, burstiness and adaptation describe the same phenomenon: words can have high frequencies, ie there are bursts of occurrences in some documents and the concepts of adaptation and burstiness have somehow been merged in the literature. Sometimes, it is not a useful distinction to stress, but it is important to keep in mind the two level of burstiness: *at the document level* and *at the corpus level*.

Church's experiments [12] brought the adaptation phenomenon to light. In a series of experiments, some documents were split in two parts: from the beginning to the middle part and from the middle part to the end of the document. This enables to measure the proclivity of words to reappear in the second part of the document knowing they have appeared in the first part. These experiments clearly demonstrated the inadequacy of the binomial model as Church [12] explains:

> "Repetition is very common. Adaptive language models, which allow probabilities to change or adapt after seeing just a few words of a text, were introduced in speech recognition to account for text cohesion. Suppose a document mentions Noriega once. What is the chance that he will be mentioned again? If the first instance has probability p, then under standard (bag-of words) independence assumptions, two instances ought to have probability p^2, but we find the probability is actually closer to $p/2$. The first mention of a word obviously depends on frequency, but surprisingly, the second does not. Adaptation depends more on lexical content than frequency; there is more adaptation for content words (proper nouns, technical terminology and good keywords for information retrieval), and less adaptation for function words, cliches and ordinary first names."

In a way, Church upholds Harter's experiments [37] in the 70s. Indeed, Harter showed that *content-bearing* words are those which diverge the most from a Poissonian behavior. Here, Church stresses that content-bearing words are the ones that tends to be repeated the most: "there is more adaptation for content words".

To sum up, Multinomial distributions have a limited capacity to model over-dispersed events (high variance) and seem inappropriate to model properly word frequency. We will now review several distributions addressing the burstiness phenomenon. Somehow, *all these models address the limited variance* problem encountered by a single Poisson or a Multinomial distribution.

2.3.3 2-Poisson Model

Harter [37] observed that *specialty*, ie content words diverge the most from a Poissonian behavior, whereas non-specialty words are close to a Poissonian behavior. Harter employed a mixture of two Poisson distributions to model term frequency in a corpora. The intuition of the 2-Poisson model can be explained in the following way: many words appear with a relatively weak frequency in many documents and appear with a greater frequency, or densely, only in one restricted set of documents. This last set is called the *Elite*[1] set (noted E) because it is supposed to contain the documents which treat mainly of the word topic. The idea is thus to model the elite set by a Poisson distribution with parameter λ_E, and the non-elite set by another Poisson distribution of parameter λ_G. Implicitly, $\lambda_E > \lambda_G$. The 2-Poisson model is then a mixture of two Poisson distributions:

$$P(X_w = x_w|\alpha, \lambda_E, \lambda_G) = \alpha\frac{e^{-\lambda_E}\lambda_E^{x_w}}{x_w!} + (1-\alpha)\frac{e^{-\lambda_G}{\lambda_G}^{x_w}}{x_w!} \tag{2.1}$$

Figure 2.5 shows 2 mixtures of 2-Poissons: the non-elite component is modeled with a Poisson of mean 3 and the elite component by a Poisson of mean 10. These two mixtures differs by their mixture parameter.
Eliteness here is directly related to the corpus-level definition of burstiness: there are few documents that contains a lot of frequencies of a particular term. Hence, the 2-Poisson mixture model is an attempt to capture burstiness at the corpus-level.

[1] Elite is not the exact term proposed by Harter but it is the one used later on in the literature

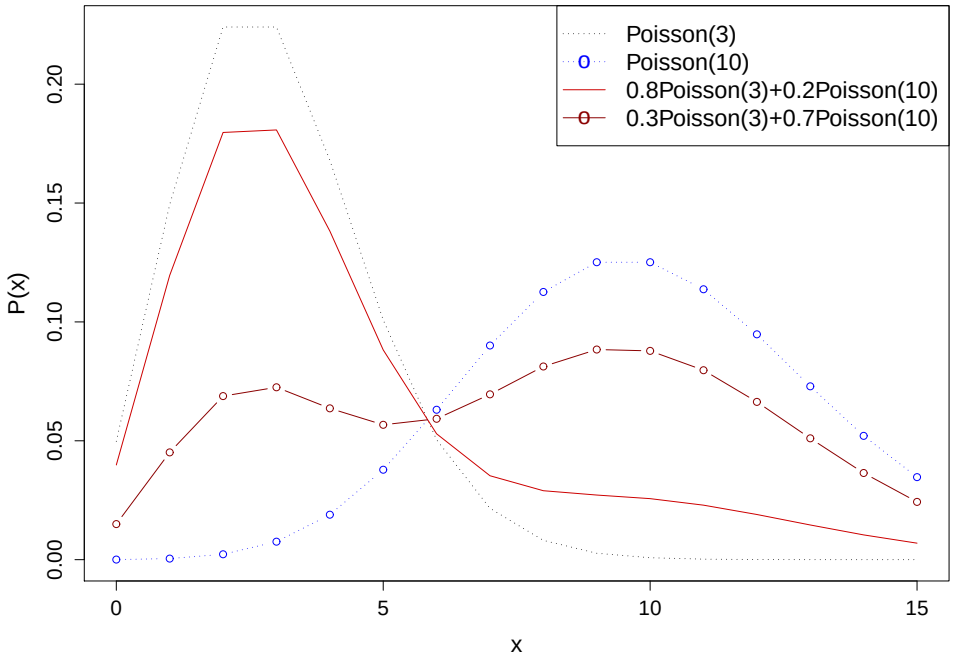


Figure 2.5: 2-Poisson Mixture Model

One of the first applications of information retrieval was the operation of indexing. The operation of indexing consists in choosing the terms regarded as good descriptors of the document: The general idea is the following one: a good descriptor of a document is a word rather frequent in the document and relatively rare in the collection. Thus, it would describe relatively well the content of a document and would be rather discriminating compared to other terms in the collection. Harter used the 2-Poisson mixture model to suggest indexing terms. The probability a document belongs to the elite set is :

$$P(d \in E | X_w = x_w) = \frac{P(X = x_w, d \in E)}{P(X_w = x_w)} = \frac{\alpha \frac{e^{-\lambda_E} \lambda_E^{x_w}}{x_w!}}{\alpha \frac{e^{-\lambda_E} \lambda_E{}^{x_w}}{x_w!} + (1-\alpha) \frac{e^{-\lambda_G} \lambda_G^{x_w}}{x_w!}} \tag{2.2}$$

Harter used this quantity to sort words likely to be indexing terms. He then proposed to measure a distance between the two Poissons with:

$$z = \frac{\lambda_E - \lambda_G}{\sqrt{\lambda_E + \lambda_G}}$$

This measure is closely related to the t-test statistic when assessing the significance of the difference between two sample means. This statistics encodes a measure of separability of the two Poisson distributions. If the elite set is well distinguishable from the non-elite set, then the word is likely to be a good descriptor. However, the 2-Poisson requires to estimate 3 parameters for each word. Harter used a method of moments in order to estimate these parameters (we describe the method in the appendix). However, this

estimation raise some problems. Indeed, Harter propose one method which often has degenerated cases. Sometimes, there is not enough observations to be able to distinguish the two Poisson distributions.

Summary

To sum up, even if the 2-Poisson model assumptions are relatively simple, this model had a significant influence in the development of IR models. It is at the heart of Okapi [72] model and has inspired partly DFR models [2] as we will see in the chapter 3

2.3.4 Negative Binomial

In order to model word frequencies, the 2-Poisson model has been extended to the case of n components by Margulis [56]. Then, Church and Gale were interested by the Negative Binomial [13] which can be viewed as an infinite mixture of Poisson distributions. Church and Gale compared the Binomial and Poisson distributions with mixtures of Poisson to model word frequencies. Their results indicate that the Negative Binomial distribution, which is an infinite mixture of Poisson distributions, fits the data better than a n-Poisson mixture. The family of Negative Binomial distributions is a two parameter family, and supports several equivalent parametrizations. A commonly used one employs two real valued parameters, β and r, with $0 < \beta < 1$ and $0 < r$, and leads to the following probability mass function:

$$P(X_w = x|r,\beta) = \frac{\Gamma(r+x)}{x!\Gamma(r)}(1-\beta)^r\beta^x$$
$$\forall x = 0, 1, 2, \cdots, \text{ where } \Gamma \text{ is the gamma function}$$

Whenever r is an integer, the Negative Binomial can be thought of as a generalization of the Geometric distribution. It stands for the number of success in a sequence of Bernoulli trial before r failure occur. We can also understand the Negative Binomial as a 'flatten' Poisson distribution where the parameter r controls the distribution variance. Figure 2.6 shows the graph of several negative binomial distributions.
The Negative Binomial can also be viewed as an infinite mixture of Poisson distribution: it can be derived from the following hierarchical model.

$$\begin{aligned} \lambda &\sim \text{Gamma}(r, \beta/(1-\beta)). \\ X_w|\lambda &\sim \text{Poisson}(\lambda) \end{aligned} \tag{2.3}$$

Then, by integrating out the Gamma distribution:

$$\begin{aligned} P(X_w = x|r,\beta) &= \int_0^\infty \frac{e^{-\lambda}\lambda^x}{x!}\lambda^{r-1}\frac{e^{-\lambda\frac{1-\beta}{\beta}}}{\Gamma(r)}\frac{(1-\beta)^r}{\beta^r}\mathrm{d}\lambda \\ &= \frac{(1-\beta)^r\beta^{-r}}{x!\,\Gamma(r)}\underbrace{\int_0^\infty \lambda^{r+x-1}e^{-\lambda/\beta}\mathrm{d}\lambda}_{\propto \text{Gamma}(r+x,\beta)} \\ &= \frac{(1-\beta)^r\beta^{-r}}{x!\,\Gamma(r)}\beta^{r+x}\Gamma(r+x) \\ &= \frac{\Gamma(r+x)}{x!\,\Gamma(r)}(1-\beta)^r\beta^x \end{aligned}$$

So, the Negative Binomial can be seen as a compound distribution: a Poisson distribution marginalized by a Gamma distribution. Several methods have been proposed in order

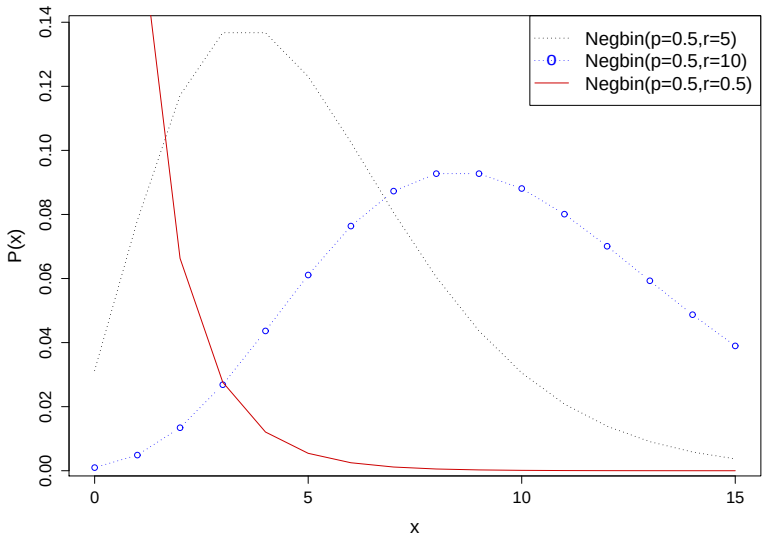


Figure 2.6: Negative Binomial Distribution

to estimate the Negative Binomial parameters. Johnson and Kotz [65] introduced two methods. The first one is the classical method of moments. The second one is a modified method of moments, where the empirical variance is replaced by the inverse document frequency. Church and Gale built on this method of modified moment and considered also the empirical mean and 5 variability measures which are *the variance, the IDF, the entropy, a burstiness measure and an adaptation measure*:

Mean $E[P(X_w|r,\theta)] = r\frac{\beta}{1-\beta}$

Variance $E[(P(X_w|r,\theta) - E[P(X_w|r,\theta)])^2] = r\frac{\beta}{(1-\beta)2}$

IDF $-\log_2 P(X_w \geq 1) = -\log(1-(1-\beta)^r)$

Entropy $-\sum_{x=0}^{+\infty} P(X_w = x)\log_2 P(X_w = x)$

Burstiness $\frac{E_P[X_w]}{P(X_w \geq 1)}$

Adaptation $\frac{P(X_w \geq 2)}{P(X_w \geq 1)}$

Church credits Katz with the burstiness measure which is simply the mean frequency in documents where the word appears at least once. The *measure of adaptation*: $\frac{P(X_w \geq 2)}{P(X_w \geq 1)} = P(X_w \geq 2 | X_w \geq 1)$, which is the probability of observing at least 2 occurrences knowing we observed at least one. To conclude on the estimation, Church observed, in practice, that the generalized method of moments with IDF was more robust than the classical method of moments.

Summary

The Negative Binomial is a generalization of the 2-Poisson mixture model. Church and Gale argued that a single Poisson is not enough to model the bursts in frequencies (ie the

elite set): one need an infinite number of Poisson. The good behavior of the Negative Binomial distribution for text processing has also been observed in several recent works. [32] uses respectively a binomial, a Poisson and a Negative Binomial distribution to model the probability of words given classes in a Naïve Bayes classifier. Rigouste [69] reproduces the experiments reported in [13] on different collections. The Negative Binomial is shown to provide a better fit to the data. Church also showed that n Poissons is not enough to model word frequencies. This may also suggest that n multinomials, as in topic models, is not sufficient.

2.3.5 K-mixture

Katz, who introduced the concept of burstiness, proposed several models of word frequencies, mainly based on Geometric distributions. The assumption of theses models is to consider the ratio $\frac{P(X_w=x+1)}{P(X_w=x)}$ to be a constant. We could say that he assumed that adaptation does not depend on x. Let's quote Katz [45] before introducing the K-mixture:

> When a particular word is used topically, occurrence of its additional instances, in the remainder of the document depends on whether or not there is anything left to be said about the concept associated with this word, not on how much has been said so far. Therefore, a high number of instances of some word that have already occurred in a document would not necessarily mean that occurring of additional instances is unlikely. For example, ten occurrences of a particular word or a phrase in one document is a very infrequent event in comparison with two occurrences of the same word or phrase in one document. But nine occurrences took place, the tenth occurrence does not seem less likely than the third one,when only two have already occurred. [....] Therefore, it would not be unreasonable to consider the conditional probabilities of repeats in a burst $P(k+1|k)$ for $k \geq 2$, as being *independent* of the number k, of previously observed occurrences and approximate them by some constant. The reasoning for that given above is by no means a proof of such independence but only an argument that it is sensible approximation to entertain, expecting that a good fit of the model, based on this approximation, to the data, will justify it.

In a nutshell, Katz suggest to approximate the ratio $P(x+1|x)$ (with our notation) by a constant, which implies the choice of the geometric distribution to model repetitions of a word in a document (ie term frequencies greater than 1). Based on this assumption, he expects to obtain a good fit to the data with the K-mixture. Formally, the K-mixture is a mixture between a Dirac distribution and a Geometric distribution. The probability of a number of occurrence x is given by:

$$P(X_w = x|\alpha,\beta) = (1-\alpha)\delta_{x,0} + \frac{\alpha}{1+\beta}(\frac{\beta}{\beta+1})^x \tag{2.4}$$

Methods to estimate the parameters of this distribution are presented in [13] and [55]. We do not detail them and simply give:

$$\begin{aligned} \beta &= \frac{F_w - N_w}{N_w} \\ \alpha &= \frac{N_w}{F_w - N_w}\frac{F_w}{N} = \frac{N_w}{N}\frac{F_w}{F_w - N_w} \end{aligned}$$

Let's look at the ratio $\frac{\beta}{\beta+1}$:

$$\begin{aligned}\frac{\beta}{\beta+1} &= \frac{F_w - N_w}{N_w} \times \frac{1}{\frac{F_w - N_w}{N_w} + 1} \\ &= \frac{F_w - N_w}{F_w}\end{aligned}$$

$\frac{\beta}{\beta+1}$ is the parameter of the geometric distribution: it serves to model the extra-occurrences in documents, where a word appears more than one time. It can be understood as the average repetition rate when a word appears multiple times.

Summary

The K-mixture is based on a geometric distribution by assuming a constant adaptation factor. The goal of Katz was to go beyond the 2-Poisson mixture model to obtain a better word frequency model. According to Church and Gale [13], the K-mixture gives rather similar fits to the Negative Binomial. And yet, this distribution is simpler to manipulate and estimate than the Negative Binomial, which offer an interesting alternative to the latter.

2.3.6 Pólya Urn Process and Dirichlet Compound Multinomial

The Pólya's Urn model is a process where balls are drawn from an urn and new balls are added gradually to the urn. The urn initially contains a black balls and b white balls. For each draw, the drawn ball is returned to the urn with c balls of the same color. When c is negative, then balls are removed from the urn. If $c = 0$, then this process amounts to the Binomial model. If $c = -1$, then it is a sampling scheme without replacement, ie an hypergeometric distribution.

We now consider the case where c is positive and l samples are drawn from this process. Let Y_i the i^{th} drawn ball. $Y_i = 1$ for black, 0 otherwise. Then, the probability of the sequence $Y_1, .., Y_l$ is given by:

$$P(Y_1, .., Y_l) = \frac{a(a+c)...(a+(x-1)c) \times b(b+c)...(b+(l-x-1)c)}{(a+b)(a+b+c)...(a+b+(l-1)c)}$$

where $x = \sum_i y_i$, ie the number of black balls drawn after l draws. With the previous equation, the probability of the sequence $(1, 1, 1, 0, 0)$ can be shown to be equal to the probability of observing $(0, 0, 1, 1, 1)$. So, the joint distribution is invariant under a permutation of the Y_i. Hence Y is an exchangeable sequence. Let $X = \sum_i^n Y_i$, the probability of X is:

$$P(X = x) = \frac{l!}{x!(l-k)!} \frac{a(a+c)...(a+(x-1)c) \times b(b+c)...(b+(n-x-1)c)}{(a+b)(a+b+c)...(a+b+(l-1)c)}$$

These two equations differ only by the factor $\frac{l!}{x!(l-k)!}$, which accounts for the number of possible sequences with x black balls.

Beta Binomial Model

When $c = 1$, the Polya Urn process becomes equivalent to the Beta-Binomial model. The Beta-Binomial model is defined by the following hierarchical model:

$$\begin{aligned}\pi &\sim \text{Beta}(a, b) \\ X &\sim \text{Binomial}(\pi, l)\end{aligned}$$

Marginalizing the Beta distribution:

$$\begin{aligned} P(X = x|a, b, l) &= \int_0^1 P(\pi|a, b)P(X = x|\pi, l)d\pi \\ &= \frac{l!}{x!(l-x)!}\frac{B(x+a, n-x+b)}{B(a, b)} \\ &= \frac{l!}{x!(l-x)!}\frac{\Gamma(a+x)\Gamma(l-x+b)}{\Gamma(a+b+l)}\frac{\Gamma(a+b))}{\Gamma(a)\Gamma(b)} \end{aligned} \tag{2.5}$$

The Beta-Binomial model is more general than the Polya urn scheme in the sense that a and b can have real values instead of integer ones, but it is more restrictive since it assumes that $c = 1$. The Beta-Binomial model can be estimated by the methods of moments as shown by Jansche [41]. First, a different parametrization is used [41]:

$$\begin{aligned} p &= \frac{a}{a+b} \\ \gamma &= \frac{1}{a+b+1} \end{aligned}$$

The mapping to the previous parametrization is given by:

$$\begin{aligned} a &= p\frac{1-\gamma}{\gamma} \\ b &= (1-p)\frac{1-\gamma}{\gamma} \end{aligned}$$

The mean and the variance are then given by:

$$\begin{aligned} E(X|p, \gamma, l) &= lp \\ Var(X|p, \gamma, l) &= lp(1-p)(1+(n-1)\gamma) \end{aligned}$$

Given that the Binomial variance is $lp(1-p)$, this shows that the Beta Binomial model can account for extra variance. Regarding estimation, the method of moments [41] gives:

$$\hat{p} = \frac{\sum_d x_d}{\sum_d l_d} \tag{2.6}$$

$$\hat{\gamma} = \frac{\sum_d (x_d - l_d\hat{p})^2/(\hat{p}(1-\hat{p})) - \sum l_d}{\sum_d (l_d)^2 - \sum_d l_d} \tag{2.7}$$

Furthermore, Jansche [41] proposed to use a mixture of a Dirac distribution modeling the zeros and a BetaBinomial for the 'true' occurrences.

Dirichlet Multinomial (DCM)

There exists a multivariate extension to the Beta Binomial model, known as the Dirichlet-Multinomial distribution. Concerning text modeling, Madsen [53] proposed to use the Dirichlet Multinomial (which they call Dirichlet Compound Multinomial (DCM)) in order to model burstiness in the context of text categorization and clustering. The Dirichlet Multinomial is defined by:

$$\begin{aligned} \theta &\sim \text{Dirichlet}([\alpha_w]) \\ X^d|\theta &\sim \text{Multinomial}(\theta, l_d) \\ P(X^d = |\alpha, l_d) &= \frac{l_d!}{\prod_{w=1}^M x_w!}\frac{\Gamma(\sum_{w=1}^M \alpha_w)}{\Gamma(\sum_{w=1}^M \alpha_w + x_w)}\prod_{w=1}^M \frac{\Gamma(x_w + \alpha_w)}{\Gamma(\alpha_w)} \end{aligned} \tag{2.8}$$

The maximum likelihood estimator of α can be shown to follow the following fixed point equation (cf Minka [59]):

$$\alpha_w = \alpha_w \frac{\sum_d \Psi(x_{wd} + \alpha_w) - \Psi(\alpha_w)}{\sum_d \Psi(x_{wd} + \sum_{w'} \alpha_{w'}) - \Psi(\sum_{w'} \alpha_{w'})} \tag{2.9}$$

where Ψ is the digamma function. In practice, this estimator is quite slow to converge, due to the presence of digamma function and to the fact that all dimensions of α are tied together in the denominator of the previous function. Thus, there are M fixed point equation to solve.

EDCM

To speed up learning time, Elkan [31] then approximated the DCM distribution by the EDCM distribution , and showed the good behavior of the model obtained on different text clustering experiments.

The motivation of the EDCM is the following: most α_w values, estimated by maximum likelihood, are closed to zero ($0 < \alpha_w \ll 1$). As,

$$\lim_{\alpha \to 0} \frac{\Gamma(x + \alpha)}{\Gamma(\alpha)} - \Gamma(x)\alpha = 0$$

the DCM distribution could be approximated by:

$$P(X^d|\alpha, l_d) \approx l_d! \frac{\Gamma(\sum_{w=1}^M \alpha_w)}{\Gamma(\sum_{w=1}^M \alpha_w + x_w)} \prod_{w=1}^M \frac{\alpha_w}{x_w} \tag{2.10}$$

The right hand side of this equation is called the *EDCM distribution*. Elkan admits it is not a proper distribution but believes that the approximation is good enough to consider this function as a probability distribution. He then used a maximum likelihood method to estimate the parameters of the EDCM 'distribution'. Let $s = \sum_{w=1}^M \alpha_w$, then s verifies the following fixed-point equation:

$$s = \frac{\sum_{w,d} I(x_{wd} > 1)}{\sum_d \Psi(s + l_d) - N\Psi(s)} \tag{2.11}$$

Once s is known, the α_w can be obtained directly by:

$$\alpha_w = \frac{\sum_d I(x_{wd} > 1)}{\sum_d \Psi(s + l_d) - N\Psi(s)} \tag{2.12}$$

Hence, the EDCM model is much faster to estimate: there is only one fixed point iteration. Elkan then proposed a mixture of EDCM distribution to model a corpus and derive an EM-like algorithm to estimate parameters.

Summary

The BetaBinomial model and its multivariate extension can be viewed as simple extension of standard multinomial models by marginalizing a Beta prior. This lead to distributions that can account for larger variance compared to the multinomial case. Nevertheless, we did not find an explicit motivation or argument for the choice of such distributions in [53] or[31] compared to the Negative Binomial, except that these distributions are supposed to better fit textual data than multinomial distributions. Pólya urn behaviour, which gradually reinforced the word probability, may lead to think that these distributions account for the adaptation phenomenon.

2.3.7 Summary

We have discussed in this section the 2-Poisson, Negative Binomial, K-mixture and Polya Urn models. All these models are claimed to be better suited for the task of modelling word frequencies and some of them explicitely aim at modelling word burstiness.

2.4 A Formal Characterization of Burstiness

In section 2.3.1, we discussed the notions of burstiness and adaptation . Recall that Katz gave this definition:

> *document-level burstiness*, i.e. multiple occurrences of a content word or phrase in a single text document, which is contrasted with the fact that most other documents contain no instances of this word or phrase at all

whereas Church [13] compared words to a contagious disease, a behavior he called adaptation. Our approach and goals are similar to the ones of Church [13], and Madsen [53] but different from Sarkar [74], who studied within document burstiness. Our goal is to obtain for each word a probability distribution for its occurrences in the collection, with the requirement that these distributions account for burstiness. To do so, we propose a *definition of burstiness as a property* of probability distributions. Then, we suggest a discrete bursty distribution: *the Beta Negative Binomial*, and a continuous distribution: *the Log-Logistic* distribution

2.4.1 Definition of Burstiness

Several models tried to take into account burstiness, but few formal definitions were proposed. More formally, for a word probability distribution $P(X_w)$, [13] measures its burstiness through the quantity:

$$B_P = \frac{E_P[X_w]}{P(X_w \geq 1)}$$

where E_P denotes the expectation with respect to P. This measure provides a way to compare two different word distributions with respect to burstiness, but does not give a clear measure on whether a given word distribution accounts or not for bursty and non-bursty words.

To introduce our definition of burstiness, we first discuss an experiment by Manning [55]. He looked at the term *soviet* and its successive ratio of $P(X_w \geq x)/P(X_w \geq x+1)$.

$P(X_w \geq 0)/P(X_w \geq 1)$	$P(X_w \geq 1)/P(X_w \geq 2)$	$P(X_w \geq 2)/P(X_w \geq 3)$
23.4	2.38	1.63

His point is to criticize the K-mixture assumptions. According to K-mixture assumption, this ratio should be constant. But, for the word *soviet*, this ratio decreases, ie its inverse $P(X_w \geq x+1)/P(X_w \geq x)$ increases.

Our definition of burstiness is also motivated by Church comparison of words to a contagious disease [12].

> "But **text** is more like a **contagious** disease [...]. If we see one instance of a contagious disease such as tuberculosis in a city, then we would not be surprised to find quite of few more"

Hence, we want to capture behavior such as *the more we have, the more we should get*. This lead us to the following definition:

Definition 1. *[Discrete case] A discrete distribution P is bursty iff for all integers $(n', n), n' \geq n$:*

$$P(X \geq n' + 1 | X \geq n') > P(X \geq n + 1 | X \geq n)$$

This definition directly translates the fact that a word is bursty if it is easier to generate it again once it has been generated a certain number of times. Note that this definition can be seen as a generalization Church's adaptation measure $P(X \geq 2|X \geq 1)$ [13]. In other words, adaptation is measured for all integers n and these adaptation rates should be increasing for a word distribution to be bursty.

In practice, however, it is not always easy to compute $P(X \geq n + 1|X \geq n)$ and determine whether a particular word distribution can account for burstiness. The following property can be used to do so:

Property 2. Characterization of Burstiness
Let $P(X_w)$ be a frequency distribution for word w and let $a_n = \frac{P(X_w=n+1)}{P(X_w=n)}$.

(i) If a_n is increasing, then w is bursty under P

(ii) If a_n is decreasing, then w is not bursty under P

Proof We have: $P(X \geq n + 1|X \geq n) = \frac{P(X \geq n+1)}{P(X \geq n)} = \frac{1}{\frac{P(X=n)}{P(X \geq n+1)} + 1}$

But:

$$\frac{P(X \geq n + 1)}{P(X = n)} = a_n + a_n a_{n+1} + \tag{2.13}$$

$$\frac{P(X \geq n + 1)}{P(X = n)} = a_n + a_n a_{n+1} + \cdots \; ; \; \frac{P(x_i \geq n + 2)}{P(X = n + 1)} = a_{n+1} + a_{n+1} a_{n+2} + \cdots$$

$$\frac{P(X \geq n + 2)}{P(X = n + 1)} \geq \frac{P(X \geq n + 1)}{P(X = n)}$$

and hence: $\forall n \in N, n \geq n_0, P(X \geq n + 2|X \geq n + 1) \geq P(X \geq n + 1|X \geq n)$ which establishes (i).
Similarly, for (ii) we obtain: $\forall n \in N, P(X \geq n + 2|X \geq n + 1) \leq P(X \geq n + 1|X \geq n)$ which proves (ii).

We now generalize the discrete definition to the continuous case as follows :

Definition 3. *[General case] Let X a random variable defined on $\mathbb{R}$ with distribution P. The distribution P is bursty iff $\forall \epsilon > 0$, the function g_ϵ defined by:*

$$g_\epsilon(x) = P(X \geq x + \epsilon | X \geq x)$$

is a strictly increasing function of x. A distribution which verifies this condition is said to be bursty. (The same definition applies to discrete distributions except that $\epsilon \in \mathbb{N}$).

This translates the fact that, with a bursty distribution, it is easier to generate higher values of X once lower values have been observed. We can develop the continuous definition of burstiness with the following equations:

$$\begin{aligned} g_\epsilon(x) \text{ strictly increasing } &\iff \Delta = \log g_\epsilon(x) \text{ strictly increasing} \\ &\iff \Delta = \log P(X > x + \epsilon) - \log P(X > x) \text{ is increasing} \end{aligned}$$

We refer to Δ as the successive difference in log probability. As $\Delta < 0$, absolute values of successive difference Δ decreases. Figure 2.7 shows the graph of two distributions: a Log-Logistic [2] and a Gaussian distribution. Distributions are plotted with coordinates $(x, \log P(X > x))$. The vertical segments indicate absolute values of successive difference, ie Δ. A condition for a distribution to be bursty is to be log-convex:

Theorem 4. *Let P be a probability distribution of class C^2. A necessary and sufficient condition for P to be bursty is:*

$$\frac{\partial^2 \log P(T > t)}{\partial t^2} > 0$$

Proof Let $f(x) = \log P(X > x)$. As the logarithm is an increasing function, the burstiness property can be expressed as:

$$g_\epsilon(x) \text{ strictly increasing} \iff \log g_\epsilon(x) \text{ strictly increasing}$$

$$\begin{aligned} &\forall\ x, \epsilon > 0 \qquad \text{the function } f(x+\epsilon) - f(x) \text{ grows} \\ &\forall\ x, \epsilon > 0 \qquad f'(x+\epsilon) - f'(x) > 0 \\ &\qquad\qquad\quad f'(x+\epsilon) > f'(x) \\ &\qquad\qquad\quad \text{ie } f' \text{ grows} \iff f'' > 0\ \square \end{aligned}$$

Under regularity assumptions, this conditions is necessary and sufficient and this convexity condition can be observed on the plots.

Application of the Theorems

Using this property, it is easy to see that the Binomial, Poisson and Geometric distributions cannot account for burstiness.

- $P(X_w) = \mathbf{Binomial}(L, p_w)$

$$P(X_w = n) = \binom{L}{n} p_w^n (1 - p_w)^{L-n}$$

$$\forall n \leq L, a_n = \frac{(L-n)p_w}{(n+1)(1-p_w)}$$

  a_n is strictly decreasing , which shows that the binomial distributions does not account for burstiness as claimed in [31].

- $P(X_w) = \mathbf{Poisson}(\lambda_w)$

$$P(X_w = n) = e^{-\lambda_w} \frac{\lambda_w^n}{n!}$$

$$\forall n, a_n = \frac{\lambda_w}{n+1}$$

  a_n is strictly decreasing so the Poisson is not bursty.

- $P(X_w) = \mathbf{Geometric}(p_w)$

$$P(X_w = n) = p_w (1 - p_w)^n$$

$$\forall n, a_n = (1 - p_w)$$

  a_n is constant. Hence Geometric distributions are neutral wrt burstiness.

[2] this distribution is detailed in section 2.4.3

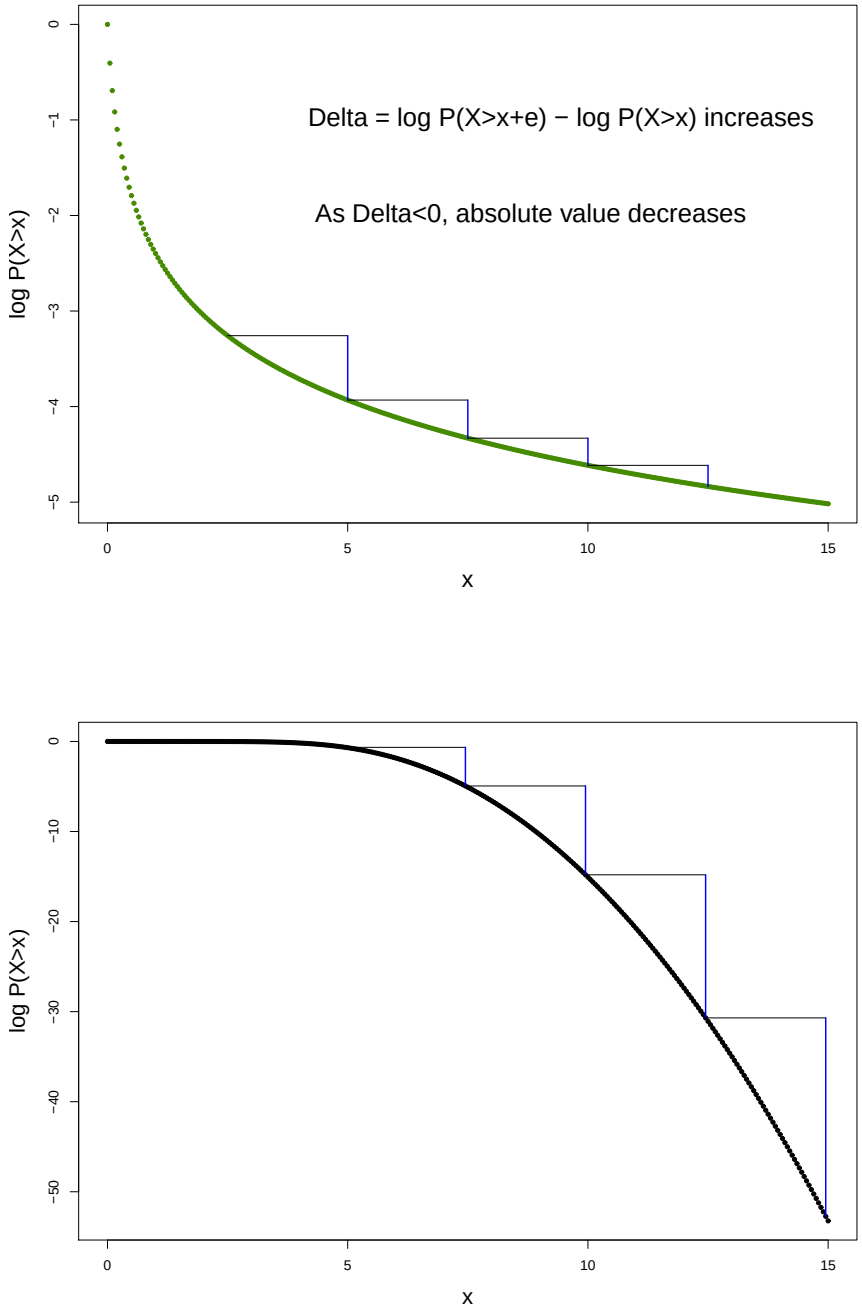


Figure 2.7: Geometrical Interpretation of burstiness: top figure shows a bursty distribution (log-logistic) and bottom figure shows that the Gaussian(mean=5, std=1) is not bursty

Negative Binomial

$$\forall n, a_n = \frac{P(X_w = n+1)}{P(X_w = n)} = \frac{\beta_w(r_w + n)}{n+1}$$

a_n is strictly increasing *iff* $r_w < 1$, strictly decreasing *iff* $r_w > 1$ and constant else. This shows that the negative binomial can account for bursty words and non-bursty words, according to the value of r.

Interestingly, the Beta Binomial model is not guaranteed to be bursty as shown in table 2.2 although it can account for more variance. We recall here the variance of the Binomial and Beta Binomial distributions:

$$\begin{aligned} \text{Binomial } Var(X|p,l) &= lp(1-p) \\ \text{Beta Binomial } Var(X|p,\gamma,l) &= lp(1-p)(1+(n-1)\gamma) \end{aligned}$$

So, extra-variability offered by the Beta-Binomial model does not always translate in burstiness. However, with setting such as $l_d = 20$, $a = 0.0004$, $b = 0.005$, the Beta Binomial model seems to be bursty. A formal proof needs to be investigated in order to find under which settings a Beta Binomial is bursty.

Table 2.2: The BetaBinomial is not guaranteed to be bursty. Beta Binomial with parameter $n = 20$ $a = 3$ $b = 7$

x	$P(x)$	$P(x)/P(x-1)$
0	0.0230	-
1	0.0531	2.3077
2	0.0806	1.5200
3	0.1008	1.2500
4	0.1118	1.1087
5	0.1138	1.0182
6	0.1084	0.9524
7	0.0975	0.9000
8	0.0834	0.8553
9	0.0680	0.8148
10	0.0528	0.7765

Summary

To conclude, table 2.3 shows whether standard distributions for text are bursty or not and the motivation of the previous definitions were:

1. to give a formal proof that state of the art models, such as Poisson and Binomial models, are unable to model the burstiness phenomenon in a collection of documents.
2. to understand when a distribution is bursty or not according to its parameter. For example, the Negative Binomial distribution [13] can be bursty or non-bursty, depending on one parameter value.
3. to help designing new distributions for word frequencies by checking their burstiness property.

Having introduced a formal definition of burstiness, we now present the Beta Negative Binomial and the Log-Logistic distributions, which are two bursty distributions according to our definition.

Table 2.3: Burstiness of Probability Distributions

Distribution	Burstiness
Poisson	No
Binomial	No
Geometric	Neutral
Negative Binomial ([13])	Depends on Parameter (Yes if $r < 1$)
Dirichlet Compound Multinomial ([53])	Depends on Parameter
Log-Logistic	Yes (when $\beta = 1$)
Exponential	Neutral
Weibull	Depends on Parameter
Pareto	Yes
Beta Negative Binomial	Yes

2.4.2 Beta Negative Binomial

Recall that the Negative Binomial distribution is given by:

$$\text{NegBin}(x|r,\beta) = \frac{\Gamma(r+x)}{x!\Gamma(r)}(1-\beta)^r\beta^x \tag{2.14}$$

An interesting extension to the Negative Binomial distribution consists in considering that the parameter β arises from a prior $Beta(a,b)$ distribution. In this case, the resulting distribution has the form:

$$P(X_w = x|r,a,b) = \frac{\Gamma(r+x)\Gamma(a+x)}{x!\Gamma(r)\Gamma(a)\Gamma(b)} \times \frac{\Gamma(a+b)\Gamma(r+b)}{\Gamma(a+b+r+x)} \tag{2.15}$$

where $x = 0, 1, 2, \cdots$, and a and b represent the two parameters of the prior Beta distribution. Assuming that this prior is uniform (ie $a = b = 1$), one obtains the following one-parameter distribution, which we will refer to as the **Beta Negative Binomial distribution** , or **BNB** in short[3]:

$$P(X_w = x|r) = \frac{r}{(r+x+1)(r+x)} \tag{2.16}$$

Figure 2.8 displays the probability plot of a BNB for several values of r. Figure **??** compares a Poisson distribution with a BNB distribution, whose parameters are equal to each other and where the power law behavior of the BNB can be observed: we can notice that the Poisson law assigns little probabilities to large frequencies, which is not the case of the BNB distribution.

Regarding the BNB distribution burstiness, $a_n = \frac{r+n}{r+n+2}$ is strictly increasing. So, the BNB distribution can model burstiness.

The maximum likelihood method leads to a fixed point equation.

$$\hat{r} = \text{argmax}_r L(\mathcal{D}, r) = \text{argmax}_r \prod_d \frac{r}{(r+x_d)(r+x_d+1)}$$

This likelihood can be rewritten in order to distinguish the contribution of zeros and

[3]This distribution is sometimes referred to as the Johnson distribution, inasmuch as it was studied by N. Johnson in [43].

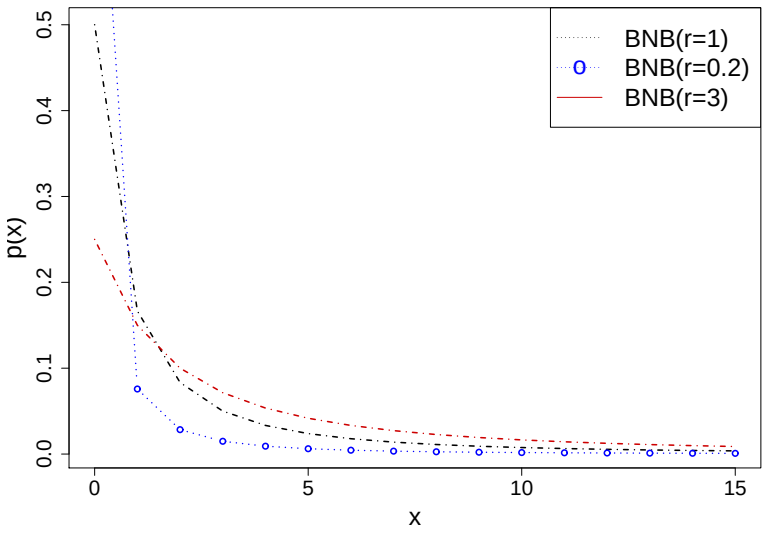


Figure 2.8: Beta Negative Binomial Distribution

non-zeros observations:

$$\begin{aligned}
\mathcal{L} &= (N - N_w)\log(\frac{1}{r+1}) + \sum_{d,x_d>0}[\log r - \log(r+x_d) - \log(r+x_d+1)] \\
\frac{\partial \log L}{\partial r} &= -\frac{N-N_w}{r+1} + \sum_{d,x_d>0}\frac{1}{r} - \frac{1}{r+x_d} - \frac{1}{r+x_d+1} \\
\frac{\partial \log L}{\partial r} = 0 \quad &\Rightarrow \\
r &= \frac{N_w}{\frac{N-N_w}{r+1} + \sum_{d,x_d>0}\frac{1}{r+x_d} + \frac{1}{r+x_d+1}}
\end{aligned} \tag{2.17}$$

Estimators

The maximum likelihood of the BNB distribution lead to a fixed point equation. We want to show here that this fixed point equation has a unique solution.
Let f the function defined by:

$$\begin{aligned}
\mathbb{R}^+ &\rightarrow \mathbb{R}^+ \\
r &\rightarrow \frac{N_w}{\frac{N-N_w}{r+1} + \sum_{d,x_d>0}\frac{1}{r+x_d} + \frac{1}{r+x_d+1}}
\end{aligned}$$

We can show that f is increasing and $f(0) > 0$ and $\text{Lim}_{r\rightarrow+\infty}\ f(r) = +\infty$

Let $F = \sum_{d,x_d>0} x_d$, then

$$\begin{aligned} \frac{1}{r+x_d} &\geq \frac{1}{r+F} \\ \sum_{d,x_d>0} (\frac{1}{r+x_d} + \frac{1}{r+x_d+1}) &\geq \frac{N_w}{r+F} + \frac{N_w}{r+F+1} \\ f(r) &\leq \frac{N_w}{\frac{N-N_w}{r+1} + \frac{N_w}{r+F} + \frac{N_w}{r+F+1}} \end{aligned} \tag{2.18}$$

This lead to the following upperbound on f:

$$\begin{aligned} f(r) &\leq \frac{1}{\frac{N-N_w}{N_w(r+1)} + \frac{1}{r+F} + \frac{1}{r+F+1}} \\ f(r) &\leq \frac{1}{\frac{1}{r+F} + \frac{1}{r+F+1}} \end{aligned}$$

But let's define the function g by:

$$g(r) = \frac{1}{\frac{1}{r+F} + \frac{1}{r+F+1}} = \frac{1}{\frac{r+F+1+r+F}{(r+F)(r+F+1)}} = \frac{r^2+2Fr+r+F^2+F}{2r+2F+1} \tag{2.19}$$

Now we want to find a solution for $g(r) = r$:

$$r^2 + 2Fr + r + F^2 + F = r(2r + 2F + 1) \tag{2.20}$$
$$r^2 + 2Fr + r + F^2 + F = 2r^2 + 2Fr + r \tag{2.21}$$
$$r^2 = F^2 + F \tag{2.22}$$

So, let $r^* = \sqrt{F^2 + F}$, then

$$f(r^*) \leq g(r^*) = r^* \tag{2.23}$$

As f is increasing, $f(0) > 0$ and $\text{Lim}_{r\to+\infty} f(r) = +\infty$, the previous inequality shows that f will cross the identity function. Thus, there exists a unique fixed point to the BNB maximum likelihood.

Note that the BNB distribution has not a finite mean nor variance. So, the methods of moments can not directly be applied here and alternative methods are needed to estimate the BNB distribution. Equating $P(X \geq 1)$ to the empirical mean document frequency with a generalized method of moments, as proposed by Church for the Negative Binomial (see section2.3.4), gives

$$\frac{N_w}{N} = 1 - P(0|r) = 1 - \frac{1}{r+1} \tag{2.24}$$

which leads to the following estimator.

$$r_w = \frac{N_w}{N - N_w} \approx \frac{N_w}{N} \text{ as } N_w \ll N \text{ for most words} \tag{2.25}$$

We have compared the maximum likelihood estimator and the mean document frequency ones for two TREC collection. Figure 2.9 shows the comparison for the ROBUST collection. Overall, there is not a significant difference between these two estimation methods.

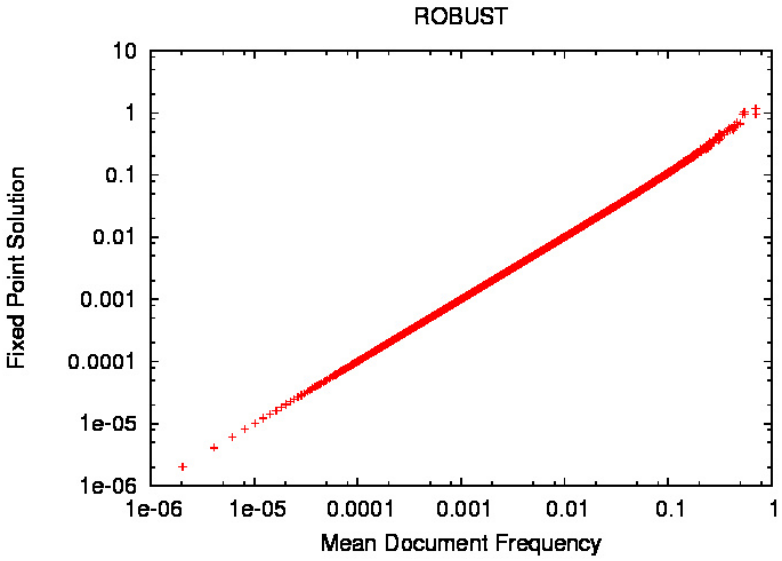


Figure 2.9: Comparison of r_w estimated by maximum likelihood to the generalized method of moments proposed for all words of the ROBUST collection. Each dot correspond to the estimated values for a given word. Correlation between the estimators is $= 0.986$, the mean difference $= 1.432e-5$, and mean relative error $= 1.3e-3$

Summary

Both the BNB distribution and DCM distribution are compound distributions. Table 2.4 shows the different distributions involved in these models. The difference between the BNB and the DCM are:

- The DCM is a multivariate model whereas the BNB is not
- The base distribution is a Negative Binomial for the BNB whereas it is a Binomial for DCM
- DCM takes into account document length on the contrary to the BNB.

	DCM	BNB
Base Distribution	Multinomial	Neg Binomial
Marginalized by	Dirichlet(Multivariate Beta)	Beta

Table 2.4: Comparison between DCM and BNB distributions

Note that there exists a multivariate extension of the Negative Binomial distribution known as Negative Multinomial [65]. The Negative Multinomial model is parametrized by the analog of the parameter r in the Negative Binomial. This r parameter is common for all words which prevents one from finely modeling the behavior of words. The second parameter is a standard multinomial parameter (θ in the Negative Binomial).

2.4.3 Log-Logistic Distribution

In IR tasks, document length normalization is a key component of an effective retrieval system. In IR models such as BM25 and DFR models, a normalized *continuous* term

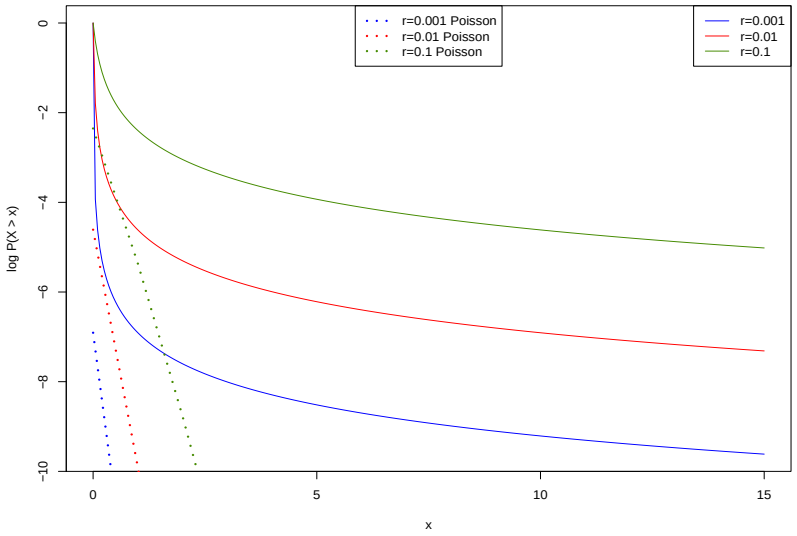


Figure 2.10: $\log P(X > x)$ Poisson(r) and Log-Logistic(r) when $r \in \{10^{-1}, 10^{-2}, 10^{-3}\}$

frequency is plugged in a discrete model. Normalization of term frequencies is a standard preprocessing step and its role is to account for the different document lengths in a collection. As most normalizations transform the frequencies in continuous values, continuous distributions seems to be more appropriate to handle word frequency data for several IR tasks.

The log-logistic distribution is the probability distribution of a random variable whose logarithm has a logistic distribution. It is similar in shape to the log-normal distribution but has heavier tails. Its cumulative distribution function can be written in closed form, unlike that of the log-normal [44]. The log-logistic distribution is defined $\forall x \in [0, +\infty)$ by:

$$P_{LL}(T < t|r, \beta) = \frac{t^\beta}{t^\beta + r^\beta}$$

Figure 2.10 compares several Poisson distributions with several Log-Logistic distributions. Tthe figure shows the power-law behavior of the Log-Logistic model.

Setting β to 1 leads to a relation between the log-logistic and the BNB distribution: $\forall x \in \mathbb{R}^+$

$$P_{LL}(t \leq T < t+1|r) = \frac{t+1}{r+t+1} - \frac{t}{r+t} = \frac{r}{(r+t+1)(r+t)} \tag{2.26}$$

which is exactly the form of the BNB distribution. This show that the Log-Logistic can be understood as a continuous variant of the BNB distribution. Furthermore, the following equation shows that the log-logistic is bursty:

$$\forall \epsilon > 0,\ g_\epsilon(x) = P_{LL}(T > t + \epsilon | T > t, r) = \frac{r+t}{r+t+\epsilon}$$

Given the relation with the BNB, we could simply estimate a log-logistic distribution by $\beta = 1$ and $r = \frac{N_w}{N}$.

Summary

The log-logistic can be seen as a continuous version of the BNB distribution, which in turn is an extension of the Negative Binomial distribution proposed by Church. The use of continuous distributions to model word frequencies in not entirely novel as Rennie [68] proposed a LogLog model. However, continuous distributions are not common for that purpose.

We proposed here a bursty distribution with a simple estimation procedure. Others continuous distributions can model word frequencies such as the Pareto distribution for example. The main benefit of continuous distributions is their ability to model normalized continuous term frequencies as DFR normalized schemas [2] or pivoted length normalization [77]. Then, the probabilistic IR model defined with continuous distributions handle valid probability values as opposed to several IR models which plug in a normalized term frequency in discrete models.

2.5 Experiments

The experiments presented here aim at assessing whether BNB and Log-logistic models are indeed appropriate to model word frequencies. We want to show that their theoretical properties (both distributions are bursty according to our definition) match empirical data. So, we want to show that these distributions are able to capture the bursty behavior of word frequencies. Criterion such as likelihood or χ^2 statistic will be examined in different experiments in order to validate the proposed models. A last experiment will look at the burstiness phenomenon against the sample size in order to assess the relation between variance and burstiness.

2.5.1 Comparison between Poisson, Katz-Mixture and BNB

Let $\mu_w = \frac{F_w}{N}$ the mean frequency of term w in a corpus. We want to illustrate here 3 different models of term occurrences:

$$\textbf{Poisson} \qquad X_w \sim Poisson(\mu_w)$$

$$\textbf{Katz-Mixture} \quad X_w \sim K\text{-}Mixture(\alpha_w, \beta_w)$$

$$\textbf{BNB} \qquad X_w \sim BNB(\mu_w)$$

These 3 models have different burstiness capacity: *non-bursty, neutral and bursty.* α_w and β_w refer to the parameter described in section 2.3.5 and correspond to the standard estimation of the Katz-Mixture. For a word in collection, we consider the following quantities:

Empirical Mean $\mu_w = \frac{F_w}{N}$ is the empirical mean of the number of occurrences.

Non-Zero Variance The empirical variance of the occurrences samples can be decomposed in two parts:

$$\sigma = \frac{1}{N+1}\Big((N - N_w)(0 - \mu_w)^2 + \sum_{d, x_{wd}>0} (x_{wd} - \mu_w)^2\Big) \tag{2.27}$$

We considered here the non-zero part of the empirical variance:

$$\sigma^* = \frac{1}{N+1} \sum_{d, x_{wd}>0} (x_{wd} - \mu_w)^2 \tag{2.28}$$

Non-Zero Likelihood For each of these models, the likelihood of non-zero observations will be computed. Recall that the likelihood for these models can be written as :

$$L(\theta) = (N - N_w) \log P(X_w = 0|\theta) + \sum_{d, x_{wd}>0} \log P(X_w = x_{wi}|\theta)$$

Focusing on non-zeros observations, this leads to:

$$L^*(\theta) = \sum_{d, x_{wd}>0} \log P(X_w = x_{wi}|\theta)$$

Term Rank The rank of term is computed by sorting $\mu_w = \frac{F_w}{N}$ in decreasing order

We now want to explain why we choose to compute non-zero likelihoods and non-zero variances. First, the Katz-mixture has a perfect fit by construction for the zero probability. Given the predominant number of zeros observations, this may bias our conclusion toward models fitting well the non-zero probability. As the burstiness phenomenon generates large frequencies, we want to capture the model performance for such events. This is why we compute the non-zero variance, so that a large variance comes only from a large frequency namely a large deviation from the mean
To sum up, 7 features can be computed for a word in a collection:

$$(rank_w, \mu_w, N_w, \sigma^*(w), L_P^*, L_G^*, L_B^*)$$

where subscript of L^* indicates the distribution (P=Poisson, G=Geometric, B=BNB).

Those 7 features were computed on different collections such CLEF 2003 Adhoc- Task CLEF 2007 Domain Specific (GIRT) and the TREC ROBUST collection.

Figures 2.11, 2.13 and 2.15 show for a given collection the term rank against the log of document frequency. A dot on these plots has coordinates $(rank_w, \log N_w)$ and a color code indicates which non-zero likelihood is maximal for this word. If L_B^* is the maximum of the three likelihoods, then the point on this graph corresponding to the word has a black color (respectively red for geometric and green for poisson). A corresponding graph also displays the same information without the color code in 3 different subplots. These plots shows that for similar ranks (ie a vertical line in the plot), words which are better modeled with a BNB have a lower document frequency. Hence, for a similar mean frequency, these words appear in less documents. They have relatively higher frequencies in documents in which they appear than words explained by a Poisson or Geometric distribution for a similar rank. This shows that the BNB distribution captures words that tend to appear with high frequency in relatively few documents and suggests that the BNB distribution is indeed *appropriate to model the burstiness phenomenon.*

A second phenomenon can also be observed on these figures. There exists different statistical behaviors for words. Some words are better explained by a Poisson model, others by a Geometric and others by a BNB distribution.

Figures 2.12, 2.14 and 2.16 show for a given collection the empirical mean against the non-zero variance. For each word w, a dot of coordinates $(x = \log \mu_w, y = \log \sigma^*(w))$ is drawn. The color of this point shows which non-zero likelihood is maximal. The same color code applies. (If L_B^* is the maximum of the three likelihoods, then the point on this graph corresponding to the word is black respectively red for geometric and green for poisson). Theses plots show again that the 3 distributions tested capture different ranges of variance for a fixed mean. For a similar empirical mean, words with larger non-zero variance are better explain by a BNB distribution. Large deviations from the mean could be explained by high frequencies, i.e. a bursty behavior of words in documents.

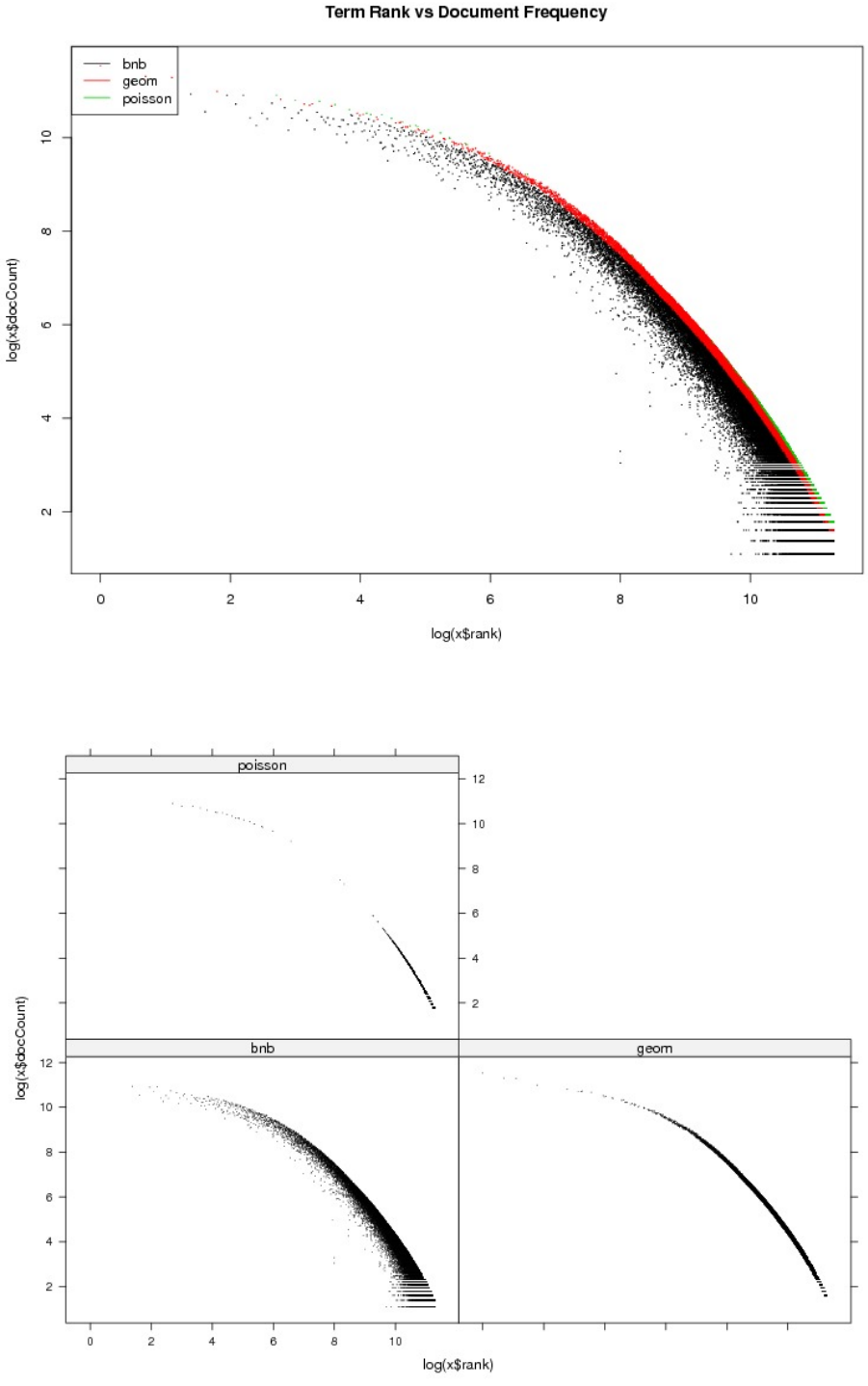


Figure 2.11: CLEF

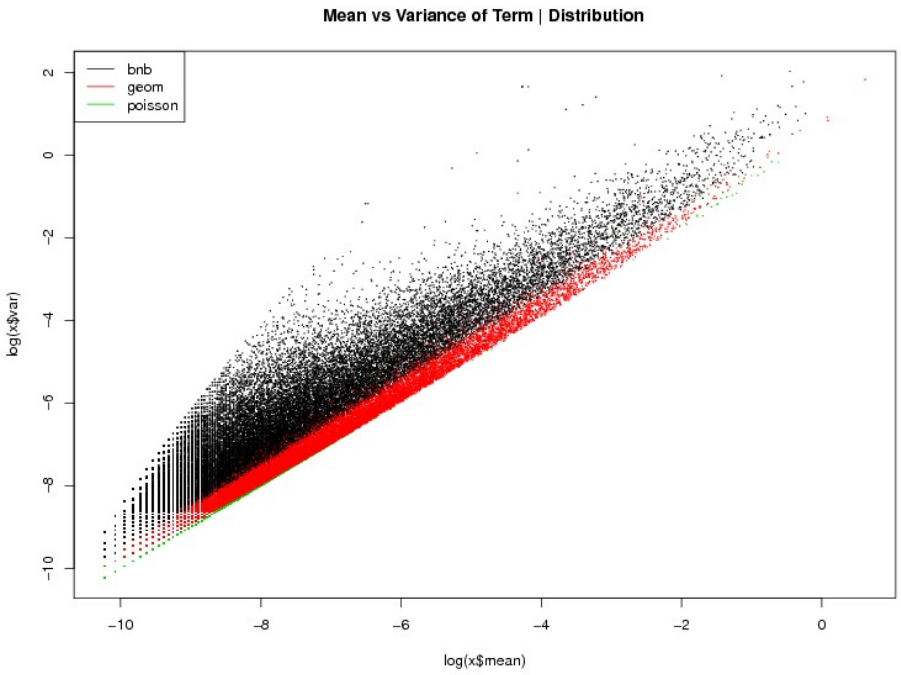


Figure 2.12: CLEF

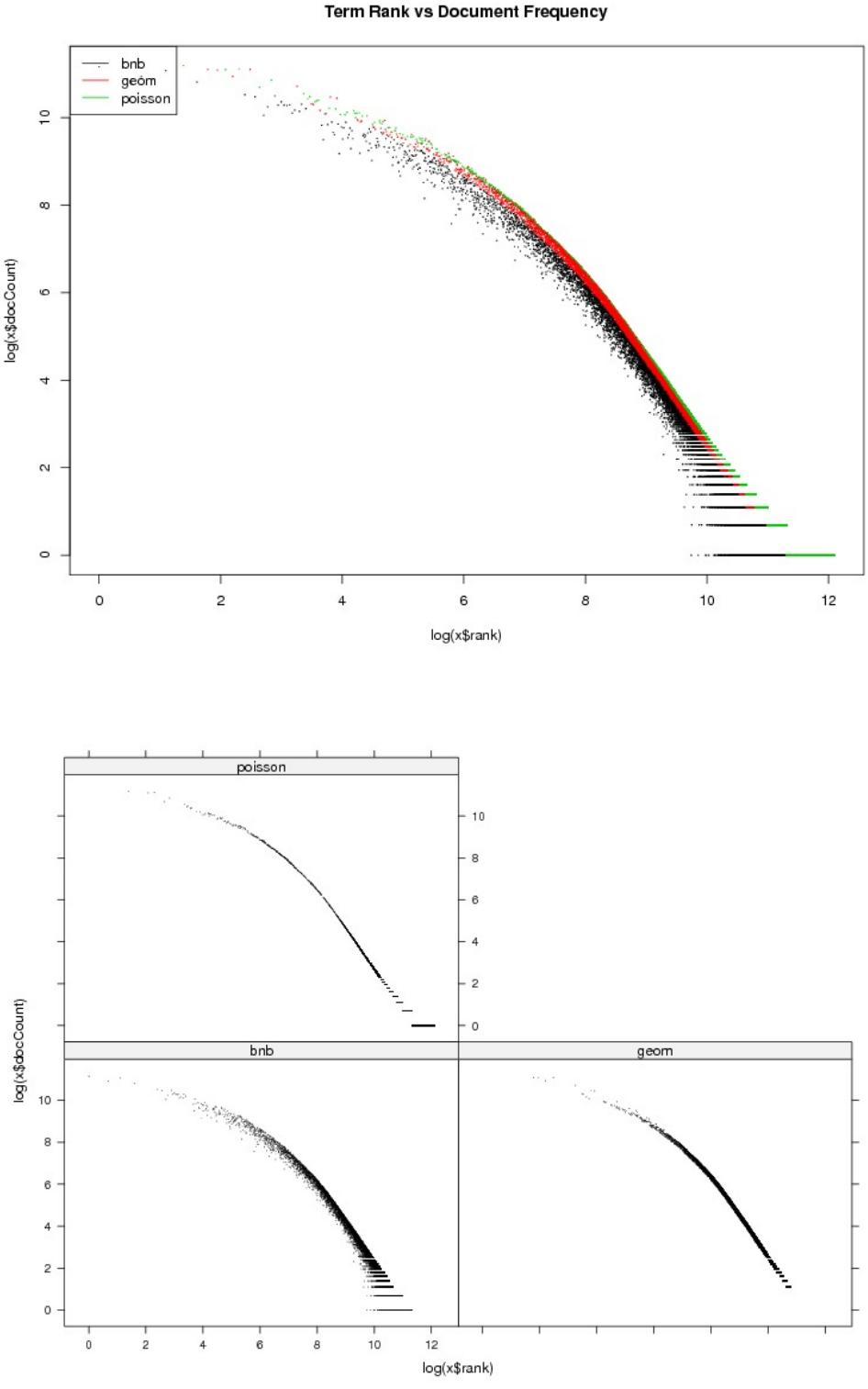


Figure 2.13: GIRT

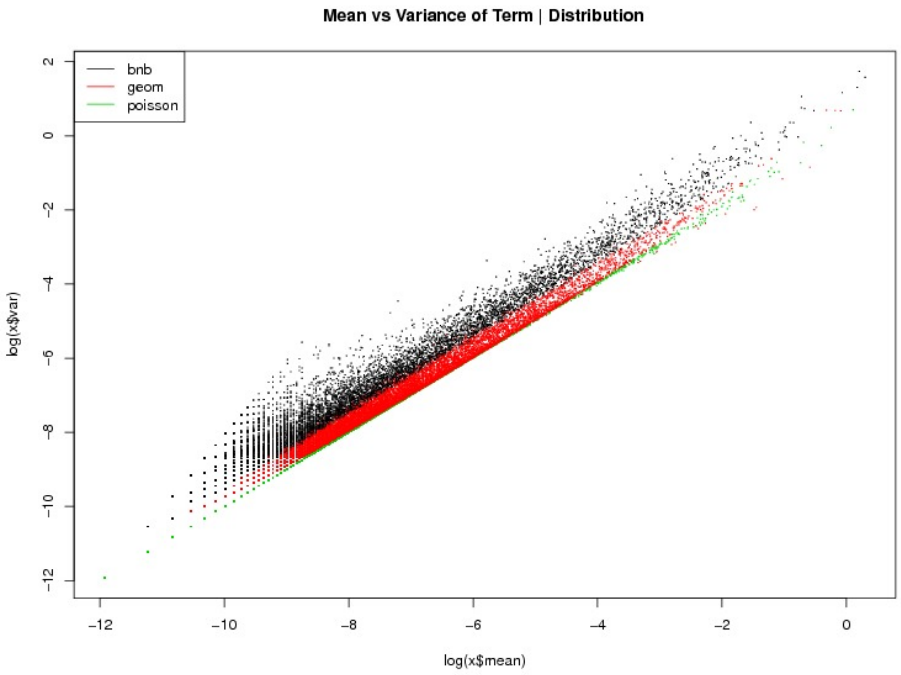


Figure 2.14: GIRT

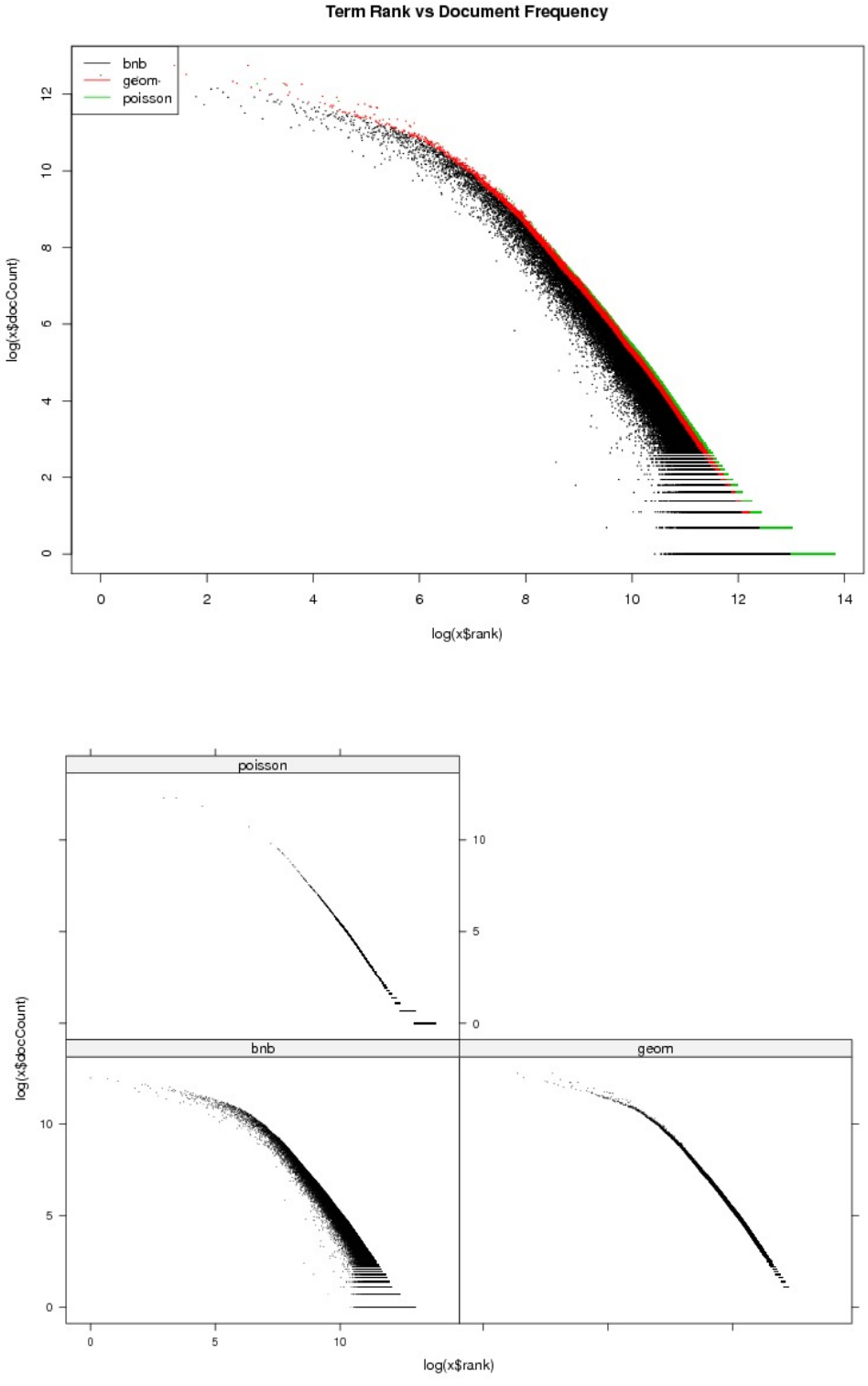


Figure 2.15: TREC-7 ROBUST

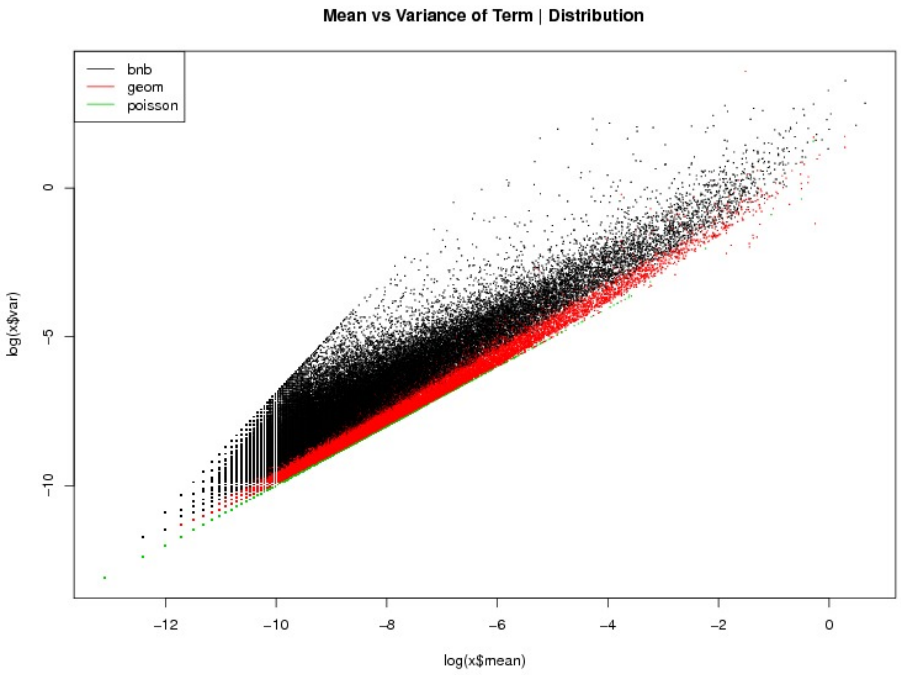


Figure 2.16: TREC-7 ROBUST

2.5.2 χ^2 Test

We illustrate here the fact that the BNB and Log-Logistic distribution, unlike others like the Poisson distribution, provides a good fit to the data. Instead or relying on the likelihood as a fit measure, we computed the Chi-square statistics for each term under both a Poisson hypothesis and a Log-Logistic one (figure 2.17). Our goal here is to see what is the fit between experimental observations and the ones predicted by these distributions: the Chi-square statistics provides us with a measure of this fit.

The Pearsonχ^2 statistic is defined by:

$$\chi^2 = \sum_i \frac{(O_i - E_i)^2}{E_i} \tag{2.29}$$

where O_i is an observed frequency and E_i an expected frequency.

We restrict our study to terms appearing at least in 100 documents of the ROBUST' TREC collection. For each selected term, we want to compare two candidate distributions modeling the term frequencies in the documents, namely the Poisson and Log-Logistic distributions. Furthermore, we assume that the parameters of these distributions are set according to:

- Poisson: $\hat{\theta_w} = \frac{F_w}{N}$
- Log-Logistic: $\hat{r_w} = \frac{F_w}{N}$

For each selected word w and document d, x_{wd} is binned into one of the following intervals: $[0, 3)$, $[3, 10)$ and $[10, 100)$. These intervals correspond roughly to low, medium and high frequency. The number of observations falling into each interval constitutes statistics that the Chi-Square compares to an expected number predicted by the assumed distribution. For each selected term, we then compute the Chi-square statistics under a Poisson hypothesis and a Log-Logistic hypothesis[4]. . We then plotted the term rank [5] against the log of the Chi-Square statistics for both the Poisson and Log-Logistic distributions. Figure 2.17 shows the log of the Chi-square statistics against the term rank for the ROBUST' TREC collection. One dot with coordinate (x, y) on the graph corresponds to a given word in the collection, where x is the term rank and y is the log of the Chi-square statistics for the distribution considered. The horizontal line is the upper critical value for the Chi-square test at the 0.05 confidence level. Note that the conditions required for χ^2 test are likely to be not satisfied for all words.

Concerning the Poisson plot, there are 2 main clouds of points. The upper left area can be explained by words from the interval $[10, 100)$: this is an extremely unlikely event under a Poisson distribution with a very small mean (ex: 0.05). The second area, which looks like a thick band, corresponds to words from the first two intervals only. As one can note, the fit provided by the BNB/log-logistic distribution is good inasmuch as the values obtained by the Chi-square statistics are small. These distributions can thus well explain the behavior of words in all the frequency ranges. The same does not hold for the Poisson, for which large values are observed over all the frequency ranges, many words getting a value above the upper critical value.

Similar results also holds for other collections. Interestingly, we do not exactly observed the same results than with the likelihood. For the likelihood method, some frequent words were better modeled with a Poisson distribution which is not the case here. The likelihood somehow computes a global behavior whereas the χ^2 statistic is very sensitive

[4] Due to relation 2.26, the Chi-square statistics *is the same for the BNB* and the log-logistic distributions on the given intervals.

[5] To display the results, we first ranked the selected terms by their frequency in the collection in order to get their term rank, as is done in Zipf's Law

to the observation of an unexpected event due to the factor $\frac{1}{E_i}$. For example, observing a frequency of 15 under a Poisson distribution of mean 0.0001 is an unlikely event. This also shows that the fit measure is an important parameter in order to evaluate word frequency models. Of course, the Poisson distribution is known for a long time to provide a poor fit to the data. It is however used in some IR models.

2.5.3 Asymptotic Behavior

The following experiments show the evolution of the average word likelihood under a Poisson model or a BNB model. For $n = 1$ to 1000, we select n documents randomly and compute non zero likelihoods. Finally, this experiment is repeated 30 times and the non-zeros likelihoods are averaged. More formally, we compute the following quantities for a set S of documents:

$$\mu_w(S) = \frac{\sum_{d \in S} x_{wd}}{|S|} \tag{2.30}$$

$$L^*_{Poisson}(S) = \sum_{d \in S} \sum_w I(x_{wd} > 0) \log P_{Poisson}(x_{wd}|\mu_w(S)) \tag{2.31}$$

$$L^*_{LL}(S) = \sum_{d \in S} \sum_w I(x_{wd} > 0) \log P_{LL}(x_{wd}|\mu_w(S)) \tag{2.32}$$

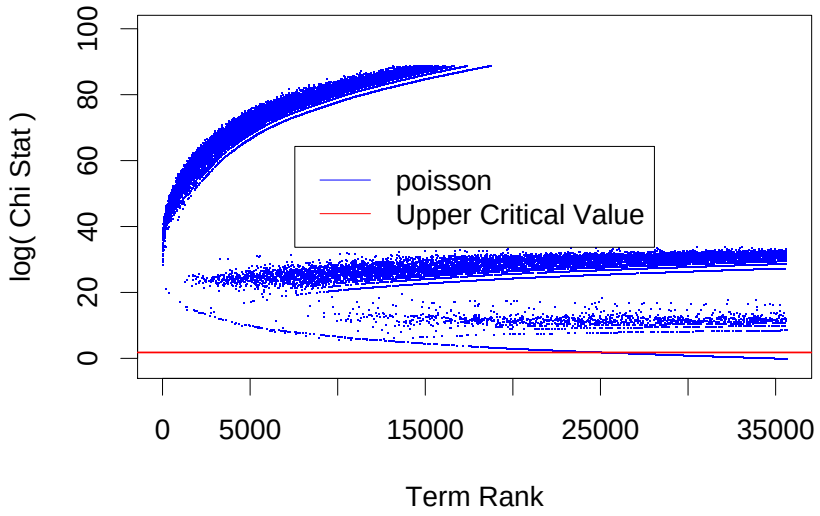
Figure 2.18 shows the difference of likelihood for three different collections. between the Poisson model and the BNB model. Those figures show that a Poisson model is much more appropriate for a small number of document, but a BNB model is more adequate for a larger sample size. As the sample size increase, the number of zeros observations also increases. It may also means that the burstiness phenomenon, described here only occurs at large sample sizes. For small samples, Poisson would fit better the data in term of non-zero likelihood. If we were to model a single document, then the best fit would probably be provided by a Poisson or Binomial, hence justifying the idea of Language Models. In the other hand, if we had to model the frequencies of a word in a collection, then the bursty distributions would probably be better.

2.6 Conclusion

This chapter surveyed the main probabilistic models of word frequencies, which are summarized in table 2.5. We mostly examined standard univariate distributions and their compounds, paying less attention to non-parametric models like Dirichlet processes [79]. Burstiness was introduced in order to reveal limitations of multinomial assumptions, in particular a limited variance range. Burstiness was discussed informally, with Katz definition and Church's experiments. We then reviewed several models adressing the burstiness phenomenon, including the Negative Binomial distribution and the Dirichlet Compound Multinomial model.

Contrary to prior studies, we have proposed a *definition of burstiness as a property* of probability distributions which relates to the log-convexity of the survival function. This definition of burstiness is acan also help to determine when a distribution is bursty or not according to its parameter values. Furthermore, it can guide the design of new distributions for word frequencies by checking their burstiness property.

Then, we have extented the Negative Binomial model with the Beta Negative Binomial distribution and the Log-Logistic model was proposed as a continuous counterpart. For both distributions, we provided constant time estimation procedure as opposed to the DCM and EDCM models. The resulting distributions provide a good fit to data compared



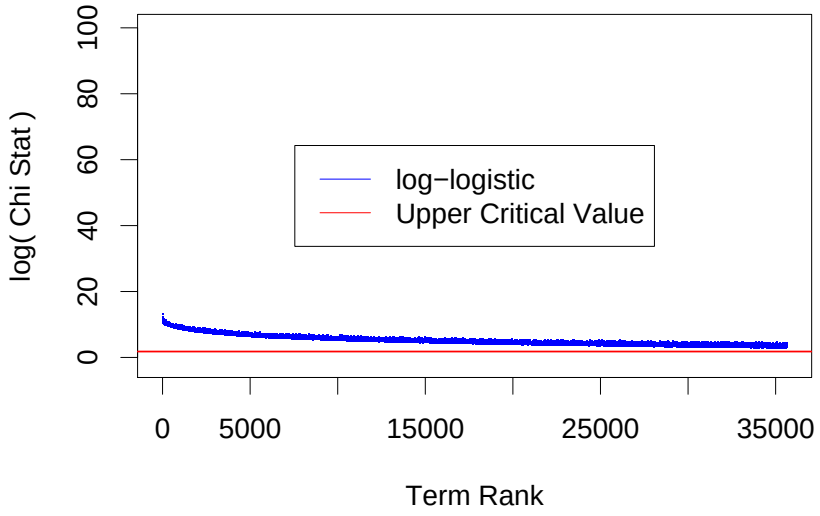


Figure 2.17: Distribution of the Chi-square statistics for the Poisson and the BNB/log-logistic distributions on the ROBUST Collection

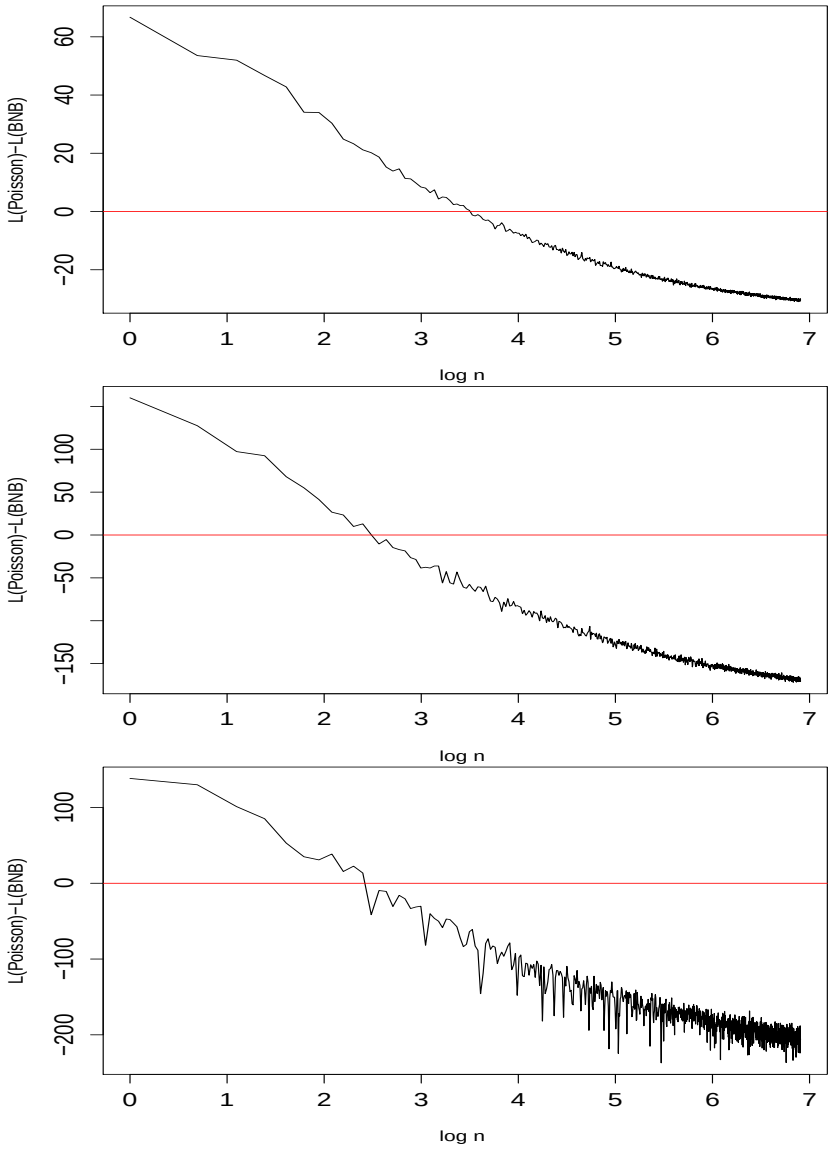


Figure 2.18: Number of Documents n against the difference of Likelihood between BNB and Poisson for the GIRT, CLEF03 and ROBUST collection

to Poisson and Katz-Mixture distributions. Overvall, the proposed models seem to have a better fit to data than distributions used in standard IR models.

Nevertheless, two important experiments are missing in our experimental design and are left for future work. First, it would be nice to compare the fit provided by a BNB model to a DCM or BetaBinomial model. Second, Durot [30] proposed a statistical test in order decide about the convexity/concavity of a survival function $P(X > x)$. As concavity is linked to the burstiness property, such a procedure is appealing because it directly tests the burstiness nature of a data sample. Preliminary experiments indicate that many words have concave empirical survival function, which suggest again that word frequencies should be modeled with bursty distributions.

Having discussed the burstiness phenomenon and selected two probability distributions, the remaining task consist in applying our candidate distributions in IR. This is why the next chapter reviews the main probabilistic IR models.

Multinomial

- $P(X^d = (x_{1d}, \ldots, x_{Md})|\theta, l_d) = \frac{l_d!}{\prod_w x_{wd}!} \prod_w \theta_w^{x_{wd}}$
- $\theta_w = \frac{x_{wd}}{l_d}$

2-Poisson

- $P(X_w = x|\alpha, \lambda_E, \lambda_G) = \alpha \frac{e^{-\lambda_E} \lambda_E^x}{x!} + (1-\alpha) \frac{e^{-\lambda_G} {\lambda_G}^x}{x!}$
- Method of Moments

Negative Binomial

- $P(X_w = x|r, \beta) = \frac{\Gamma(r+x)}{x!\Gamma(r)}(1-\beta)^r \beta^x$
- Generalized Method of Moments

K-Mixture

- $P(X_w = x|\alpha, \beta) = (1-\alpha)\delta_{x,0} + \frac{\alpha}{1+\beta}(\frac{\beta}{\beta+1})^x$
- $\alpha = \frac{N_w}{N}$ $\beta_w = \frac{F_w}{F_w - N_w}$

Beta Binomial

- $P(X_w = x|l, a, b) = \frac{l!}{x!(l-x)!} \frac{\Gamma(a+x)\Gamma(l-x+b)}{\Gamma(a+b+l)} \frac{\Gamma(a+b))}{\Gamma(a)\Gamma(b)}$
- Moment Estimators

Dirichlet Compound Multinomial

- $P(X^d = [x_w]|\alpha) = \frac{l_d!}{\prod_{w=1}^M x_w!} \frac{\Gamma(\sum_{w=1}^M \alpha_w)}{\Gamma(\sum_{w=1}^M \alpha_w + x_w)} \prod_{w=1}^M \frac{\Gamma(x_w+\alpha_w)}{\Gamma(\alpha_w)}$
- M Fixed Point Equations

EDCM

- $P(X^d|\alpha, l_d) \approx l_d! \frac{\Gamma(\sum_{w=1}^M \alpha_w)}{\Gamma(\sum_{w=1}^M \alpha_w + x_w)} \prod_{w=1}^M \frac{\alpha_w}{x_w}$
- 1 fixed point equation

Beta Negative Binomial

- $P(X_w = x|r) = \frac{r}{(r+x+1)(r+x)}$
- Generalized Method of Moments: $r_w = \frac{N_w}{N - N_w} \simeq \frac{N_w}{N}$
- Maximum Likelihood: Fixed Point Equation

Log-Logistic

- $P_{LL}(T < t|r, \beta) = \frac{t^\beta}{t^\beta + r^\beta}$
- $\beta = 1$ $r_w = \frac{N_w}{N - N_w} \simeq \frac{N_w}{N}$

Table 2.5: Main Word Frequencies Probability Distributions with their pdf and estimation methods

Chapter 3

Review of Probabilistic Information Retrieval Models

Contents

3.1 Introduction

All information retrieval systems include a query model, a document model and a function to match queries and documents. These are the 3 required ingredients of an IR engine. Figure 3.1 depicts the different components of an IR system. This chapter deals with the function matching documents and queries in a probabilistic way. Despite the fact that the machine learning approach to IR has been one of the major breakthrough in IR recently, we do not give an overview of methods à la Learning to Rank such as RankSVM, RankBoost, LambdaRank [83, 35, 81, 9]. We first focus on ad hoc retrieval since the performance of 'generative' and discriminative approaches are similar in ad hoc scenario as shown in [61]. So, we review three families of probabilistic IR models: the Probabilitic Ranking Principle, the Language models, and the Divergence From Randomness family, which are state of the art *ad hoc* information retrieval models.

These three different information retrieval families rely on word probability distributions with their own specificities. In Okapi, for example, it is assumed that word frequencies follow a mixture of two Poisson distributions. The Divergence from Randomness

(DFR) framework proposed by Amati and van Rijsbergen [2] makes use of several distributions, among which the geometric distribution, the Poisson distribution and Laplace law of succession play the major role. Language models are, for themselves, built upon the multinomial distribution, which amounts to consider binomial distributions for individual words.

Among the three families, it is the Divergence From Randomness framework that retained our attention fo reasons we will explain later. The DFR framework will serve us as a starting point for a formal analysis of IR models in the next chapter and to the elaboration of a suitable framework for the BNB and Log-Logistic distributions. Before moving on to the detailed presentation of these three families of IR models, we give a short description of the three main IR models families:

Probabilistic Ranking Principle (PRP): These models suppose the existence of a class of relevant documents and a class of non-relevant documents for a query. This idea results in ordering document with the estimated probability of relevance. This principle will be presented in the part 3.2. The pre-eminent model in this family is called *BM25* or *Okapi*

Languages Models (LM): The core idea of languages models is to estimate the probability a query is generated from a document modle $P(q|d)$. The language models are nowadays very popular. These models will be the subject of section 3.3

Divergence From Randomness: These models try to quantify the importance of a term in a document compared to its behavior in the collection. Thus, the weight of a term in a document can be measured thanks to a function of the Shannon information. These models will be presented in the section 3.4

3.2 Probability Ranking Principle

All the models based on the probability ranking principle [71] make the following assumption:

Hypothesis. *The relevance of a document to a query can be encoded by a random variable. The benefit of this formulation is to reconsider certain deficiencies of the concept of relevance; namely that the relevance is not easily definable and especially partially observable.*

This assumption deals with the function matching queries and documents. We will note R_q the random variable of relevance specific to the request q. This assumption, developed in the 70's, had a considerable impact on information retrieval models. This assumption results in ordering the documents by order of decreasing probability of relevance. This principle is called *Probability Ranking Principle* [71]. This principle results in sorting the documents for a request according to the probability $P(R = 1|X^d)$ where $X^d = (x_w)$ a representation of a document, in the form of a feature vector. In general X^d is a vector of words where the component x_{wd} represents the frequency/presence of the word w in the document d. The Probability Ranking Principle can be stated as follows [71]:

> **The probability ranking principle:** If a reference retrieval system's response to each request is a ranking of the documents in the collection in order of decreasing probability of relevance to the user who submitted the request, where the probabilities are estimated as accurately as possible on the basis of whatever data have been made available to the system for this

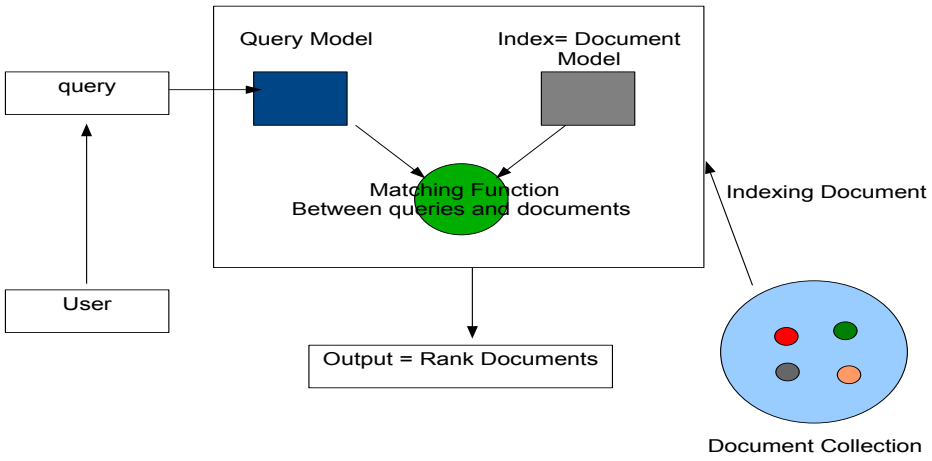


Figure 3.1: Information Retrieval System Architecture. An Information Retrieval system is composed by a query model, ie a query language/formalism, a component to index documents and a function matching queries and documents. The output of IR engine is in general a ranked list of documents.

purpose, the overall effectiveness of the system to its user will be the best that is obtainable on the basis of those data.

The probability ranking principle is a direct consequence of Bayes decision rule. Let us suppose that the probability of making a bad decision is following form:

$$P(error|X^d) = \begin{cases} P(R=1|X^d) & \text{if one chooses R=0} \\ P(R=0|X^d) & \text{if one chooses R=1} \end{cases}$$

Then, if one decides $R=0$ (document non-relevant) and that $P(R=1|X^d) > P(R=0|X^d)$, this decision leads to an error larger than the decision opposite. Thus, it is enough to choose the assumption which maximizes $P(R|X^d)$ to minimize the probability of error. By supposing that the documents are independent (statistically), this rule results in ordering the documents by decreasing probability of relevance. Figure 3.2 tries to illustrate the Probability Ranking Principle approach to IR.

One of the main limitations of the PRP is the assumption that one can calculate the probabilities $P(R|X^d)$ and this with a certain precision. This assumption is rather problematic. In general, one does not know which are the relevant documents, nor their number, or distribution. However, one could test guess these probabilities and, by test and successive corrections , improve their estimation. Nevertheless, this principle can be sub optimal as [36] shows, when probabilities are not properly calibrated.

In summary, these probabilistic models try to estimate the probability that a document is relevant. By assuming that some relevant and irrelevant documents are known, *an assumed probabilistic model* could estimate the probability that a new document is relevant or not. After a first retrieval step, users can annotate documents and probability of relevance can be updated.

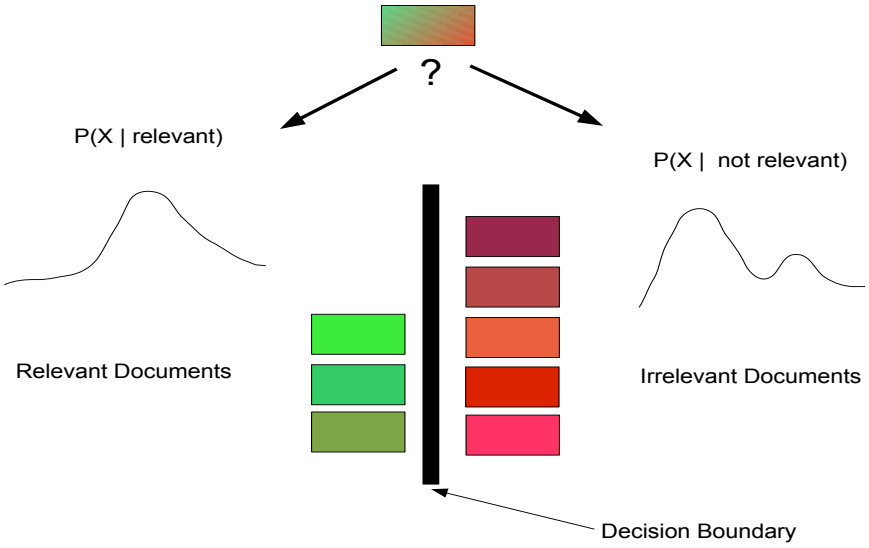


Figure 3.2: The Probability Ranking Principle. Two class of documents exists: relevant and not relevant. For each class, a probability distribution models documents. When a new document comes in, class probabilities are computed in order to decide which class this document belongs to.

It turns out that the PRP can be reformulated in a simplified form after some assumptions:

$$\text{Sort Documents by decreasing order of: } \log(\frac{P(X^d = x|R = 1)}{P(X^d = x|R = 0)})$$

This formulation will make possible to simplify this criterion by taking into account new assumptions on the dependencies between the variables X^d and R. In particular, one can suppose independence between the different terms of a document, which is similar to suppose the orthogonality of the terms in the vectorial model. Then, $P(X^d|R = r)$ can be written like a product of probabilities and the function of score as a sum on the common terms between the query and the documents. For example, Robertson centered this function so that empty documents get a null score. Thus, it leads to consider the following family of retrieval functions:

$$RSV(q, d) = \sum_{w \in q \cap d} \log(\frac{P(X_w^d = x|R = 1)}{P(X_w^d = 0|R = 1)} \frac{P(X_w^d = 0|R = 0)}{P(X_w^d = x|R = 0)})$$

Nevertheless, it always remains to clarify the probability $P(X^d|R)$, which we will do for the forthcoming models.

3.2.1 Binary Independence Model (BIR)

The Binary Independence Retrieval (BIR) supposes that the weights of the terms in the documents and request are binary ($X^d = (1010 \cdots 010 \cdots)$). Each term is characterized

by a binary variable, A_w, which indicates the probability that the word appears in a document. Moreover, each term is conditionally independent of each other given R. For two words w_1, w_2, then $P(X^d_{w_1} = x, X^d_{w_2} = y|R = r) = P(X^d_{w_1} = x|R = r)P(X^d_{w_2} = y|R = r)$

Let us note $a_w = P(A_w = 1|R = 1)$, probability that the word w appear in a relevant document and $b = P(A_w = 1|R = 0)$ probability that the word w appear in a non-relevant document.

$$P(X^d = (x_1, \ldots, x_M)|R = 1) = \prod_w a_w^{x_w}(1 - a_w)^{(1-x_w)}$$

$$P(X^d = (x_1, \ldots, x_M)|R = 0) = \prod_w b_w^{x_w}(1 - b_w)^{(1-x_w)}$$

In other words, the documents are modeled by independent Bernoulli laws. The fact that a document is relevant or not, is described by different values from the parameters of these probabilities (a_w or b_w). Under these assumptions, the PRP is expressed as:

$$RSV(q, d) = \sum_{w \in q \cap d} \log(\frac{P(X^d_w = 1|R = 1)}{P(X^d_w = 0|R = 1)} \frac{P(X^d_w = 0|R = 0)}{P(X^d_w = 1|R = 0)})$$

Under these assumptions the ranking function is:

$$RSV(q, d) = \sum_{w \in q \cap d} \log(\frac{a_w}{1 - a_w} \frac{1 - b_w}{b_w}) \tag{3.1}$$

The estimate of the probabilities a_w and b_w is done by an iterative processes:

1. Initial values are defined (for example $a^0_w = 0.5$, $b^0_w = \frac{N_w}{N}$)
2. A retrieval step is performed with the parameter current values
3. Parameters are updated. If, V is the number of relevant document found at this stage and V_w the number of relevant document containing w, then the re-estimation of the parameters becomes:

$$a_w = \frac{V_w}{V}\ ,\ b_w = \frac{N_w - V_w}{N - V}$$

The advantages of this model are a theoretically well-founded and a clear concept of relevance. Moreover, information retrieval is cast as an iterative process which involves the user. However, the model is rather sensitive to the initial values and its major disadvantage remains the binary representation of the occurrences of the words in the documents, which limits largely its performance.

3.2.2 Okapi/BM25

The BM25 [72] model reconsiders certain deficiencies of the BIR model. First, BM25 supposes that the frequencies of the words are distributed according to a mixture of 2 Poisson distribution. Moreover, it makes the assumption that in the relevant set ($R = 1$), the distribution of Poisson representing the Elite component has a weight stronger than in the not-relevant class. More formally, these assumptions result in:

$$X_d = x|R = 1 \sim 2Poisson(\alpha, \lambda_E, \lambda_G) \qquad X_d = x|R = 0 \sim 2Poisson(\beta, \lambda_E, \lambda_G)$$

$$\alpha > \beta$$

Recall that $\lambda_E > \lambda_G$

Let us recall reformulation of the PRP by Robertson:

$$RSV(q,d) = \sum_{w \in q \cap d} \log(\frac{P(X_w^d = x|R=1)}{P(X_w^d = 0|R=1)} \frac{P(X_w^d = 0|R=0)}{P(X_w^d = x|R=0)})$$

which gives by adding the assumptions on the 2-Poisson model:

$$RSV(q,d) = \sum_{w \in Q} \log(\frac{\alpha \frac{e^{-\lambda_E} x_w^{\lambda_E}}{x_w!} + (1-\alpha)\frac{e^{-\lambda_G} x_w^{\lambda_G}}{x_w!}}{\beta \frac{e^{-\lambda_E} x_w^{\lambda_E}}{x_w!} + (1-\beta)\frac{e^{-\lambda_G} x_w^{\lambda_G}}{x_w!}} \frac{\beta e^{-\lambda_E} + (1-\beta)e^{-\lambda_G}}{\alpha e^{-\lambda_E} + (1-\alpha)e^{-\lambda_G}}) \quad (3.2)$$

This model suffers from same the problems as the 2-Poisson model, ie the difficulty in estimating its parameters. Nevertheless, Robertson studies the properties of the following weighting function:

$$h(x_w) = \log(\frac{\alpha \frac{e^{-\lambda_E} x_w^{\lambda_E}}{x_w!} + (1-\alpha)\frac{e^{-\lambda_G} x_w^{\lambda_G}}{x_w!}}{\beta \frac{e^{-\lambda_E} x_w^{\lambda_E}}{x_w!} + (1-\beta)\frac{e^{-\lambda_G} x_w^{\lambda_G}}{x_w!}} \frac{\beta e^{-\lambda_E} + (1-\beta)e^{-\lambda_G}}{\alpha e^{-\lambda_E} + (1-\alpha)e^{-\lambda_G}}) \quad (3.3)$$

Knowing that $\alpha > \beta$, one can show that this function is an increasing function of the frequency of the x_w term. Moreover, the limit of h, when x_w tends towards the infinite, exists and takes the following value:

$$\lim_{x \to +\infty} h(x) = \log(\frac{\alpha}{\beta} \frac{(\beta e^{-\lambda_E + \lambda_G} + 1 - \beta)}{(\alpha e^{-\lambda_E + \lambda_G} + 1 - \alpha)}) \approx \log(\frac{\alpha}{\beta} \frac{1-\beta}{1-\alpha}) \quad (3.4)$$

The approximation of this limit uses the fact that $\lambda_E > \lambda_G$ The idea of Robertson and Walker [72] was to find a function which would have similar properties of the function h. Initially, he proposes to use a function of the type $r(X) = \frac{X}{x+K}$, which is increasing but which tends towards 1. Then, he proposed to multiply this last function by the weights which the model BIR would give, which is similar to the approximated limit of function h.

$$\begin{aligned} h^*(x_w) &= \frac{x_w}{x_w + K} \log(\frac{P(X_w^d = 1|R=1)}{P(X_w^d = 0|R=1)} \frac{P(X_w^d = 0|R=0)}{P(X_w^d = 1|R=0)}) \\ h^*(x_w) &= \frac{x_w}{x_w + K} \log(\frac{a_w}{1-a_w} \frac{1-b_w}{b_w}) \end{aligned} \quad (3.5)$$

Again, a_w and b_w can be estimated repeatedly.

Lastly, Robertson and Walker make some modifications to the original model:

1. It is necessary to take into account the length of the documents in the renormalization of the frequencies . Thus, instead of using a function of the type $\frac{X}{x+K}$, they choose a function of the form

$$\frac{(k_1+1)x_{wd}}{k_1((1-b) + b\frac{l_d}{avgl}) + x_{wd}}$$

where l_d is the length of document d and $avgl$ the mean document length in the collection. k_1 is set by default to 1.2 and b to 0.75.

2. They renormalize the frequency of the words of the request in the following way:

$$\frac{(k_3+1)q_w}{k_3 + q_w}$$

By default, $k_3 = 1000$

Finally, with the initial default values of a_w and b_w, BM25 model can be written as:

$$RSV(q,d) = \sum_{w \in Q} \frac{(k_3+1)q_w}{k_3+q_w} \frac{(k_1+1)x_{wd}}{k_1((1-b)+b\frac{l_d}{m})+x_{wd}} \log(\frac{N-N_w+0.5}{N_w+0.5}) \tag{3.6}$$

The formula of BM25 is rather complex and involves 3 parameters $(k1, k3, b)$ which can be possibly optimized on particular dataset. This model appeared around the years 1995 and known a strong success in surveys like TREC. It is still regarded as a model of reference.

3.2.3 Dirichlet Multinomial and PRP

Xu and Akella [82] proposed recently a retrieval model built upon the PRP with Dirichlet Multinomial distributions . Xu and Akella first argued that the multinomial model is not appropriate under the PRP paradigm and that a better model (the DCM) accounting for burstiness should be used: If the class of relevant document is modeled by a multinomial distribution with parameter θ_R and irrelevant class with parameter θ_N, then, the PRP ranking function gives:

$$RSV(q,d) = \sum_{w \in q} q_w x_{wd} \log(\frac{\theta_{Rw}}{\theta_{Nw}})$$

Then, Xu and Akella [82] explain that this model is inappropriate and that the DCM distribution should be used:

> Consequently, the multinomial distribution is not an appropriate distribution for the probabilistic model. Because the multinomial distribution assumes the independence of the word repetitive occurrences, it results in a score function which incorporates undesired linearity in term frequency. To capture the concave property and penalize document length in the score function, a more appropriate distribution should be able to model the dependency of word repetitive occurrences (burstiness) that is if a word appears once, it is more likely to appear again. The Dirichlet compound multinomial (DCM) distribution [11, 10], which is motivated by the Polya urn scheme, is able to capture word burstiness, and thus better addresses the need to capture score function concavity and document length.

This is the motivation of the DCM distribution within the PRP. We now detail the model. First, the irrelevant class is represented by the whole collection. A DCM model with parameters (β_w) is optimized to fit the collection. Then, the relevant class is modeled with a DCM distribution whose parameters are $(\beta_w + q_w)$, where q_w are the query word frequencies. In other words, query term frequencies increase the parameters of the relevant distribution. The model assumptions can be summarized by:

$$\gamma > 0$$
$$X_d = x|R=0 \sim DCM((\beta_w)_w, l_d) \qquad X_d = x|R=1 \sim DCM((\beta + \gamma q_w)_w, l_d)$$

The resulting IR model is then given by:

$$\begin{aligned} RSV(q,d) &= \sum_{w \in q \cap d} \sum_{j=0}^{x_{wd}-1} \log(1+\gamma \frac{q_w}{\beta_w + j}) \\ &\quad - \sum_{i=0}^{l_d-1} \log(1+\gamma \frac{l_q}{i + \sum_w \beta_w}) \end{aligned} \tag{3.7}$$

Xu and Akella proposed several strategies to estimate β, either with the EDCM distribution or with a leave one out likelihood for the DCM distribution. They also proposed to approximate the relevant class with the set of documents containing all query terms in order to optimize γ. Finally, they extended their model for pseudo relevance feedback. This model has an higher computational cost than other state of the art models, coming from the estimation of DCM distributions and the double sum ($\sum\sum$) in the matching function. Overall, the ad-hoc model performs similarly to language models.

The presentation of the PRP under Dirichlet Multinomial models ends the part on the models developed under the PRP auspices. We now will move to language models for information retrieval.

3.3 Language Models

Language models comes from the speech processing community and were defined as *a probability distribution on a sequence of words*

The core idea of language models in information retrieval is to rank documents by the probability $P(q|d)$- the probability the query could be generated from a document model d. Hence, most relevant documents would be the most likely to generate the query. Analogies with the vectorial space model [73, 5] are straightforward. Instead of representing a document by a vector, a document is represented by a probability distribution; instead of computing euclidean distances, probabilities or KL-divergences are computed. Figure 3.3 illustrates the principle of language models for IR. Thus, for each document one need to associate a language model, namely a probability distribution.

Ponte and Croft [67] proposed the first language models for IR, which had then been extended in many ways [38]. Zhai et al. give a good overview in the language modeling approach in [85]. Most of theses models make the choice of the multinomial distributions to model documents.

One of the fundamental assumptions of the language modeling approach is that for each document there exists a document language model, namely θ^d such that

$$P(X^d = (x_{wd})|\theta_d, l_d) = \frac{l_d!}{\prod_{i=1}^{n} x_{wd}!} \prod_{w \in d} (\theta_{wd})^{x_{wd}}$$

The problem now is to estimate the document language model (θ_{wd}) for each document. To do so, the maximum likelihood estimator (m.l.e) is often employed:

$$\hat{\theta}_{wd}^{mle} = \frac{x_{wd}}{\sum_w x_{wd}} = \frac{x_{wd}}{l_d}$$

Then, the probability the query is generated by a document model d can be computed as follows

$$RSV(q,d) = \log P(q|\hat{\theta_d}, l_q) = \sum_{w \in q} q_w \log(\hat{\theta_{wd}}) + h(q) \tag{3.8}$$

However, the m.l.e raises a major issue: if a query word does not appear in a document, a zero probability is assigned for the document, so that the log probability is not defined. To overcome theses issues, smoothing methods are employed to add some background knowledge during the estimation of the document language model.

3.3.1 Smoothing Methods

The most popular smoothing methods are Dirichlet smoothing and Jelinek-Mercer smoothing . An overview of smoothing methods for language models is presented in [85].

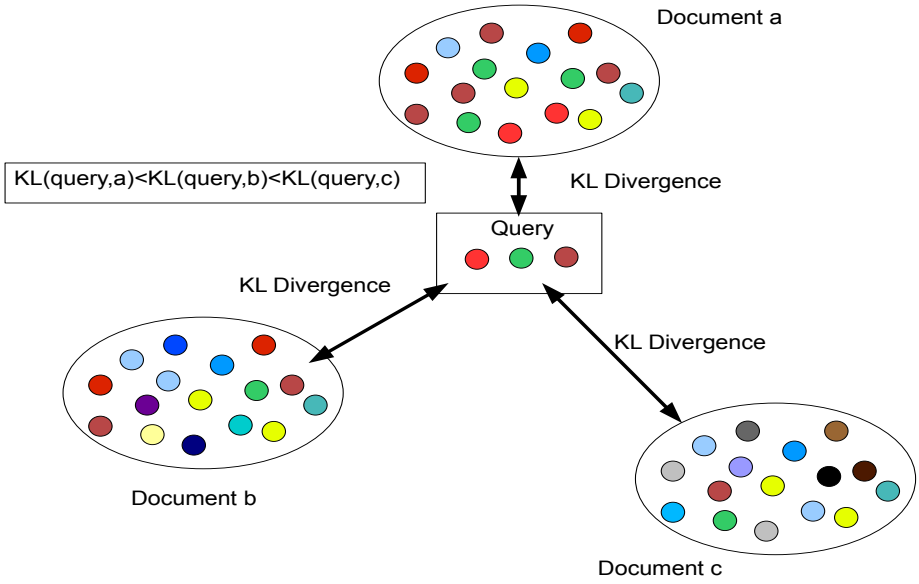


Figure 3.3: The language modeling approach to Information Retrieval. Document are represented by bag of words, ie multinomial distributions. Then, KL-Divergence serves as a distance between queries and documents. In the figure, document a has a lower KL divergence compared to other documents because it contains more query terms.

Jelinek-Mercer Smoothing

The collection of document can also be represented with a language model. Let $\mathcal{C}$ be the collection of documents, then the (multinomial) language model is given by:

$$\beta_w = P(X_w = 1|\mathcal{C}) = \frac{\sum_d x_{wd}}{\sum_d l_d} = \frac{F_w}{L} \tag{3.9}$$

Jelinek-Mercer smoothing proceeds by interpolating the maximum likelihood estimator (m.l.e.) of the document language model with the collection's one.

$$\theta_{wd} = \alpha \hat{\theta_{wd}}^{mle} + (1-\alpha)\beta_w \tag{3.10}$$

which is sometimes noted in the literature as:

$$P(w|d) = \alpha P^{mle}(w|d) + (1-\alpha)p(w|\mathcal{C})$$

Thus, α is a parameter of the model. Generally, α is set after maximizing some performance measure on a given dataset.

The score of a document can be decomposed in two parts: the first part deals with words that both belongs to the query and document. and the second one with words which do not appear in the document. Hence, the latter words would be explained

by the corpus language model. This lead to reformulate the model as:

$$
\begin{aligned}
RSV(q,d) &= \sum_{w\in q} q_w \log(\theta_{wd}) \\
RSV(q,d) &= \sum_{w\in q, x_{dw}>0} q_w \log(\theta_{wd}) + \sum_{w\in q, x_{dw}=0} q_w \log(\theta_{wd}) \\
RSV(q,d) &= \sum_{w\in q, x_{wd}>0} q_w \log(\alpha\theta_{wd}^{mle} + (1-\alpha)\beta_w) + \sum_{w\in q, x_{dw}=0} q_w \log((1-\alpha)\beta_w) \\
RSV(q,d) &= \sum_{w\in q, x_{dw}>0} q_w \log(\frac{\alpha\theta_{wd}^{mle} + (1-\alpha)\beta_w}{(1-\alpha)\beta_w}) + \sum_{w\in q} q_w \log((1-\alpha)\beta_w) \\
RSV(q,d) &= \sum_{w\in q, x_{dw}>0} q_w \log(1 + \frac{\alpha\theta_{wd}^{mle}}{(1-\alpha)\beta_w}) + h(q) \qquad (3.11)
\end{aligned}
$$

This formulation shows that the trade off between the document language model and the corpus is set by the factor $\frac{\alpha}{(1-\alpha)}$. Finally, this formulation is also useful to implement the model in an efficient way with an inverted index.

Dirichlet Smoothing

The next smoothing methods adopts a Bayesian view for the estimation of language models. For each document language model, the following a priori is assumed:

$$\theta_d|\beta,\mu \sim Dirichlet([\mu\beta_w]) \qquad (3.12)$$

where β_w is defined by equation 3.9. So, μ is the parameter which sets the strength of this a priori. Knowing that the Dirichlet and Multinomial are conjugated, the posterior distribution of $\theta_{d.}$ is given by:

$$\theta_d|X^d, l_d, \beta, \mu \sim Dirichlet([\mu\beta_w + x_{wd}]) \qquad (3.13)$$

Finally, the mean value of the posterior distribution is chosen as the language model of the document.

$$\hat{\theta_{wd}} = E(\theta_{d.}|X^d, l_d, \beta, \mu) = \frac{x_{wd} + \mu\beta_w}{l_d + \mu} \qquad (3.14)$$

The bigger μ, the smaller the variance of θ and the less the observed frequency x_{wd} impacts the mean value of θ_{wd}. The decomposition of the score in two terms can also be applied in the case of Dirichlet Smoothing:

$$
\begin{aligned}
RSV(q,d) &= \sum_{w\in q, x_{dw}>0} q_w \log(\frac{x_{wd} + \mu\beta_w}{l_d + \mu}) + \sum_{w\in q, x_{dw}=0} q_w \log(\frac{\mu\beta_w}{l_d + \mu}) \\
RSV(q,d) &= \sum_{w\in q, x_{dw}>0} q_w \log(\frac{x_{wd} + \mu\beta_w}{\mu\beta_w}) + \sum_{w\in q} q_w \log(\frac{\mu\beta_w}{l_d + \mu}) \\
RSV(q,d) &= \sum_{w\in q, x_{dw}>0} q_w \log(1 + \frac{x_{wd}}{\mu\beta_w}) + l_q \log\frac{\mu}{l_d + \mu} + h(q) \qquad (3.15)
\end{aligned}
$$

This formulation shows that μ appears as an a-priori score for a document (in $l_q \log\frac{\mu}{l_d+\mu}$) Furthermore, the ratio $\frac{l_d}{\mu}$ is the analog of $\frac{\alpha}{1-\alpha}$ for Jelinek-Mercer smoothing. As α, μ is optimized according to some performance criteria. Nevertheless, Zhai [85] proposed to estimate μ as the optimal value of the leave one likelihood of the document collection. This method is particularly interesting as relevance judgment are not necessary to carry this estimation.

Smoothing methods penalize common terms compared to rare terms: this was the *IDF* effect in the vectorial model. So, smoothing is a key component of the language modeling approach. Hence, smoothing sets the discrimination power between terms. In the Dirichlet case, it also enables to add a prior score on documents.

3.3.2 KL Retrieval model

The basic language model for IR given by equation 3.8 consists in computing the query likelihood for each document in the collection. This model can be generalized by considering the query as a sample from a random variable [46]. As for each document in the collection, a query is considered as a sample from a multinomial distribution:

$$q|\theta_q, l_q \sim Multinomial(\theta_q, l_q)$$

Queries and documents can be compared with a probabilistic distance, the KL-divergence:

$$RSV(q,d) = -KL(\theta_q, \theta_d)$$

$$\begin{aligned}
RSV(q,d) &= -\sum_w P(w|\theta_q) \log \frac{P(w|\theta_q)}{P(w|\theta_d)} \\
RSV(q,d) &= \sum_w P(w|\theta_q) \log P(w|\theta_d) - \sum_w P(w|\theta_q) \log P(w|\theta_q) \\
RSV(q,d) &= \sum_w \theta_{wq} \log \theta_{wd} - \sum_w \theta_{wq} \log \theta_{wq} \\
RSV(q,d) &= \sum_{w \in q} \theta_{wq} \log \theta_{wd} + h(q)
\end{aligned}$$

where $h(q)$ is the entropy of the query language model. θ_q can be estimated by m.l.e.:

$$\theta_{wq} = \frac{q_w}{\sum_w q_w} = \frac{q_w}{l_q}$$

Then, the KL retrieval model becomes rank equivalent to the query likelihood mode when the document and query model are multinomial distributions:

$$RSV(q,d) =_{\text{rank}} \sum_{w \in q} q_w \log \theta_{wd}$$

So, the KL retrieval model is generalization of the likelihood model. When the query is represented by a language model (a distribution over words), pseudo feedback and query expansion methods become more natural, more valid from a theoretical perspective: the query can then be considered as an incomplete information, which can be updated or enriched with other sources.

3.3.3 Summary

Language models can easily be understood with a vector space model analogy. Documents are represented by probability distributions and and probabilities or divergences are computed between documents and queries instead of distances. Besides, these divergences seems to better fit textual data as language model retrieval model outperform the cosine retrieval function or several tf-idf retrieval functions. Language models are relatively easy to extend and to adapt to several problems in information retrieval.

For example, the Poisson distribution is chosen instead of the multinomial [58]. Several works extended the language modeling approach to cross-lingual IR [47, 63]. Lastly, some works try to take into account the document neighborhood or to use topic models in order to better smooth documents models. To sum up, there exists many extensions of the language modeling approach. These models are the most popular in the field nowadays.

3.4 Divergence From Randomness

Divergence From Randomness (DFR) models [2] reconsider the 2-Poisson underlying idea. Instead of regarding a word as significant or not for a document, these models try to quantify the importance of a word in a document. Harter and Church [37] [11] basically observed that 'good keywords' are far from a Poissonian behavior. The idea of DFR models build on this observation in order to derive weights for words in documents. The cornerstone of these model consists in using Shannon information to measure the importance of a word in a document. and this is why all DFR models relies on a function of a first information: $Inf_1 = -\log P(X_w = x_{wd}|\lambda_w)$ to weigh words in documents. As $P(X_w = x_{wd}|\lambda_w)$ represent the probability of x occurrences of term w in a document d according to parameters λ_w estimated on the collection, the information Inf_1 has the following interpretation:

- If $P(X_w = x_{wd}|\lambda_w)$ is low, then the distribution of w in d deviates from its distribution in the collection, and w is important to describe the content of d. In this case, Inf_1 will be high and word w might be a good descriptor for a document d.

- On the contrary, if $P(X_w = x_{wd}|\lambda_w)$ is high, then w behaves in a document$d$ as expected from the whole collection and, thus, does not provide much information on d (Inf_1 is low).

To sum up, Inf_1 thus captures the importance of a term in a document through its deviation from an average behavior estimated on the whole collection. Figure 3.4 illustrates the principle of Shannon Information to measure informative content.

Measuring word importance with Shannon information in this way is a powerful idea but it contains 2 limitations that are further corrected in DFR models. These corrections are called 'normalization principle':

First Normalization Principle . It aims are renormalizing the first information quantity: Inf_1. The rationale for this normalization is somehow related to the burstiness phenomenon. It is well known fact that many words do not follow Poisson distribution. Although the Poisson distribution can help distinguishing good content words, it can also overestimate the word importance in documents. For instance, if a word occurs many times in a document, then the Poisson distribution will give very high values for the information. Therefore, DFR models proceed with a second probability model $Prob_2$ which renormalizes the previous informative content as follows:

$$(1 - \text{Prob}_2(t_{wd}))\text{Inf}_1(t_{wd})$$

Second Normalization Principle It aims at normalizing the number of occurrences of words in documents by the document length, as a word is more likely to have more occurrences in a long document than in a short one. The different normalizations considered in the literature transform raw number of occurrences. DFR models

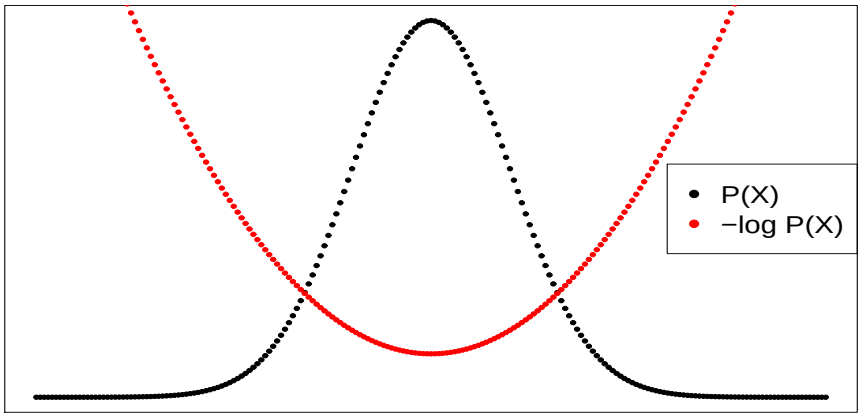


Figure 3.4: Shannon Information Illustration. Probability model and its corresponding information. When an observation is close to the mean/mode of the distribution, then its information is low. On the contrary, an observation with a low probability gives a lot of information

usually adopt one of the two following term frequency normalizations:

$$t_{wd} = cx_{wd}\frac{avgl}{l_d} \tag{3.16}$$

$$t_{wd} = x_{wd}\log(1 + c\frac{avgl}{l_d}) \tag{3.17}$$

where $avgl$ is mean document length and c parametrizes the normalization function.

It is important to stress that DFR models use these normalized term frequencies in discrete distributions in order to compute the Inf_1 and $Prob_2$ quantities.

All in all, the resulting DFR models have the general form:

$$RSV(q,d) = \sum_{w \in q \cap d} q_w(1 - \text{Prob}_2(t_{wd}))\text{Inf}_1(t_{wd})$$

We will discuss the different probability distributions used in DFR models either for measuring the informative content or for modeling the risk of using a word as document descriptor .

3.4.1 Inf_1 Model

Geometric (G)

This model assumes that word frequencies are distributed in the collection according to a geometric distribution: $\lambda_w = \frac{F_w}{N}$ (ie the mean frequency)

$$Inf_1(t_{wd}) = \log_2(1 + \lambda_w) + t_{wd}\log_2(\frac{1 + \lambda_w}{\lambda_w}) \tag{3.18}$$

Poisson (P)

Similarly with a Poisson law: $\lambda_w = \frac{F_w}{N}$ (ie the mean frequency)

$$\begin{aligned} Inf_1(t_{wd}) &= -\log_2(\frac{e^{-\lambda_w}\lambda_w^{t_{wd}}}{t_{wd}!}) \quad (3.19) \\ Inf_1(t_{wd}) &\approx t_{wd}\log_2(\frac{t_{wd}}{\lambda_w}) + \left(\lambda_w + \frac{1}{12t_{wd}} - t_{wd}\right)\log_2(e) + 0.5\log_2(2\pi t_{wd}) \end{aligned}$$

The last approximation use Stirling approximation

3.4.2 $Prob_2$ Model (First Normalization Principle)

Laplace (L)

The Laplace normalization consists in estimating the probability of observing one more occurrence of a term in a document.

$$\begin{aligned} P(X_w = t_{wd}|X_w = t_{wd}-1) &\sim Bernoulli(\frac{t_{wd}}{t_{wd}+1}) \quad (3.20) \\ Prob_2(t_{wd}) &= \frac{t_{wd}}{t_{wd}+1} \quad (3.21) \end{aligned}$$

Binomial Ratio (B)

Let's suppose the number of documents N_w where a word occurs is known. Then, all the occurences of the term are suppose to be uniformly distributed among this set of document. The occurrences follow a Binomial law with parameter $\frac{1}{N_w}$ Then, the probability to have x_{wd} occurrences in a document is given by:

$$P(X_w = x_{wd}|N_w, F_w) = \binom{x_{wd}}{F_w}(\frac{1}{N_w})^{x_{wd}}(\frac{N_w-1}{N_w})^{F_w - x_{wd}} \quad (3.22)$$

Amati then considers the variation of probability when one extra occurrence is added to a document.

$$\frac{P(X_w = x_{wd}|N_w, F_w) - P(X_w = x_{wd}+1|N_w, F_w+1)}{P(X_w = x_{wd}|N_w, F_w)} = 1 - \frac{F_w+1}{N_w(x_{wd}+1)} \quad (3.23)$$

Using the the normalized frequencies t_{wd} instead of x_{wd} leads to the following normalization:

$$Prob_2(t_{wd}) = 1 - \frac{F_w+1}{N_w(t_{wd}+1)} \quad (3.24)$$

3.4.3 Models

DFR models results from the choice of a first probability model $Prob_1$ and renormalization function $Prob_2$. Most DFR models adopts the second term frequency normalization given by equation 3.17. For example, the Geometric-Laplace model with the second normalization (called officially GL2) is written as:

$$\begin{aligned} t_{wd} &= x_{wd}\log(1 + c\frac{avgl}{l_d}) \quad (3.25) \\ \lambda_w &= \frac{F_w}{N} \\ RSV(q,d) &= \sum_{w \in q \cap d} q_w \frac{1}{t_{wd}+1}\big(\log_2(1+\lambda_w) + t_{wd}\log_2(\frac{1+\lambda_w}{\lambda_w})\big) \quad (3.26) \end{aligned}$$

Similarly, the Poisson-Laplace with the second normalization (PL2) is written as:

$$\begin{aligned}
t_{wd} &= x_{wd}\log(1+c\frac{avgl}{l_d}) \\
\lambda_w &= \frac{F_w}{N} \\
RSV(q,d) &= \sum_{w\in q\cap d}\frac{q_w}{t_{wd}+1}\big(\\
&\quad t_{wd}\log_2(\frac{t_{wd}}{\lambda_w}) + (\lambda_w + \frac{1}{12t_{wd}} - t_{wd})\log_2(e) + 0.5\log_2(2\pi t_{wd})\big) \\
RSV(q,d) &\approx \sum_{w\in q\cap d} q_w(\frac{t_{wd}\log_2(\frac{t_{wd}}{\lambda_w})}{t_{wd}+1} + \frac{\log_2(e)(\lambda_w - t_{wd})}{t_{wd}+1} + \frac{0.5\log_2(2\pi t_{wd})}{t_{wd}+1})
\end{aligned}$$

Amati [2] proposed many others models following this principles such as, *PB2, GB2, I(n)B2, I(F)L2*. In practice, these different models get very similar performance, even if $PL2$ et $I(n)L2$ are among the most popular. For all these models, c is the parameter which normalizes frequencies of word in documents. As most IR models, this parameter is set empirically according to performance measures.

3.5 Conclusion

We have reviewed in this chapter the main families of probabilistic IR models. The BM25 model follow the Probability Ranking Principle and assume two poisson mixture models for word frequencies. Language Models are based mostly on multinomial distributions whereas DFR models involves Poisson or Geometric distributions for instances. All these 'basic' retrieval models are often extended toward particular IR need, such as accounting for document structure for instance and we have not mentionned here extensions of these models as we choose to focus on the assumptions of the particular IR models.

A first conclusion of this chapter is that none of the leading IR model (except the DCM model use within the PRP cf section 3.2.3) rely on bursty distributions. Therefore, our goal will be to try to define well performing IR model relying on bursty distributions. The BNB and Log-Logistic distributions, we have introduced in the previous chapter, model term occurrences on the collection whereas language models rather look at what happens at the document level. PRP models need to choose a distribution of occurrence in the relevant class and we have no indication that the burstiness phenomenon still hold in the relevant class. Overall, it is the Divergence From Randomness framework that seems closer to our requirements. This is why we will discuss and analyze DFR models in the context of burstiness in the next chapter.

Albeit probabilistic, these three families of IR models relies on a different framework and the underlying probability laws modeling word occurrences also differs from one family to another in most cases. However, all the resulting weighting functions do have some properties in common as we will see in the next chapter. As well as performance, the easiness to understand and extend an IR model, its assumptions adequacy are other significant features of an IR model.

Chapter 4

Retrieval Heuristic Constraints

Contents

4.1 Introduction

Although the main probabilistic IR models differentiate from each other, either by a underlying theorical framework or by a distinct choice of a word frequency distribution, all these well performing models share common properties which would allow one to describe these models in a *single* framework. This is precisely the aim of retrieval heuristic constraints , which were pioneered by Fang et al [33].

Heuristic constraints aim at describing *formally* some constraints that all IR models share. Describing IR functions with constraints is referred to as the *axiomatic approach* to IR. One can view these constraints as 'necessary' conditions or as general properties that can help us to understand, from a theoretical point of view, the behavior of IR models. As an introduction to retrieval constraints, we give some examples of the properties captured by the main retrieval constraints:

1. It is important that documents with more occurrences of query terms get higher scores than documents with less occurrences (Term Frequency effect).

2. However, the increase in the retrieval score should be smaller for larger term frequencies, inasmuch as the difference between say 110 and 111 is not as important as the one between 1 and 2 since the number of occurrences has doubled in the second case, whereas the increase is relatively marginal in the first case (Concave effect).

3. In addition, longer documents, when compared to shorter ones with exactly the same number of occurrences of query terms, should be penalized as they are likely to cover additional topics than the ones present in the query (Document Length effect).

4. Lastly, it is important, when evaluating the retrieval score of a document, to weigh down terms occurring in many documents, ie which have a high document/collection frequency, as these terms have a lower discrimination power (IDF effect).

This chapter is structured as follows: first, the retrieval heuristic constraints are presented. Then, Divergence From Randomness models are analyzed in order to better assess the effect of their different components.

4.2 Analytical Formulation of Heuristic Constraints

Axiomatic methods were pioneered by Fang et al [33] and followed by many works including [34, 26]. We first present in this section an analytical version of heuristic retrieval constraints which underlie most IR models. We consider here retrieval functions noted RSV which the following form:

$$RSV(q,d) = \sum_{w \in q} a(q_w) h_0(x_{wd}, l_d, z_w, \theta)$$

where q_w is the query term frequency, x_{wd} is the number of occurrence of w in w, l_d the document length, z_w a corpus statistic for word w and θ is a set of parameters.

The function h_0, the form of which depends on the IR model considered, is assumed to be of class[1] C^2 and defined over $(\mathbb{R}^+)^3 \times \Omega$. Ω represents the domain of the parameters in θ. The function a is often the identity function.

In many cases, the above weighting function h_0 can be written as:

$$h_0(x_{wd}, l_d, z_w, \theta) = h(tz_w, \theta) \text{ where } t(w,d) = t(x_{wd}, l_d)$$

where t is the normalized frequency associated to a given normalization function $t(x,l)$.

Language models [86], Okapi [72] and Divergence from Randomness [2] models as well as vector space models [73] all fit within the above form. For example, the Jelinek-Mercer language model can be written as (cf section 3.3.1):

$$\begin{aligned} t(x,l) &= \frac{x}{l} \\ h(t, z = p(w|C), \lambda) &= \log(\lambda t + (1-\lambda) z) \end{aligned}$$

Similarly, the InL2 DFR model can be written as :

$$\begin{aligned} t(x,l) &= x \log(1 + c\frac{avgl}{l}) \\ h(t, z = \frac{N_w}{N}, \lambda) &= -\frac{t}{t+1} \log(z) \end{aligned}$$

We recall here Fang's criteria and provide an analytical version of them which leads to conditions on h which can be easily tested.

The names of the different criteria are directly borrowed from Fang et al. The presentation of the conditions is quite short but dense in mathematical notations. Two conditions deals with the behavior of the function h wrt t, ie the behavior wrt to term frequency.

[1] A function of class C^2 is a function for which second derivatives exist and are continuous.

These conditions are the criterion **TFC1** and **TFC2**. One condition, **TDC**, encodes the IDF effect of vectorial models. In addition, the **LNC1** conditions ensures that longer document get penalized compared to shorter ones. Lastly, the conditions **TFLNC1** and **LNC2** regulates the interaction between the term frequency and the document length.

4.2.1 TF Effect

The first constraint is:
TFC1: Let $q$be a query with only word w, ie $q = w$ and two documents $d1$ and $d2$ such that $l_{d1} = l_{d2}$ (same length).

$$\text{If } x_{wd1} > x_{wd2}, \text{ then } RSV(d1, q) > RSV(d2, q)$$

TFC1 $\iff \forall(l, z, \theta),\ n \in \mathbb{N}^*,\ h_0(n, l, z, \theta)$ is increasing in n. A sufficient condition is:

$$\forall(l, z, \theta), \quad \frac{\partial t(x, l, \theta)}{\partial x} > 0 \text{ and } \frac{\partial h(t, z, \theta)}{\partial t} > 0 \quad (\textbf{TF Effect})$$

This constraint translates the fact that documents with more occurrences of query terms get higher scores than documents with less occurrences and is illustrated in figure 4.1. For example, the function $\log(1 + x)$ captures the increase in term frequency for language models, whereas for DFR models, it is often a function with a pattern as $\frac{x}{x+1}$.

4.2.2 Concave Effect

The next constraint presented by Fang is:
TFC2: Let $q = w$ and 3 documents such that $l_{d1} = l_{d2} = l_{d3}$ and $x_{wd1} > 0$.
If $x_{wd2} - x_{wd1} = 1$ and $x_{wd3} - x_{wd2} = 1$, then

$$RSV(d2, q) - RSV(d1, q) > RSV(d3, q) - RSV(d2, q)$$

TFC2 $\iff \forall(l, z, \theta),\ n \in \mathbb{N}^*,\ h_0(n + 1, l, z, \theta) - h_0(n, l, z, \theta)$ is decreasing. A sufficient condition is:

$$\forall(z, \theta), \frac{\partial^2 h(t, z, \theta)}{\partial t^2} < 0 \quad (\textbf{Concave Effect})$$

This constraint guarantees that the increase in the retrieval score should be smaller for larger term frequencies and is illustrated in figure 4.2.

We propose to illustrate and to discuss further the implications of the concave effect with the following developpement. Let a and b be two words with similar *idf* or collection frequency, ie $z_a = z_b$. Imagine that all documents in the collection have the same length l, let s a constant, representing the number of occurences of word a and b, ie $t_a + t_b = s$. We want to show that concave functions favor a uniform distribution of occurences in documents. Let f the univariate function defined by $f(t) = h(t, z, \theta)$ Now, consider the following optimization problem:

$$\begin{aligned} \text{argmax} \quad & \mathcal{A} = f(t) + f(s - t) \\ \text{subject to} \quad & t \geq 0\ , t \leq s \end{aligned}$$

So, $\mathcal{A}$ gives the score of a document whose frequencies for word a and b are equal to $t_a = t$ and $t_b = s - t$. The solution of this problem gives the preferred repartition of frequencies for both words in documents. The Lagrangian of this problem is then:

$$\Lambda = f(t) + f(s - t) - \lambda t - \delta(s - t)$$

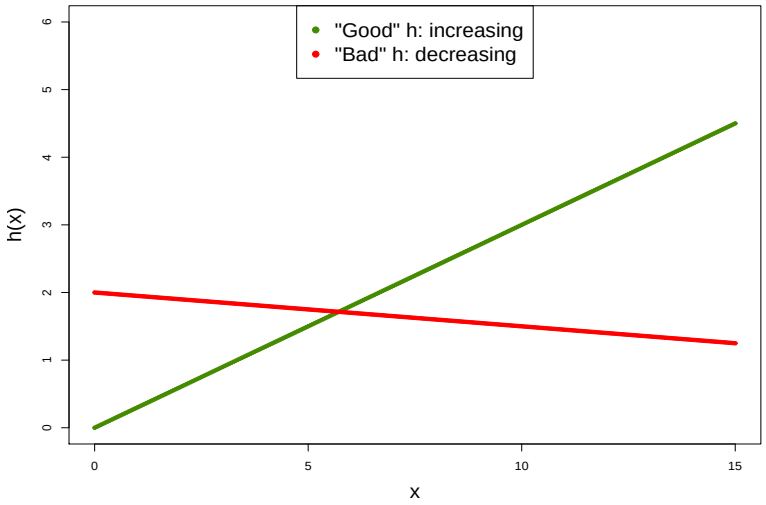


Figure 4.1: Illustration of TF Effect

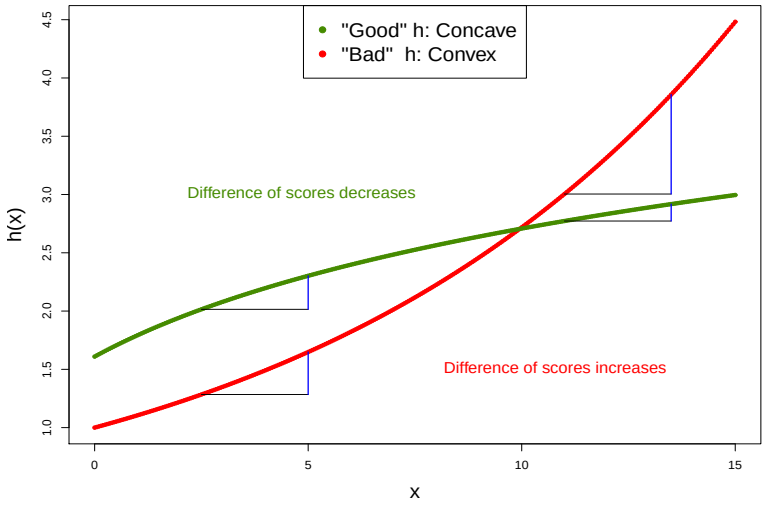


Figure 4.2: Illustration of Concave Effect

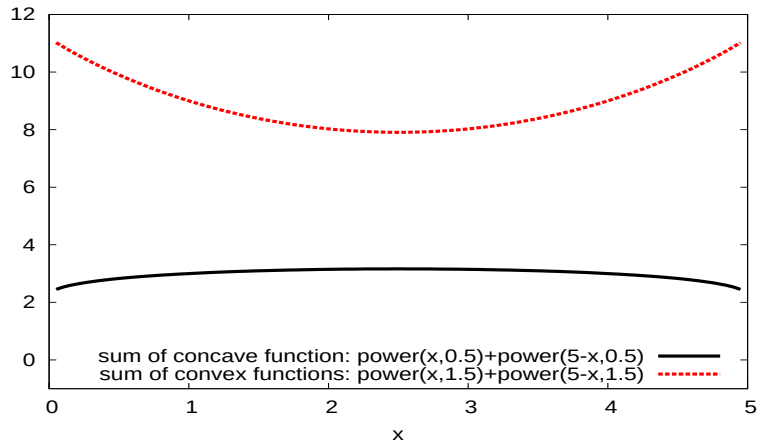


Figure 4.3: Illustration of the equipartition property of concave functions as opposed to convex functions

The Kuhn-Tucker conditions then gives:

$$\begin{aligned} f'(t) - f'(s-t) - \lambda - \delta &= 0 \\ \lambda &\leq 0 \\ \delta &\leq 0 \\ \lambda t + \delta(s-t) &\leq 0 \end{aligned}$$

Either the constraints are active and $t = 0$ or $t = s$, or they are inactive and $\lambda = \delta = 0$, which gives $f'(t) = f'(s-t)$ and $t = \frac{s}{2}$. Overall this gives the two possible solutions:

$$\begin{aligned} t = \frac{s}{2} \qquad & \mathcal{A}(t) = 2f(\frac{s}{2}) \\ t = 0 \text{ or } t = s \qquad & \mathcal{A}(t) = f(0) + f(s) \end{aligned}$$

As f is concave the optimal solution is $t^* = \frac{s}{2}$. Hence, concave functions favors the equipartition of frequencies. In other words, concave functions favor documents with as many occurrences of word a as word b that is to say documents that cover both aspects of a query. On the contrary, if f was convex, it would favor the other solution, when we choose only one word. In other words, convex functions favor documents with either word a alone or word b alone. Note that these arguments are only valid for a fixed document length l and a predetermined s and that they could be generalized with more than two words. Figures 4.3 illustrates the equipartition property of concave functions.

4.2.3 Document Length Effect

The next constraint deals with penalizing longer documents:
LNC1: Let $q = w$ be a query and $d1$, $d2$ two documents.
If, for a word $w' \notin q$, $x_{w'd2} = x_{w'd1} + 1$ but for the query word w, $x_{wd2} = x_{wd1}$, then:

$$RSV(d1, q) \geq RSV(d2, q)$$

$\forall(x, z, \theta),\ n \in \mathbb{N}^*,$ Let $b_n = h_0(x, n, z, \theta)$
LNC1 $\iff \forall(x, z, \theta),\ n \in \mathbb{N}^*,\ h_0(x, n, z, \theta)$ is decreasing. A sufficient condition is:

$$\forall(x,z,\theta), \frac{\partial h_0(x,l,z,\theta)}{\partial l} < 0$$

which translates in the term frequency normalization function as:

$$\forall(x,\theta), \frac{\partial t(x,l,\theta)}{\partial l} < 0 \quad (\textbf{Document Length Effect})$$

This constraint penalizes long documents compared to shorter ones. For example, language models have a term frequency normalization of the form:

$$t(x,l) = \frac{x}{l}$$

due to the constraint on the Multinomial parameter. DFR models rather choose a normalization parametrized by the mean document length $avgl$ and an additional parameter c:

$$t(x,l) = x\log(1 + c\frac{avgl}{l})$$

whereas BM25 rely on the pivoted length normalization [77].

4.2.4 IDF Effect

The next constraint aims at capturing the IDF effect of vectorial models.
TDC: Let q a query and $w1$, $w2$ two query words.
Suppose that $l_{d1} = l_{d2}$, $x_{w1d1} + x_{w2d1} = x_{w1d2} + x_{w2d2}$.

$$\text{If } idf(w1) \geq idf(w2) \text{ and } x_{w1d1} \geq x_{w1d2}, \text{ then } RSV(d1,q) \geq RSV(d2,q).$$

A special case of TDC corresponds to the case where $w1$ occurs only in document $d1$ and $w2$ only in $d2$. In such a case, the constraint can be written as:

speTDC: Let q a query and $w1$, $w2$ two words.
Suppose that $l_{d1} = l_{d2}$, $x_{w1d1} = x_{w2d2}$, $x_{w1d2} = x_{w2d1} = 0$.

$$\text{If } idf(w1) \geq idf(w2), \text{ then } RSV(d1,q) \geq RSV(d2,q).$$

A sufficient condition for *speTDC* is:

$$\forall(t,\theta),\ h(0,z,\theta) = 0 \text{ and } \frac{\partial h(t,z,\theta)}{\partial z} < 0 \Longrightarrow speTDC \quad (\textbf{IDF Effect})$$

This constraint accounts for the IDF effect and is illustrated in figure 4.4. Note that $\frac{\partial h(t,z,\theta)}{\partial z} < 0$ alone is not a sufficient condition. For example, language models with Jelinek-Mercer smoothing are such that $\frac{\partial h}{\partial z} > 0$ even if it does verify the speTDC condition.

More generally, the situation of the *TDC* constraint is unclear, and in fact we show that **several state-of-the-art IR models do not comply** with the general *TDC* constraint, but do satisfy the *speTDC* one. We use a similar development to the one illustrating the equipartition property of concave functions to do so. Let us consider the case where the weighting function is a rank equivalent Jelinek-Mercer smoothing, namely :

$$h(t,z_w,\theta) = \log(1 + \theta\frac{t}{z_w})$$

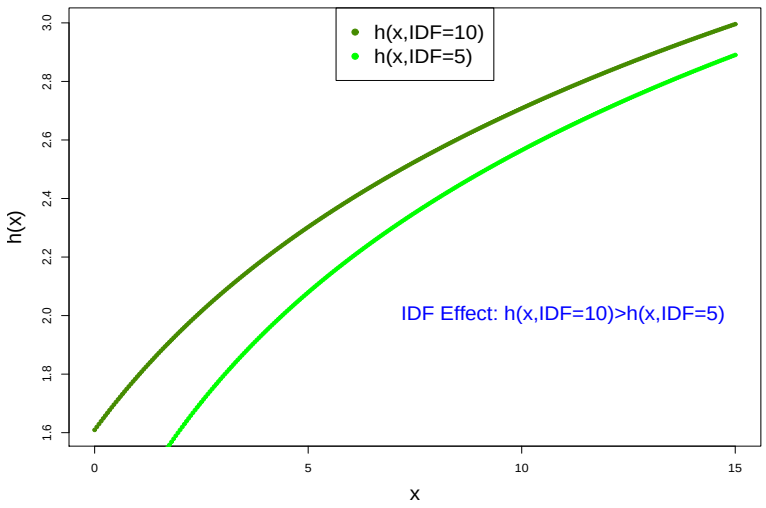


Figure 4.4: Illustration of IDF effect

with $z_w = p(w|C)$. Let a and b two words, let $s > 0$ a constant, representing the number of occurences of words a and b, ie $t_a + t_b = s$ as in the TDC condition. Consider the following optimization problem:

$$\begin{array}{ll} \text{argmax} & \mathcal{A} = \log(1 + \theta \frac{t}{p(a|C)}) + \log(1 + \theta \frac{s-t}{p(b|C)}) \\ \text{subject to} & t \geq 0 \ , t \leq s \end{array}$$

So, $\mathcal{A}$ gives the score of a document whose frequencies for word a and b are equal to $t_a = t$ and $t_b = s - t$. The solution of this problem gives the preferred distribution of frequencies for both words in documents. The Lagrangian of this problem is then:

$$\Lambda = \log(1 + \theta \frac{t}{z_a}) + \log(1 + \theta \frac{s-t}{z_b}) - \lambda t - \delta(s - t)$$

The Kuhn-Tucker conditions then gives:

$$\begin{aligned} \theta(\frac{1}{z_a + \theta t} - \frac{1}{z_b + \theta(s-t)}) - \lambda - \delta &= 0 \\ \lambda &\leq 0 \\ \delta &\leq 0 \\ \lambda x + \delta(s - x) &\leq 0 \end{aligned}$$

which gives the following solution [2]:

$$t^* = \frac{s}{2} + \frac{z_b - z_a}{2\theta}$$

and the corresponding distribution of frequencies for word a and b are:

$$t_a^* = \frac{s}{2} + \frac{z_b - z_a}{2\theta}, \ t_b^* = \frac{s}{2} - \frac{z_b - z_a}{2\theta}$$

[2] As $z_a \ll t$ and $z_b \ll s - t$, $t > 0$ and $t < s$ if s is large enough

Table 4.1: Pair of query terms (short query) below mean corpus language model

Collection	m	μ	$diff < m$
robust	0.0003	500	62.2 %
trec1-2	0.0005	1000	62.2 %

Now let us consider a query q with two words (a and b) occurring only once, and let d_1 and d_2 be two documents of equal length. Let us furthermore assume that: $z_a < z_b$, and:

$$t_{ad_1} = t_a^* + \epsilon, \quad t_{bd_1} = t_b^* - \epsilon$$
$$t_{ad_2} = t_a^*, \qquad t_{bd_2} = t_b^*$$

for ϵ sufficiently small for all the quantities to be positive. In this case, all the conditions of the *TDC* constraint are verified, and thus one should observe that $RSV(q, d_1) \geq RSV(q, d_2)$, which is in contradiction with the fact that the values for d_2 are the ones that maximize $\mathcal{A}$ which corresponds in this case to the retrieval status value. This shows that the Jelinek-Mercer model is not compliant with the *TDC* constraint. However, it is compliant with the *speTDC* constraint, which represents a stricter version of the *TDC* constraint.

In addition, the Dirichlet language model was shown to agree with the *TDC* constraint in [33] when:

$$\mu \geq \frac{x_{ad_1} - x_{bd_2}}{p(b|C) - p(a|C)}$$

Table 4.1 shows for several collections the mean value of $p(w|C)$ for query terms (denoted m), the optimal values obtained for the Dirichlet smoothing parameter μ and the percentage of pairs of query terms for which the corpus language model absolute difference ($|p(w'|C) - p(w|C)|$) is below m (denoted $diff < m$). As one can note, in almost two third of the cases, the numerator of equation 4.1 is very small. So, for the bound given in the above equation to hold, one needs to rely on large values for μ (larger than 2,000 when the numerator is one). As shown in table 4.1, we are far from these values in practice and the Dirichlet language model is in general not compliant with the *TDC* constraint. Furthermore, using the analytical formulation of the *speTDC* constraint, one can show that the Dirichlet language model is compliant with the *speTDC* constraint.

To sum up, several state of the art IR models satisfy *speTDC*, a stricter version of the *TDC* constraint to directly formalize the IDF effect but do not fulfill TDC. Because of the good behavior of the models we have reviewed, we believe that the above development suggests that the *TDC* constraint is too strong, and should be replaced with the *speTDC* one.

4.2.5 Adjustement Conditions

The two following criterion aim at regulating the interaction between the term frequency variable, namely x and the document length l.

LNC2: Let q be a query. $\forall k > 1$, if $d1$ and $d2$ are two documents such that $l_{d1} = k \times l_{d2}$ and for all words w, $x_{wd1} = k \times x_{wd2}$, then $RSV(d1, q) \geq RSV(d2, q)$.

$LNC2 \iff$

$$\forall (z, \theta),\ (x, l) \in \mathbb{N}^*,\ k > 1,\ h_0(kx, kl, z, \theta) \geq h(x, l, z, \theta) \quad \textbf{(condition 5)}$$

TF-LNC: Let $q = w$ be a query with only word w. If $x_{wd1} > x_{wd2}$ and $l_{d1} = l_{d2} + x_{wd1} - x_{wd2}$, then $RSV(d1, q) > RSV(d2, q)$.

According to Fang, the TF-LNC constraint captures the intuition that if the document $d1$ is generated by adding more occurrences of a query term to document $d2$, then the score of $d1$ should be higher than $d2$

$TF\text{-}LNC \iff$

$$\forall(z,\theta),\ (x,l,p) \in \mathbb{N}^*,\ h_0(x+p, l+p, z, \theta) > h_0(x, l, z, \theta) \quad \textbf{(condition 6)}$$

These two constraints basically say that the increase in x always come with an increase in l. As we want to promote the increase in x, the gain coming from the increase in x should be superior to the loss incurred by a longer document. This could be formulated as the variation in x should be bigger than the variation in l.

Moreover, Cummins et al. [26] introduce a new constraint which impacts the document length penalty.

$TFLNC4\ Constraint \iff \forall(q, d, w),\ w \notin q$

$$|RSV(q, d+w) - RSV(q, d)| < |RSV(q, d+2w) - RSV(q, d+w)|$$

Cummins explains that

> The above constraint avoids over penalizing longer documents by ensuring that the normalization aspect (measured in repeated terms) is sublinear. Therefore, as non-query terms appear in a document they should be penalized less with successive occurrences. Essentially, the inverse of the score reduction due to non-query terms being added should be sub-linear.

The concavity condition has a sublinear effect wrt to x whereas this constraint deals with document length. Cummins et al. actually obtained from a Genetic Algorithm a normalization of the form $x \times \sqrt{(\frac{m}{l})}$ and found it effective on several collections. This constraint was suggested by the analysis of the normalization found by the Genetic Algorithm.

4.2.6 Summary

To sum up, the main retrieval constraints are:

- TF Effect: h increases with t
- Concave Effect: h is concave with t
- Document Length Effect: h decreases with l
- IDF Effect : h increases with idf

Lastly, conditions 5 and 6 regulate the interaction between frequency and document length, i.e. between the derivatives wrt to x and l. They allow to adjust the functions h satisfying the above conditions. In the remainder, we will refer to the above conditions as the **form conditions** and the remaining oness as the **adjustment conditions**. We distinguish the two sets of conditions because we believe that form conditions capture a more general behavior of weighting function whereas the adjustment conditions describe a more subtil behavior.

We now need to discuss the axiomatic approach to IR. Although most state of the art algorithms do meet these constraints, it is possible to design IR ranking function that meet all the previous conditions but that will perform poorly, or worst than some other functions that do not meet one of the constraints. So, these axiomatic conditions could be viewed as necessary conditions but not sufficient conditions. Therefore, these conditions do not guarantee a good performance if they are satisfied. Despite these drawbacks, the axiomatic approach to IR provide an unified framework in order to study retrieval models. If the axiomatic theory has some limitations and might be only at its beginnings, we do believe it remains an interesting framework that yield valuable insights when elaborating new IR models.

4.3 Analysis of DFR Models

The Divergence From Randomness (DFR) framework p was introduced in the previous chapter and defines a family of IR models such that:

$$RSV(q,d) = \sum_{w \in q \cap d} q_w (1 - \text{Prob}_2(t_{wd})) \text{Inf}_1(t_{wd})$$

DFR models rely on two normalization principle and we will review them with respect to the retrieval constraints we have defined First, we will drop notation subscripts here because the context does not need such notations

- x refers to x_{wd} (respectively for t and t_{wd})
- l refers l_d
- Here z means either F_w or N_w. It amounts to a corpus frequency.

4.3.1 The Second Normalization Principle

The second normalization principle aims at normalizing the number of occurrences of words in documents by the document length, as a word is more likely to have more occurrences in a long document than in a short one. The different normalizations considered in the literature transform raw number of occurrences DFR models usually adopt one of the two following term frequency normalizations (c is a multiplying factor):

$$\begin{aligned} t = t(x,l) &= xc\frac{avgl}{l} \\ t = t(x,l) &= x \log(1 + c\frac{avgl}{l}) \end{aligned}$$

These normalizations behave as standard normalizations, namely their derivatives satisfies:

$$\frac{\partial t(x,l)}{\partial x} > 0$$
$$\frac{\partial t(x,l)}{\partial l} < 0$$

The important point about the second normalization principle is that, to be fully compliant with these definitions, the probability distribution functions at the basis of DFR models should be continuous distributions as the considered variables are continuous[3]. This is not the case for DFR models proposed so far which rely on discrete distributions.

[3]Furthermore, as these variables are positive, the support of the distributions to be considered should be (or included in) $[0; \infty)$.

4.3.2 The First Normalization Principle

The intuition behind Inf_1 component is simple. Let $P(t|\theta_w)$ represent the probability of t (normalized) occurrences of term w in document d according to parameters θ_w which are estimated or set on the basis of a random distribution of w in the collection. If $P(t|\theta_w)$ is low, then the distribution of w in d deviates from its distribution in the collection, and w is important to describe the content of d. In this case, Inf_1 will be high. On the contrary, if $P(t|\theta_w)$ is high, then w behaves in d as expected from the whole collection and, thus, does not provide much information on d (Inf_1 is low). Inf_1 thus captures the importance of a term in a document through its deviation from an average behavior estimated on the whole collection. The question which arises is *why one should need to normalize it.* In other words, what is the role of the first normalization principle?

Amati and van Rijsbergen [2] consider several basic IR models for $Prob_1$: the binomial model, the Bose-Einstein model, which can be approximated by a geometric distribution, the *tf-idf* model (denoted *I(n)*), the tf-itf model (denoted *I(F)*) and the *tf-expected-idf* model (denoted $I(n_e)$). For the last four models, Inf_1 takes the form:

$$Inf_1(t) = \begin{cases} t\log(1+\frac{N}{z}) + \log(1+\frac{z}{N}) \\ t\log(\frac{N+1}{z+0.5}) \end{cases}$$

where the first line corresponds to the geometric distribution, and the second one to *I(n)*, *I(F)* and $I(n_e)$ (z being respectively equal to n_w, F_w and $n_{w,e}$, the latter representing the expected number of documents containing term w). We assume in the remainder that t is given either by equation 3.16 or 3.17. The conclusions we present below are the same in both cases.

Were we to base a retrieval function on the above formulation of Inf_1 only, then it is straightforward to see that models *I(n)*, *I(F)* and $I(n_e)$ meet the TF, Doc Length and IDF effect and that the model for the geometric distribution verifies the TF condition , but only partly the IDF ones, as the derivative can be positive for some values of z, N and t.

All models however fail the concave effect, in all cases, $\frac{\partial^2 h(x,l,z,\theta)}{\partial x^2} = 0$. Hence, Inf_1 alone, for the geometric distribution and the models *I(n)*, *I(F)* and $I(n_e)$, is not sufficient to define a valid IR model[4]. One can thus wonder whether Inf_2 serves to make the model concave . We are going to see that this is indeed the case.

Two quantities are usually used for $Prob_2$ in DFR models: the normalization L or the normalization B. They both lead to the following form:

$$1 - \text{Prob}_2 = \frac{a}{t+1}$$

where a is independent of t. Thus integrating Inf_2 in the previous models gives:

$$h(t,z,\theta) = \begin{cases} \left(\frac{at}{t+1}\log(1+\frac{N}{z}) + \log(1+\frac{z}{N})\right) \\ \left(\frac{at}{t+1}\log(\frac{N+1}{z+0.5})\right) \end{cases}$$

and :

$$\frac{\partial^2 h(t,z,\theta)}{\partial t^2} = -\frac{b}{(t+1)^3}$$

with $b > 0$, which shows that the models are now compatible with the concave effect. The above development thus explains why the Inf_1 models considered previously need to be resized with an Inf_2 model.

[4]The same applies to the binomial model, for which $\frac{\partial^2 h(x,l,z,\theta)}{\partial x^2} > 0$. For the sake of clarity, we do not present here this derivation which is purely technical.

So, the first normalization principle in DFR models is motivated by the retrieval constraints. This enables to better understand the different components of DFR models. To sum up, $Inf1$ only models comply with all the main retrieval constraint but concavity. With the correction of the $Prob2$ model, the resulting model comply with all the main IR constraints.

4.3.3 Experiments with the First Normalization Principle

After having analyzed DFR models, we want to better demonstrate the effect of the first normalization principle in DFR models. To do so, we tested several variant of DFR models on a CLEF collection. For example, we tested the Geometric $Inf1$ model **with and without** the Laplace normalization.

For example, with only a Geometric distribution, the DFR model would be:

$$RSV(q,d) = \sum_{w \in q} q_w \ \left(\ t_{wd} \log(\frac{1+\lambda_w}{\lambda_w}) + \log(1+\lambda_w) \ \right)$$

where λ_w is the parameter for the Geometric distribution for word w.

As one can see, this model is linear wrt t_{wd} so it is not concave with term frequencies. The experiment will clearly test the concavity impact on the performances, as the model without Laplace normalization is linear whereas the model is concave with Laplace normalization.

In addition, we also test a BNB distribution as $Inf1$ model **with and without** the Laplace normalization. For example, with the Laplace normalization for Inf_2 and the BNB distribution, the model would be:

$$RSV(q,d) = \sum_{w \in q} \frac{q_w}{t_{wd}+1} \ \left(\ \log(\lambda_w + t_{wd}) + \log(\lambda_w + t_{wd} + 1) - \log(\lambda_w) \right)$$

One can see that the resulting model does not increase with the term frequency in the document: it violates the TF effet previously mentioned.

Table 4.2 shows the results of the different models. The results shows that the Geometric "only" model performs poorly, probably because it is linear. This experimental result stresses how important the first normalization principle is for DFR models. Another interesting finding deals with the BNB model. The BNB distribution *alone* reaches state of the art performances. However, the Laplace normalization severely degrades the performances. The BNB Laplace model can be shown not to respect the TF effect, ie to increase with larger term frequencies.

So, the DFR framework is *not appropriate* for bursty distributions as the BNB. Overall, the experiments shows that $Inf1$ and $Prob2$ models are not independent one from another.

4.4 Conclusion

Axiomatic methods hold a central role in IR theory as retrieval constraints enable to better understand, analyze and synthesize IR models. Even if this theory has limitations such as the validation of theses constraints, one could think of these axiomas as *necessary conditions*. We have reviewed these conditions, reformulated them into analytical criterion to ease their use and showed that the general TDC contraint was not satisfied by several IR models.

We then examined DFR models in the light of retrieval constraints in order to better understand the role of the first normalization principle. The analysis revealed that the

		MAP	P10
query-title	Geometric Laplace	0.3610	0.2833
	Geometric	0.1661	0.1433
	BNB Laplace	0.0865	0.0633
	BNB	0.3617	0.2767
query-desc	Geometric Laplace	0.4905	0.3433
	Geometric	0.2479	0.2000
	BNB Laplace	0.1328	0.1183
	BNB	0.4682	0.3327

Table 4.2: Mean average precision (MAP) and precision at 10 documents (P10) for the different models on the CLEF2003 English corpus for short and long queries

first normalization principle ensures the model concavity . $Inf1$ models comply with all the main retrieval constraint except concavity. With the correction of the $Prob2$ model, the resulting model comply with all the main IR constraints. However, the correction of the $Prob2$ model is harmful when using a BNB distribution as shown in the experiments. It suggests that $Inf1$ and $Prob2$ model are dependent from each other.

Even if concavity and the burstiness property are both ways to better take into account large frequencies either in a IR model or for a probability distributions, it seems that there is no direct alignment between the concavity of IR models and the burstiness property of the probability distributions used in IR models. Most state of the art models are concave functions with term frequency but most of the distributions used are not bursty. It could mean that current paradigms for IR models are not fully compatible with the bursty distributions we want to use. In the next chapter, we will introduce information models in order to correct several problems of DFR models and to reveal a connection between the concavity of an IR model and the burstiness property of the probability distribution modelling term frequencies.

Chapter 5

Information Based Model

Contents

5.1 Introduction

The previous chapters reviewed the main probabilistic IR models and their significant features described by axiomatic constraints. This chapter introduces the family of information-based IR models as we want to find a suitable framework for the distributions we have selected in chapter 2, the Beta Negative Binomial and the Log-Logistic distributions.

The most recent and widely used information retrieval models rely on word probability distributions with their own specificities as we saw in chapter 3. In Okapi, for example, it is assumed that word frequencies follow a mixture of two Poisson distributions, in both the relevant and irrelevant sets. The Divergence from Randomness (DFR) framework proposed by Amati and van Rijsbergen [2] makes use of several distributions, among which the geometric distribution, the binomial distribution and Laplace law of succession play the major role. Language models are, for themselves, built upon the multinomial distribution, which amounts to consider binomial distributions for individual words.

Empirical findings on how words behave in text collections however suggest that none of the above distributions is appropriate for accurately describing word frequencies, as shown in chapter 2. This legitimates the question whether one can define a well performing IR model based on bursty distributions. Even if this question was addressed with the use of the Dirichlet Compound Multinomial within the Probability Ranking Principle (cf section 3.2.3), the formal framework we have developed lead us to a different solution.

Although none of the common distributions used in IR models seem appropriate to model burstiness, the very same IR models are concave function with term frequency. As we mentionned in the thesis introduction, burstiness and the IR model concavity in term frequency seem to be two sides of the same coin. Such are the motivations for introducing a new family of IR models. Above all, these models have a remarkable property: a direct relationship between the burstiness property of the probability distributions used and the concavity of the resulting IR model.

This chapter is structured as follows. First, information-based models are introduced before presenting two models within this family: the Log-Logistic and the Smooth Power Law models. Then, experiments validates the good behavior of these models. Finally, several extensions of information models are discussed in section 5.

5.2 Information Models

Information models draw their inspiration from a long-standing hypothesis in IR, namely the fact that the *difference in the behaviors of a word at the document and collection levels brings information on the significance* of the word for the document. This hypothesis has been exploited in the 2-Poisson mixture model, in the notion of eliteness in BM25, and more recently in DFR models. By information, we refer to Shannon information [76] when observing a statistical event. The informativeness of a word in a document has a rich tradition in information retrieval since the influential indexing methods developed by Harter ([37]). The idea that the respective behaviors of words in documents and in the collection bring information on word type is, *de facto*, not a novel idea in IR. It has inspired the 2-Poisson mixture model, the concept of eliteness in BM25 models and is at the heart of DFR models.

Several researchers have observed that the distribution of significant, "specialty" words in a document deviates from the distribution of "functional" words. The more the distribution of a word in a document deviates from its average distribution in the collection, the more likely is this word significant for the document considered. This can be easily captured in terms of information:

$$\text{Info}(x) = -\log P(X = x|\lambda) = \text{Informative Content} \tag{5.1}$$

If a word behaves in the document as expected in the collection, then it has a high probability $P(X = x|\lambda)$ of occurrence in the document, according to the collection distribution, and the information it brings to the document, $-\log P(X = x|\lambda)$, is small. On the contrary, if it has a low probability of occurrence in the document, according to the collection distribution, then the amount of information it conveys is greater. In a nutshell, information could be understood as a *deviation from an average behavior*.

We make use of this notion to define information-based IR models. Indeed, we consider here the family of IR models satisfying the following equation:

$$RSV(q, d) = -\sum_{w \in q} q_w \log P(T_w > t_{wd}|\lambda_w) \tag{5.2}$$

where T_w is a random variable modeling normalized term frequencies and λ_w is a set of parameters of the probability distribution considered. This ranking function corresponds

to a mean information a document brings to a query or, equivalently, to the average of the document information brought by each query term and is similar to the Inf_1 part of DFR models. We will refer to models in this family as information-based IR models. Note that the retrieval function defined by equation 5.2, that words not occurring in a document bring a null information[1].

Few words are needed to explain the choice of the probability $P(T_w \geq t_{wd})$ in the information measure. Shannon information was originally defined on discrete probability and the information quantity from the observation of x was measured with $-\log P(X = x|\Theta)$. As the normalized frequencies t_{wd} are continuous variables, we can not directly apply Shannon information. Differential entropy extends the idea of Shannon entropy to continuous random variables. Basically, differential entropy takes the expectation of $-\log f(x)$ where $f(.)$ is the probability density function. One problem with the differential entropy and $-\log f(x)$ is that it can be either positive or negative as opposed to the discrete case. Moreover, probability density functions are not bounded in general, so comparing two pointwise differential informations from two different distributions might be problematic due to differents scales.

A possible solution is to measure information on a probablility of the form $P(t_{wd} - a \leq T_w \leq t_{wd} + b|\lambda_w)$. However, one has to choose values for a and b and we have chosen $a = 0$ and $b = +\infty$ for the natural handling of zeros and the relation with the burstiness property as we will see later. We have to admit that we chose to measure the information on the survival probability function because it seems convenient and work well in practice and without considering too much theoretical aspects.

A question that can flash through the reader mind: is this definition of information can still be understood as a deviation from an average behavior or a surprise measure ? First, the mean frequency of most words is close to 0. Second, for any word large frequencies are typically less likely than smaller frequencies on average. The larger the term frequency is, the smaller $P(T_w \geq t_{wd})$ is and the bigger $-\log P(T_w \geq t_{wd})$. Hence, the use of the survival function $P(T > t)$ seems compatible with the notion of information we have discussed previously. Figure 5.1 illustrates a probability model given by its survival function $P(T > t)$ and the corresponding information we chose to represent.

Overall, the general idea of the information-based family is the following:

1. Due to different document length, discrete term frequencies (x_{wd}) are renormalized into continuous values $t_{wd} = t(x_{wd}, l_d)$
2. For each term w , we assume that the renormalized values t_{wd} follow a probability distribution P on the corpus. Formally, $T_w \sim P(.|\lambda_w)$.
3. Queries and documents are compared through a measure of surprise, or a mean of information of the form

$$RSV(q, d) = \sum_{w \in q} -q_w \log P(T_w > t_{wd}|\lambda_w)$$

So, information models are specified by *two main components*: a function which normalizes term frequencies across documents, and a probability distribution modeling the normalized term frequencies. Information is the key ingredient of such models since information measures the significance of a word in a document.

5.2.1 Axiomatic Constraints

We have just introduced the family of Information-Based models but we need to check that this family yield valid IR models from a theoretical point of view. To do so, we

[1] with probability distribution whose support is [0,+∞)

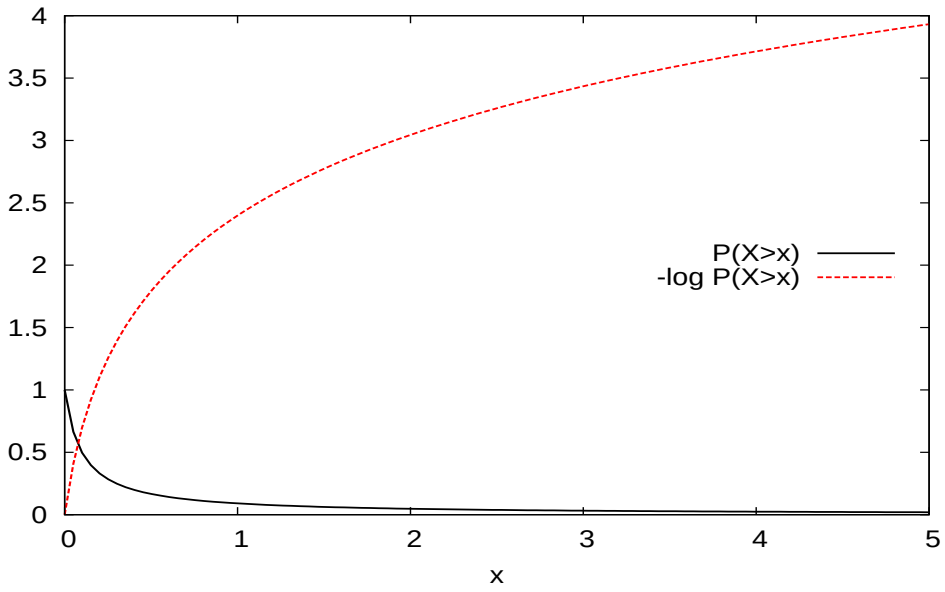


Figure 5.1: Illustration of the Shannon information measure on the Survival function. For small frequencies, then this information is low. On the contrary, for large frequencies, ie a low probability gives a lot of information

analyze information models with respect to retrieval constraints presented in chapter 4. We recall that the four main conditions for an IR model should satisfy deals with *the Term Frequency, the Convavity, the IDF and Document Length.* This analysis will *reveal an important connection between burstiness and the concavity property* of the IR model .

In the case of information models, the weighting function h corresponds actually to the Shannon information computed on the Survival function:

$$RSV(q,d) = \sum_{w \in q} q_w \overbrace{(-\log P(T_w > t_{wd}|\lambda_w))}^{\text{function } h} \tag{5.3}$$

An information models have two degrees of freedom, or two main components which are: *the term frequency normalization function and the probability distribution modeling normalized term frequencies.* Axiomatic constraints actually gives conditions on these two components in order to define a valid IR model.

As far as **term frequency normalization** is concerned, we can show that if we choose a term frequency normalization that is both increasing with term frequency and decreasing in document length, then the resulting IR model satisfies the TF condition and Document Length condition. More formally, $P(T > t_{wd}|\theta_w)$ is a decreasing function of t_{wd}. So, as long as t_{wd} is an increasing function of x_{wd} and a decreasing function of l_d, the TF and Document Length conditions are satisfied for this family of models.

The choice of the **probability distribution** is constrained by the concavity condition. The concavity condition can be expressed for information models as:

$$\frac{\partial^2 h(t,z,\omega)}{\partial t^2} < 0 \quad \Leftrightarrow \quad -\frac{\partial^2 \log P(T > t_{wd})}{\partial (t_{wd})^2} < 0 \tag{5.4}$$

In other words, the IR model is concave if and only if the survival function of the distribution modeling normalized term frequencies is log-convex. This is *exactly* the

characterization of bursty distributions we have given in section 2.4. We recall here this theorem

Theorem 5. *Let P be a probability distribution of class C^2. A necessary and sufficient condition for P to be bursty is:*

$$\frac{\partial^2 \log P(T > t)}{\partial t^2} > 0$$

This means that if the distribution P is **bursty**, then the information model defined with P is guaranteed to be **concave**. So, information models have a direct relationship between the burstiness property of the probability distributions used and the concavity of the resulting IR model. This relationship is not true for state-of -art- IR models, where burstiness and concavity seem to be two sides of the same coin.

More formally, information models can be characterized by the following three elements:

1. **Normalization function** The normalization function t, function of x_{wd} and l_d (respectively the number of occurrences of the word in the document and the length of the document), satisfies:

$$\frac{\partial t}{\partial x_{wd}} > 0; \quad \frac{\partial t}{\partial l_d} < 0$$

2. **Probability distribution** The probability distribution at the basis of the model has to be:
   - Continuous, the random variable under consideration, t_{wd}, being continuous;
   - Compatible with the domain of t_{wd}, i.e. if $t_{\min}$ is the minimum value of t_{wd}, then $P(T_w \geq t_{\min}|\lambda_w) = 1$ (because of the first inequality above, $t_{\min}$ is obtained when $x_{wd} = 0$);
   - Bursty according to our definition
3. **Retrieval function** The retrieval function satisfies equation 5.2, i.e.:

$$\begin{aligned} RSV(q,d) &= -\sum_{w \in q} q_w \log P(T_w > t_{wd}|\lambda_w) \\ &= -\sum_{w \in q \cap d} q_w \log P(T_w > t_{wd}|\lambda_w) \end{aligned}$$

   where the second equality derives from the fact that the probability function verifies $P(T_w \geq t_{\min}|\lambda_w) = 1$, with $t_{\min}$ obtained when $x_{wd} = 0$. The above ranking function corresponds to the mean information a document brings to a query (or, equivalently, to the average of the document information brought by each query term).

Hence, information models satisfy three (out of four) form conditions. The status of the remaining conditions should be checked on each particular model.

5.2.2 Link with DFR

DFR models, as we saw in chapter 3, are specified by:

$$RSV(q,d) = \sum_{w \in q \cap d} q_w \overbrace{(1 - \text{Prob}_2(t_{wd}))}^{Inf_2} \text{Inf}_1(t_{wd})$$

The above form shows that DFR models can be seen as information models, as defined by equation 5.2, with a correction brought by the Inf_2 term. A first important difference between the two models is that DFR models make use of discrete distributions for real-valued variables, a conceptual flaw that information models do not have. Furthermore, if $Inf_2(t_{wd})$ was not used in DFR models, the models with Poisson, Geometric, Binomial distributions would not be concave. In contrast, the use of bursty distributions in information models, together with the conditions on the normalization functions, ensure the concavity of the resulting IR model.

5.3 Two Power Law Instances

We present here two power law distributions which are bursty and lead to information models satisfying all form conditions. The use of power law distributions to model burstiness is not entirely novel, as other studies ([6, 10]) have used similar distributions to model preferential attachment, a notion equivalent to burstiness.

5.3.1 The log-logistic model

The Log-Logistic (LL) distribution is defined by, for $x \geq 0$:

$$P_{LL}(T > t|\lambda, \beta) = \frac{\lambda^\beta}{t^\beta + \lambda^\beta}$$

We consider here a restricted form of the log-logistic distribution where $\beta = 1$.

As explained previously, information models are specified by two main components: a function which normalizes term frequencies across documents, and a probability distribution modeling the normalized term frequencies. A good candidate function for the normalization is the second DFR normalization. Next, the log-logistic distribution is chosen to model the normalized term frequencies. The log-logistic motivation resorts to its relation with the Beta Negative Binomial. However, it parameter λ has to be estimate for each word. The mean document frequency ($\frac{N_w}{N}$) is chosen as parameter value. Such a setting is motivated by the relation to the BNB distribution and the estimation procedure proposed in section2.4.3. So, the log-logistic information model, we will call *LGD*, is defined by:

$$\begin{aligned} t_{wd} &= x_{wd} \log(1 + c\frac{avgl}{l_d}) \\ \lambda_w &= \frac{N_w}{N} \\ RSV(q,d) &= \sum_{w \in q \cap d} q_w (\log(\lambda_w + t_{wd}) - \log(\lambda_w)) \end{aligned} \tag{5.5}$$

where c is the model parameter, controlling the term frequency normalization.

Relation to Language Models

The log-logistic model introduced has fortuitous connections with the Jelinek-Mercer language model (cf section 3.3.1). Let L be the number of tokens in the collection. Following [86], the scoring formula for a language model using Jelinek-Mercer smoothing can be written as:

$$RSV(q,d) = \sum_{w \in q \cap d} q_w \log(1 + s\frac{\frac{x_{wd}}{l_d}}{\frac{F_w}{L}}) \tag{5.6}$$

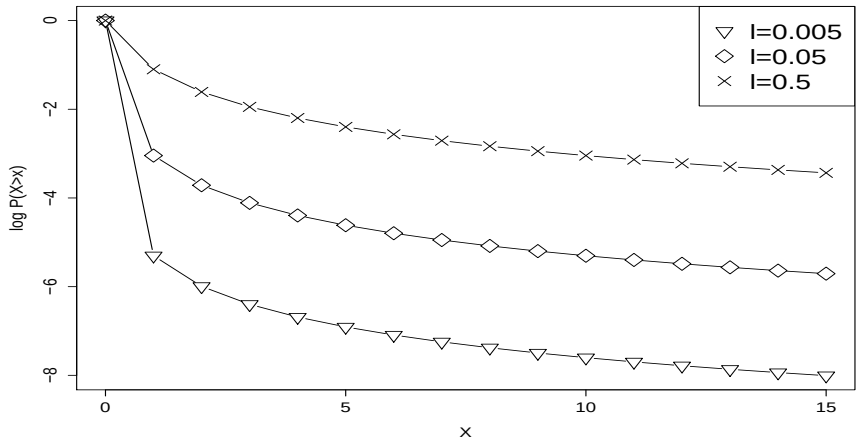


Figure 5.2: $\log P(X > x)$ for $\lambda \in \{0.5, 0.05, 0.005\}$

Using the log-logistic model introduced previously with $\lambda_w = \frac{F_w}{N}$ and the DFR1 length normalization given by:

$$t_{wd} = t(x_{wd}, l_d) = c x_{wd} \frac{avgl}{l_d}$$

We have:

$$RSV(q,d) = \sum_{w \in q \cap d} q_w \log(1 + c \frac{\frac{x_{wd} \times avgl}{l_d}}{\frac{F_w}{N}}) \tag{5.7}$$

Given that $\frac{F_w}{N} = avgl \times \frac{F_w}{L}$, equation 5.6 is equivalent to equation 5.7. The LM model with Jelinek-Mercer smoothing can thus be seen as a log-logistic model with a particular length normalization

This result may seem surprising and contradictory with the research questions that motivated this study. It seems that bursty distributions (a Log-Logistic) can be rank equivalent with non-bursty distributions (Multinomial).

We have previously mentionned in the conclusion of Chapter 4 that there is not a there is not a direct alignment between the concavity of IR models and the burstiness property of the probability distributions used in IR models. This could be one explanation but not the only one.

Thinking of this apparent paradox leads to claim that the language modelling approach in information retrieval do account for the burstiness phenomenon on the contrary of [53],[31], [60]. Although the multinomial distribution is not bursty in the sense of we have defined, we could argue that estimating a parameter for each document amounts to give to a memory to each document. This procedure seems similar to language model estimation in speech processing where a fixed window around the current word enable to adapt the language model. After estimation and smoothing, the probability that a word reoccurs is higher if it was present in the document than if it was absent: Recall that the Jelinek-Mercer smoothing model is given by:

$$P(w|d) = \lambda P(w|\theta_d) + (1 - \lambda) P(w|C)$$

Hence, if a word has appear once in a document (and we have estimated the document

language model), then it is much more likely to appear again since $P(w|\theta_d)$ is typically larger than any $P(w'|C)$.

Furthermore, language models in IR could be consider as non-parametric probabilistic models on documents. The probability of having x occurences of word w from l draws according to the language modelling assumptions in the collection could be written as follows:

$$P(X_w = x|l, \theta_1, ..., \theta_N) = \sum_{d=1}^{N} P(d)P(X_w|l, \theta_d)$$

Hence, the language modeling approach to IR could be considered to use a N mixture binomial model with N very large, which is different from a single Binomial distribution. Recall that the Negative Binomial can be viewed as as infinite mixture of Poisson distributions. So this N mixture binomial can be thought as similar to the Negative Binomial distribution and can account for more variance than a single Multinomial distribuion. For a formal proof of burstiness, we would need to study the ratio $\frac{P(X_w=x+1|l,\theta_1,...,\theta_N)}{P(X_w=x|l,\theta_1,...,\theta_N)}$, which seems non-trivial.
Nevertheless, our experiments and previous studies showed that modeling $P(w|C)$ with a multinomial distribution is a 'bad' modeling assumption. This 'error' may be balanced with the large number of multinomial distributions used in the mixture. These are potential and tentative answers to this apparent paradox.

More generally, in the language modeling approach to IR, one starts from term distributions estimated as the document level, and smoothed by the distribution at the collection level. In contrast, DFR and information-based models uses a distribution the parameters of which are estimated on the whole collection to get a local document weight for each term. Despite the different views sustaining these two approaches, the above development shows that they can be reconciled through appropriate word distributions, in particular the log-logistic one. Lastly, the above connection also indicates that term frequency or length normalizations are related to smoothing.

So, the Jelinek-Mercer model can also be derived from a log-logistic model. However, the Jelinek-Mercer language model and the LGD model differ on the following points:

1. The term frequency normalization:
2. The collection parameter ($p(w|C)$ vs $\frac{N_w}{N}$)
3. The theoretical framework they fit in

It is because we adopted a new theoretical framework, the information-based family, that we could easily use others term frequency normalizations or settings of λ_w. In fact, a language model with the same term frequency normalization as LGD is clearly not straightforward to obtain in the language modeling approach to IR when using multinomial distributions to model documents

5.3.2 Smoothed Power Law (SPL) Model

The second information-based model we are about to introduce draw its inspiration from (again) DFR models. Despite of the poor assumption of word frequency distribution in $Inf1$ model in DFR models, those models, once corrected by an Inf_2 component, lead to state-of the art performance. A legitimate question is then : *are DFR models good models of IR because they are good models of word frequencies ?* In the following, we are going to see that it is possible to approximate, interpret the InL DFR model with a probability distribution. We call Smoothed Power Law, SPL in short, the distribution defined for

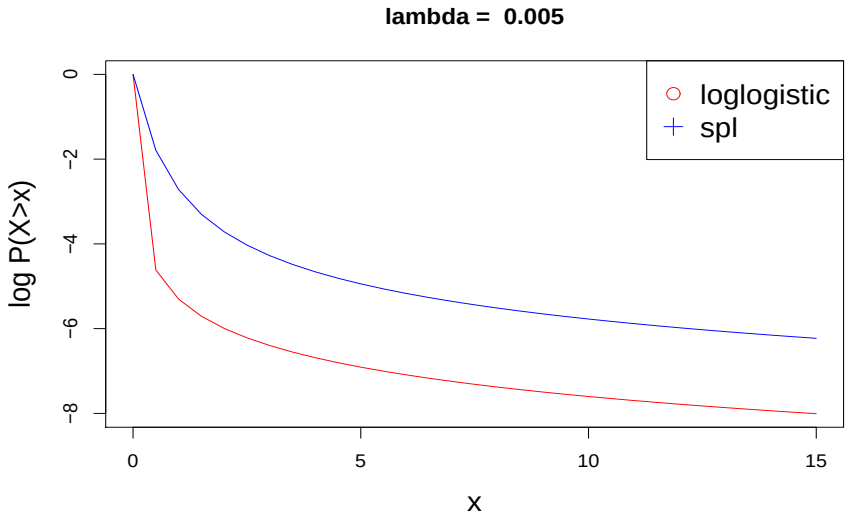


Figure 5.3: Smoothed Power Law distribution vs Log-Logistic distribution with parameter equal to 0.005

$t \geq 0$ and parametrized by λ such that $0 < \lambda < 1$ by:

$$\begin{aligned} f(t;\lambda) &= \frac{-\log\lambda}{1-\lambda}\frac{\lambda^{\frac{t}{t+1}}}{(t+1)^2} \\ P(T > t|\lambda) &= \int_t^\infty f(x;\lambda)dx = \frac{\lambda^{\frac{t}{t+1}} - \lambda}{1-\lambda} \end{aligned} \tag{5.8}$$

where f notes the probability density function.

Figures 5.3 compares a log-logistic distribution with $\lambda = 0.005$ with an SPL distribution with $\lambda = 0.005$. The SPL distribution decrease near 0 is smaller than the log-logistic distribution. Hence, the sharp decrease at 0 has been smoothed. This is what motivates the name of smooth power law. The *key component* of the SPL distribution is the ratio $\frac{t}{t+1}$, which is reminiscent of factors in DFR models and is also close to the ratio $\frac{t}{t+K}$ in Okapi.

As for the log-logistic model, we use the same term frequency normalization to design an IR model. However, the SPL parameter has to be set. We use the connection with the InL2 model to set the parameter to a smoothed mean document frequency. So, the SPL information model takes the form:

$$\begin{aligned} t_{wd} &= x_{wd}\log(1 + c\frac{avgl}{l_d}) \\ \lambda_w &= \frac{N_w}{N + 0.5} \\ RSV(q,d) &= \sum_{w \in q \cap d} -q_w \log(\frac{\lambda_w^{\frac{t_{wd}}{t_{wd}+1}} - \lambda_w}{1-\lambda_w}) \end{aligned} \tag{5.9}$$

Probability Transformation

Note that, the SPL distribution can be generalized with a different step for example.

$$f(t;\lambda,\delta) = \frac{1}{1-\exp(\frac{\log(\lambda)}{\delta})}\frac{-\log(\lambda)}{(t+\delta)^2}\lambda^{\frac{t}{t+\delta}}$$

More generally, the SPL distribution can be seen as a probability transformation. Let Y a random variable, then a transformation of Y can be obtained by:

$$P(T > t|\lambda,\theta) = \frac{\lambda^{P(Y<t|\theta)} - \lambda}{1-\lambda} \tag{5.10}$$

The SPL model, we have derived is actually a change of random variable when one considers a log-logistic distribution. Let $Y \sim LogLogistic(\lambda = 1)$, then $P(Y < t) = \frac{t}{t+1}$, then the probability transformation gives:

$$P(T > t|\lambda) = \frac{\lambda^{\frac{t}{t+1}} - \lambda}{1-\lambda}$$

which is exactly the form of SPL model. This shows that the SPL model can be generalized in many ways.

5.3.3 Comparison with DFR InL model

Figure 5.4 illustrates the behavior of the log-logistic model, the SPL model and the InL2 DFR model (referred to as *INL* for short). To compare these models, we used several values of corpus frequencies (λ_w), ie several IDF values in order to compute term weight obtained for term frequencies varying from 0 to 15. For information models, the weight corresponds to the quantity $-\log P(T_w > t_{wd}))$, whereas in the case of DFR models, this quantity is corrected by the Inf_2 part, leading to, with the underlying distributions retained:

$$\text{weight} = \begin{cases} -\log(\frac{\lambda_w}{t_{wd}+\lambda_w}) & \text{(log-logistic)} \\ -\log(\frac{\lambda_w^{\frac{t_{wd}}{t_{wd}+1}}-\lambda_w}{1-\lambda_w}) & \text{(SPL)} \\ -\frac{t_{wd}}{t_{wd}+1}\log(\frac{N_w+0.5}{N+1}) & \text{(INL)} \end{cases}$$

As one can note, the weight values obtained with the two information models are always above the ones obtained with the DFR model, the log-logistic model having a sharper increase than the other ones for low frequency terms. The plot illustrates that the SPL distribution is very close to the INL models with low values of x, thus confirming that the SPL model can be partly considered as an approximation of the InL model.

5.4 Experimental validation

We now proceed to the evaluation of the log-logistic and smooth power law models in an adhoc scenario. To assess the validity of our models, we used standard IR collections, from two evaluation campaigns: TREC (trec.nist.gov) and CLEF (www.clef-campaign.org). Table 5.1 gives the number of documents (N), number of unique terms, average document length and number of test queries for the collections we retained: ROBUST (TREC), TREC3, CLEF03 AdHoc Task, GIRT (CLEF Domain Specific Task, from the years 2004 to 2006). For the ROBUST and TREC3 collections, we used standard Porter stemming. For the CLEF03 and GIRT collections, we used lemmatization, and an additional decompounding step for the GIRT collection which is written in German.

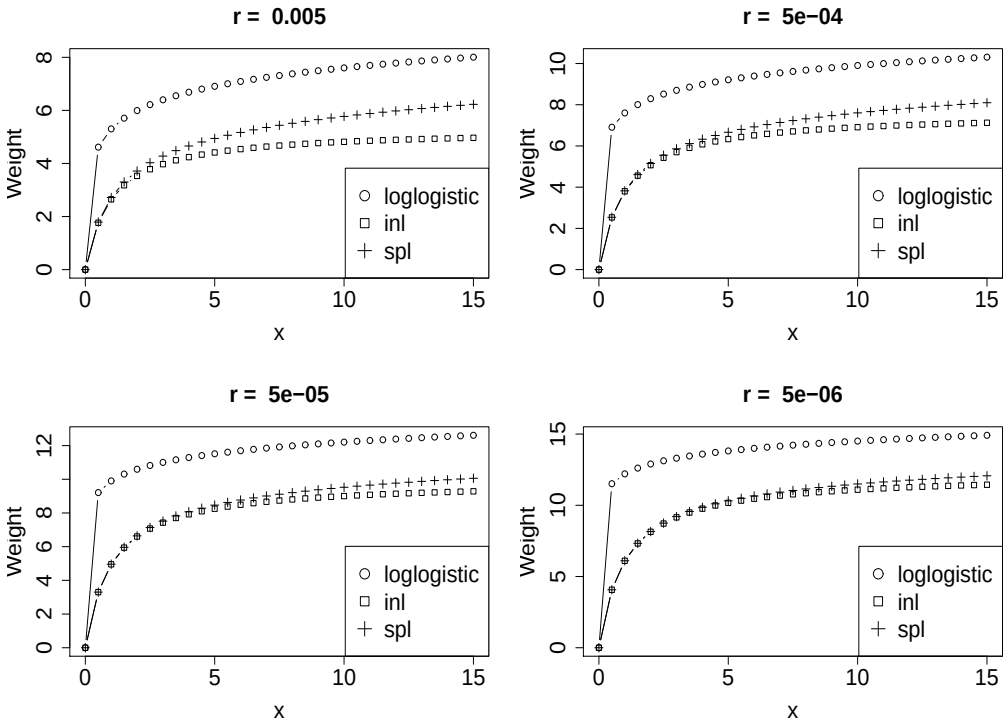


Figure 5.4: Plot of Retrieval Functions with different values of corpus frequencies $r_w = \exp(-IDF(w))$

Table 5.1: Characteristics of the different collections

	#Docs	#Distinct Words	Avg Doc Length	# Queries
ROBUST	490 779	992 462	289	250
TREC-3	741 856	668 648	438	50
CLEF03	166 754	79986	247	60
GIRT	151 319	179 283	109	75

We evaluated the log-logistic and the SPL model against language models, with both Jelinek-Mercer and Dirichlet Prior smoothing, as well as against the standard DFR models and Okapi BM25. The experimental design is the following:

1. For each dataset, we randomly split queries in train and test (half of the queries are used for training, the other half for testing). We then performed 10 such splits on each collection.
2. Learning best parameter (μ, c, k_1) to optimize MAP or P10 on the training set.
3. Measure MAP or P10 on the 10 test sets and test difference with a t-test at 0.05 level.

As the term frequency normalization parameter c is not bounded, we have to define a set of possible values from which to select the best value on the training set. We make use of the typical range proposed in works on DFR models. The set of values we retained is:

$$\{0.5, 0.75, 1, 2, 3, 4, 5, 6, 7, 8, 9\}$$

All our experiments were carried out thanks to the Lemur Toolkit [66]. In all the following tables, *ROB-t* represents the robust collection with query titles only, *ROB-d* the robust collection with query titles and description fields, *CL-t* represent titles for the CLEF collection, *CL-d* queries with title and descriptions and T3-t query titles for TREC-3 collection. The GIRT queries are just made up of a single sentence. Table summarizes 5.2 the different notations for the results tables:

Table 5.2: Notations for the result tables

Collection	query title	query title and descriptions
ROBUST	ROB-t	ROB-d
TREC-3	T3-t	-
CLEF03	CL-t	CL-d
GIRT	GIR	-

5.4.1 Comparison with Jelinek-Mercer and Dirichlet language models

As the smoothing parameter of the Jelinek-Mercer language model is comprised between 0 and 1, we use a regular grid on $[0, 1]$ with a step size of 0.05 in order to select, on the training set, the best value for this parameter. Table 5.3 shows the comparison of our models, LGD and SPL, with the Jelinek-Mercer language model (LM). On all collections, on both short and long queries, the LGD model significantly outperforms the Jelinek-Mercer language model. This is an interesting finding as the complexity of the two models is the same. Furthermore, as the results displayed are averaged over 10 different splits, this shows that the LGD model consistently outperforms the Jelinek-Mercer language model and thus yields a more robust approach to IR. Lastly, the SPL model is better than the Jelinek-Mercer model for most collections for MAP and P10.

In order to assess the relative behaviors of the log-logistic and Jelinek-Mercer models wrt to their parameter (λ for the Jelinek-Mercer model and c for the log-logistic one), we display in Figure 5.5 the MAP scores obtained with different values of these parameters, c being set to $c = \frac{\lambda}{1-\lambda}$, which allows one to compare the two models for any λ in $[0, 1]$. As one can note, with the exception of small values of λ, the log-logistic model dominates the Jelinek-Mercer model, which again shows that the log-logistic model is consistently better than the Jelinek-Mercer one.

Given the link between the Jelinek-Mercer language model and the log-logistic model, such results indicate that the second DFR term frequency normalization is more efficient than the one used in Jelinek-Mercer as the term frequency normalization is the most discriminant factor between the two models.

For the Dirichlet prior language model, we optimized the smoothing parameter from a set of typical values, defined by: {10, 50, 100, 200, 500, 800, 1000, 1500, 2000, 5000, 10000}. Table 5.4 shows the results of the comparison between our models and the Dirichlet prior language model (DIR). These results parallel the ones obtained with the Jelinek-Mercer language model on most collections, even though the difference is less marked. For the ROB collection with short queries, the Dirichlet prior language model outperforms in average the log-logistic model (the difference being significant for the precision at 10 only). On the other collections, with both short and long queries and on both the MAP and the precision at 10, the log-logistic model outperforms in average the Dirichlet prior language model, the difference being significant in most cases. The Dirichlet model has a slight advantage in MAP over the SPL model, but SPL is better for precision. Overall, the information-based models compares favorably to language models.

Table 5.3: LGD and SPL versus LM-Jelinek-Mercer after 10 splits; bold indicates significant difference

MAP	ROB-d	ROB-t	GIR	T3-t	CL-d	CL-t
JM	26.0	20.7	40.7	22.5	49.2	36.5
LGD	**27.2**	**22.5**	**43.1**	**25.9**	**50.0**	**37.5**
P10	ROB-d	ROB-t	GIR	T3-t	CL-d	CL-t
JM	43.8	35.5	67.5	40.7	33.0	26.2
LGD	**46.0**	**38.9**	**69.4**	**52.4**	**33.6**	**26.6**

MAP	ROB-d	ROB-t	GIR	T3-t	CL-d	CL-t
JM	26.6	23.1	39.2	22.3	**47.2**	37.2
SPL	26.7	**25.2**	**41.7**	**26.6**	44.1	37.7
P10	ROB-d	ROB-t	GIR	T3-t	CL-d	CL-t
JM	44.4	39.8	66.0	43.9	34.0	25.6
SPL	**47.6**	**45.3**	**69.8**	**56.0**	34.0	25.6

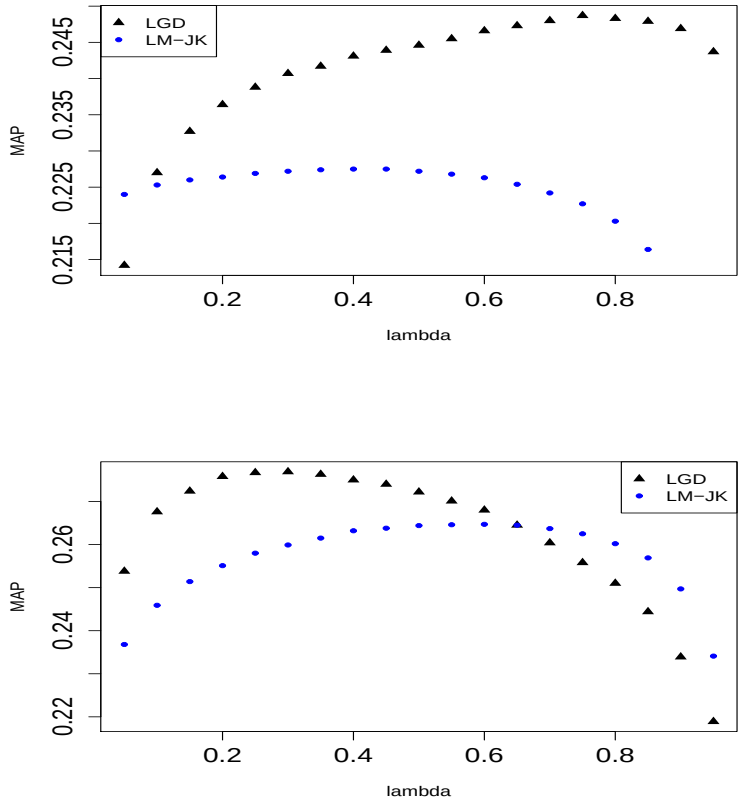


Figure 5.5: MAP against lambda. ROB-t are plot on the top and ROB-d at the bottom

Table 5.4: LGD and SPL versus LM-Dirichlet after 10 splits; bold indicates significant difference

MAP	ROB-d	ROB-t	GIR	T3-t	CL-t	CL-d
DIR	27.1	25.1	41.1	**25.6**	36.2	48.5
LGD	**27.4**	25.0	**42.1**	24.8	**36.8**	**49.7**
P10	ROB-d	ROB-t	GIR	T3-t	CL-t	CLF-d
DIR	45.6	43.3	68.6	54.0	28.4	33.8
LGD	**46.2**	43.5	**69.0**	**54.3**	**28.6**	**34.5**

MAP	ROB-d	ROB-t	GIR	T3-t	CL-t	CL-d
DIR	**26.7**	25.0	40.9	**27.1**	36.2	**50.2**
SPL	25.6	24.9	**42.1**	26.8	36.4	46.9
P10	ROB-d	ROB-t	GIR	T3-t	CL-t	CL-d
DIR	45.2	43.8	68.2	52.8	27.3	32.8
SPL	**46.6**	**44.7**	**70.8**	**55.3**	27.1	32.9

Table 5.5: LGD and SPL versus BM25 after 10 splits; bold indicates best performance significant difference

MAP	ROB-d	ROB-t	GIR	T3-t	CL-t	CL-d
BM25	26.8	22.4	39.8	25.4	34.9	46.8
LGD	**28.2**	**23.5**	**41.4**	**26.1**	34.8	48.0
P10	ROB-d	ROB-t	GIR	T3-t	CL-t	CL-d
BM25	45.9	42.6	62.6	50.6	28.5	33.7
LGD	46.5	**44.3**	**66.6**	**53.8**	28.7	34.4

MAP	ROB-d	ROB-t	GIR	T3-t	CL-t	CL-d
BM25	26.9	24.2	38.5	25.3	35.1	47.3
SPL	27.1	**25.4**	**40.5**	**26.8**	34.5	47.0
P10	ROB-d	ROB-t	GIR	T3-t	CL-t	CL-d
BM25	45.7	41.4	62.8	51.0	**28.5**	36.1
SPL	**47.6**	**44.1**	**67.9**	**57.0**	28.0	35.4

5.4.2 Comparison with BM25

We adopt the same methodology to compare information models with BM25. We choose only to optimize the k_1 parameter of BM25 among the following values: {0.3, 0.5, 0.8, 1.0, 1.2, 1.5, 1.8, 2, 2.2, 2.5}. The others parameters b and k_3 take their default values implemented in Lemur (0.75 and 7). Table 5.5 shows the comparison of the log-logistic and SPL models with Okapi BM25. The log-logistic is either better (4 collections out of 6 for mean average precision, 3 collections out of 6 for P10) or on par with Okapi BM25. The same thing holds for the SPL model, which is 3 times better and 3 times on par for the MAP, and 4 times better, 1 time worse and 1 time on a par for the precision at 10 documents. Overall, information models outperform in average Okapi BM25 with such parameter settings. Note that Okapi model could reach better performance by varying the parameter b.

Table 5.6: LGD and SPL versus INL after 10 splits; bold indicates significant difference

MAP	ROB-d	ROB-t	GIR	T3-t	CL-t	CL-d
INL2	27.7	24.8	42.5	27.3	37.5	47.7
LGD	**28.5**	**25.0**	**43.1**	27.3	37.4	48.0
P10	ROB-d	ROB-t	GIR	T3-t	CL-t	CL-d
INL2	**47.7**	43.3	67.0	52.4	27.3	33.4
LGD	47.0	43.5	**69.4**	53.2	27.2	33.3

MAP	ROB-d	ROB-t	GIR	T3-t	CL-t	CL-d
INL	**26.9**	24.3	40.4	24.8	**35.5**	49.4
SPL	26.6	**24.6**	40.7	**25.4**	34.6	48.1
P10	ROB-d	ROB-t	GIR	T3-t	CL-t	CL-d
INL	47.6	42.8	63.4	52.5	28.8	33.8
SPL	47.8	**44.1**	**68.0**	**53.9**	28.7	33.6

5.4.3 Comparison with DFR models

To compare our model with DFR ones, we chose, in this latter family, the InL2 model, based on the Geometric distribution and Laplace law of succession, and the PL2 model based on the Poisson distribution and Laplace law. These models have been used with success in different works ([2, 14, 80] for example). All the models considered here make use of the same set of possible values for c, namely: $\{0.5, 0.75, 1, 2, 3, 4, 5, 6, 7, 8, 9\}$. It is however interesting to note that both PL2 and InL2 make use of discrete distributions (Geometric and Poisson) over continuous variables (t_{wd}) and are thus theoretically flawed. This is not the case of information models which rely on a continuous distribution.

The results obtained, presented in tables 5.6 and 5.7 are more contrasted than the ones obtained with language models and Okapi BM25. In particular, for the precision at 10, LGD and InL2 perform similarly (LGD being significantly better on GIRT whereas InL2 is significantly better on ROB with long queries, the models being on a par in the other cases). For the MAP, the LGD model outperforms the InL2 model as it is significantly better on ROB (for both sort and long queries) and GIRT, and on a par on CLEF. SPL is better than InL2 for precision but on a par for MAP. Moreover, LGD and PL2 are on a par for MAP, while PL2 is better for P10. Lastly, PL2 is better than SPL for MAP but not for the precision at 10 documents.

Overall, DFR models and information models yield similar results. This is all the more so interesting that information models are simpler than DFR ones: They rely on a single information measure (see equation 5.2) without the re-normalization (Inf_2 part) used in DFR models.

5.4.4 Comparison between LGD and SPL

Table 5.8 compares the log-logistic to the smooth power law model in order to better understand their difference. The log-logistic model tends to achieve better mean average precision whereas the smooth power law model achieve better early precision. Having the same term frequencies normalization, the two model differs mainly by their variation for the t variable ie $h(t)$. However, the SPL achieve better early precision consistently on most collections. We can not explain this difference in behavior yet.

Table 5.7: LGD and SPL versus PL2 after 10 splits; bold indicates significant difference

MAP	ROB-d	ROB-t	GIR	T3-t	CL-t	CL-d
PL2	26.2	24.8	40.6	**24.9**	36.0	47.2
LGD	**27.3**	24.7	40.5	24.0	36.2	47.5
P10	ROB-d	ROB-t	GIR	T3-t	CL-t	CL-d
PL2	46.4	**44.1**	**68.2**	55.0	28.7	33.1
LGD	46.6	43.2	66.7	53.9	28.5	33.7

MAP	ROB-d	ROB-t	GIR	T3-t	CL-t	CL-d
PL2	26.3	25.2	42.8	**25.8**	37.3	**45.7**
SPL	26.3	25.2	42.7	25.3	37.4	44.1
P10	ROB-d	ROB-t	GIR	T3-t	CL-t	CL-d
PL2	46.0	45.2	69.3	54.8	26.2	32.7
SPL	**47.0**	45.2	69.8	55.4	25.9	32.9

Table 5.8: LGD vs SPL after 10 splits; bold indicates significant difference

MAP	ROB-d	ROB-t	GIR	T3-t	CL-t	CL-d
LGD	**28.2**	25.5	41.0	25.1	**32.9**	**48.1**
SPL	27.0	25.6	40.9	**25.8**	32.3	45.4
P10	ROB-d	ROB-t	GIR	T3-t	CL-t	CL-d
LGD	46.5	43.4	67.3	54.6	29.0	**32.8**
SPL	**47.5**	**44.3**	**68.4**	**57.1**	28.7	31.7

5.5 Extensions of Information Models

The applications and extensions of information models presented here aim at answering the following questions:

- Can we use other term frequency normalization other than the second DFR normalization ?
- Can we adjust the concavity/convexity of information models ?

5.5.1 Term Frequency Normalization

This section aims at assessing the dependency of information models to the second DFR term frequency normalization . We want to test if others term frequency normalizations achieve reasonable performances. These performances can be compared to the standard log-logistic model (LGD) in table 5.9.

Pivoted Length Normalization

A naive normalization of term frequencies amounts to dividing by document length. The corresponding pivoted length normalization [77] is then given by

$$t_{wd} = x_{wd} \frac{1}{(1-c) + c\frac{l_d}{avgl}}$$

Table 5.10 shows the MAP for logistic model and smooth power law model with pivoted length normalization. Except on the TREC-3 collection, this normalization does

Table 5.9: Mean Average Precision for Log-Logistic Model with the second DFR normalization , ie the LGD model

c	ROB-d	ROB-t	GIRT	CLEF-d	CLEF-t	T3-t
0.25	27.58	23.64	41.06	49.48	35.87	23.95
0.5	27.64	24.15	42.07	49.36	36.19	24.91
0.8	27.41	24.38	42.2	48.74	36.13	25.39
1	27.22	24.46	42.26	48.44	36.46	25.53
2	26.28	24.76	42.17	46.15	36.84	25.84
3	25.58	24.87	42.0	45.22	36.52	25.82
5	24.71	24.83	41.72	44.35	35.06	25.5
8	23.63	24.73	41.37	42.67	34.57	25.13
10	23.16	24.66	41.2	42.19	34.42	24.91

not improve significantly the results. For the SPL model, the pivoted length normalization achieve a good precision at 10 but does not bring significant improvements. Cummins et al. work [26] actually suggest we could correct the pivoted length normalization with

$$t_{wd} = x_{wd} \frac{1}{(1-c) + c\sqrt{\frac{l_d}{avgl}}}$$

We will call this normalization SQRT-PLN as, square root PLN. Table 5.11 shows the results with a corrected pivoted length normalization. This normalization is only interesting for short queries, as the model performances are almost constant. However, such a normalization severely degrades the performance with long-queries. Another normalization of the pivoted length schema can be obtained by smoothing the variation thanks to the logarithm: We will this normalization LOGPLN. Table 5.12 shows the results for a log-logistic model with LOGPLN. Performances for short queries are relatively stable.

$$t_{wd} = x_{wd} \log(1 + \frac{1}{(1-c) + c\frac{l_d}{avgl}})$$

.

Tables 5.9, 5.10,5.11, 5.12, 5.13 show the MAP for a log-logistic model with different term frequency normalizations. The pivoted length normalization gives similar results to the LGD model for short queries but performs worse for long queries. The SQRTPLN normalization perform poorly compared to the DFR normalization and the LOGPLN behaves as the pivoted len normalization. Overall, these three normalizations do not bring significant advantages over the LGD model.

Another Normalization

We propose to use this term frequency normalization noted TF3:

$$t_{wd} = x_{wd} \frac{c}{c + \frac{l_d}{avgl}} \tag{5.11}$$

We choose c in $\{0.1, 0.25, 0.5, 0.75, 1, 1.5, 3, 5, 7, 9\}$. We change values of c because this normalization may needs values closer to 1. Tables 5.13 and 5.14 showa new the results of this model. This term frequency normalization seems to be beneficial for the log-logistic model. However, it severely degrades the performance with a smooth power law model.

This development showed that others term frequency normalization could be used in information models even though theses new normalizations do not always provide

Table 5.10: Mean Average Precision for Log-Logistic Model with Pivoted Length Normalization

c	ROB-d	ROB-t	GIRT	CLEF-d	CLEF-t	T3-t
0.1	21.3	23.5	41.3	37.5	32.7	23.7
0.2	23.3	24.3	41.7	40.6	33.3	25.2
0.3	24.6	24.7	42.0	42.8	34.6	25.9
0.4	25.3	24.9	42.2	45.5	35.0	26.2
0.5	25.9	24.9	42.3	46.4	36.5	26.3
0.6	26.4	24.8	42.2	46.5	36.7	26.1
0.7	26.8	24.5	42.2	48.0	36.1	25.6
0.8	26.9	24.1	41.9	48.4	36.1	25.0
0.9	26.9	23.7	41.4	48.6	35.7	24.2
Best LGD	27.6	24.9	42.2	49.4	36.8	25.8

Table 5.11: Log-logistic Model with SQRTPLN normalization

c	ROB-d	ROB-t	GIRT	CLEF-d	CLEF-t	T3-t
0.1	19.73	24.21	39.61	39.37	34.67	22.39
0.2	20.68	24.3	40.23	40.11	34.84	22.63
0.3	21.16	24.35	40.53	40.38	34.96	22.75
0.4	21.37	24.38	40.68	40.72	34.97	22.82
0.5	21.43	24.39	40.74	40.77	34.98	22.84
0.6	21.37	24.38	40.68	40.72	34.97	22.82
0.7	21.16	24.35	40.53	40.38	34.96	22.75
0.8	20.68	24.3	40.23	40.11	34.84	22.63
0.9	19.73	24.21	39.61	39.37	34.67	22.39

Table 5.12: Log-logistic Model with LOGPLN normalization

c	ROB-d	ROB-t	GIRT	CLEF-d	CLEF-t	T3-t
0.1	20.9	23.24	41.52	37.15	32.59	22.98
0.2	22.81	24.04	41.96	39.34	32.98	24.79
0.3	24.02	24.5	42.15	41.95	33.5	25.58
0.4	24.96	24.81	42.37	43.09	34.63	26.02
0.5	25.6	24.98	42.5	44.61	34.86	26.38
0.6	26.03	25.02	42.61	46.17	35.33	26.51
0.7	26.42	25.01	42.56	46.7	36.52	26.54
0.8	26.76	24.91	42.59	48.32	36.7	26.33
0.9	27.05	24.71	42.48	47.95	36.85	26.0

spl

Table 5.13: Mean Average Precision for a Log-Logistic Model with TF3 normalization

c	ROB-d	ROB-t	GIRT	CLEF-d	CLEF-t	T3-t
0.1	26.9	23.0	39.2	47.8	35.5	23.0
0.25	**27.9**	24.1	41.5	**49.8**	36.0	25.
0.5	27.8	24.7	42.5	49.4	36.7	26.2
0.75	27.3	**25.0**	42.6	48.8	36.7	26.6
1	26.8	**25.0**	**42.7**	47.2	36.5	**26.7**
1.5	26.0	**25.0**	42.6	45.9	35.2	26.6
3	24.4	24.6	42.2	42.3	33.7	25.8
5	23.0	24.2	41.7	39.7	33.0	25.0
Best LGD	27.6	24.9	42.2	49.4	36.8	25.8

Table 5.14: LGD vs LGD-TF3 after 10 splits: bold indicates statistical significance

MAP	ROB-d	ROB-t	GIRT	CLEF-d	CLEF-t	TREC-3
LGD-TF3	**28.1**	**25.7**	**43.8**	49.4	34.0	**28.9**
LGD	27.9	25.5	43.2	49.9	34.3	28.1

significant improvements. In particular , the TF3 normalization is the most interesting with a Log-Logistic model . Finally, it shows that term frequency normalization is not totally independent from the distribution modeling term frequencies in order to achieve optimal performance.

5.5.2 Q-Logarithm

The purpose of this section is to add a parameter adjusting the concavity/convexity a newof information models . Such a parameter enables to play with the analytical properties of a weighting function. This parameter comes from a generalization of the logarithm function: the q-deformed logarithm [62]. We change the notations here for η-logarithm to avoid confusion with the query notation. The η-logarithm is defined by $\forall t > 0$:

$$\ln_\eta(t) = \frac{1}{1-\eta}(t^{1-\eta} - 1) \tag{5.12}$$

The interesting properties of this curved logarithm are:

- $\ln_\eta(1) = 0$
- $\frac{\partial \ln_\eta(t)}{\partial t} = \frac{1}{t^\eta} = t^{-\eta}$
- $\eta = 1$ leads to the familiar log function.

Figure 5.6 shows the graph of different η-logarithm.

We can define a generalized information model of the form:

$$RSV(q,d) = \sum_{w \in q \cap d} -q_w \ln_\eta P(T_w > t_{wd} | \lambda_w) \tag{5.13}$$

This model, noted QLN, has then two parameters:

- c which normalizes the term frequencies ($t_{wd} = x_{wd} \log(1 + c\frac{avgl}{l_d})$)

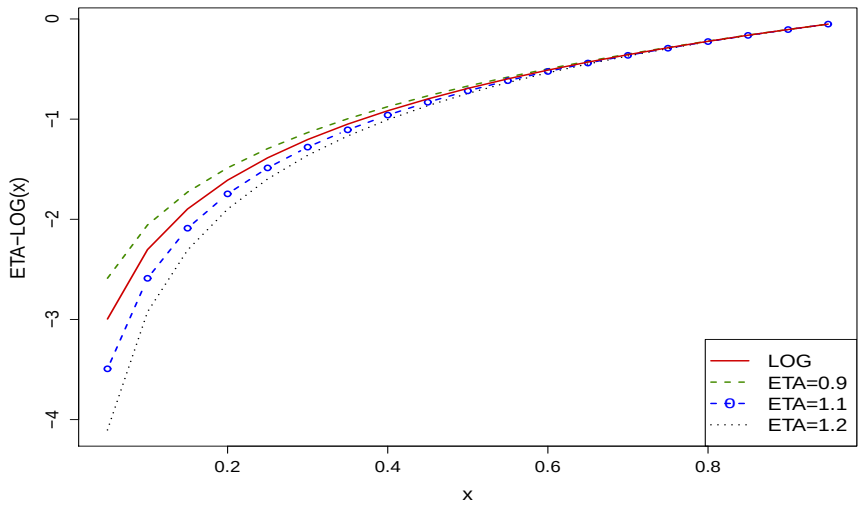


Figure 5.6: $\log_\eta$ for $\eta \in \{0.9, 1, 1.1, 1.2\}$

- η which sets the curvature of the logarithm

The benefit of such model is the ability to play with the curvature of the model in order to assess the role of concavity/convexity. Furthermore, there is no intrinsic reason for the log function to have the best analytical properties in an IR setting. Let's analyze this model with analytical constraints. The weighting function of this model with a log-logistic distribution is given by:

$$\begin{aligned} h(t,z,c,\eta) &= -\frac{1}{1-\eta}((\frac{\lambda}{\lambda+t_{wd}})^{1-\eta}-1) \\ \frac{\partial h}{\partial t} &= -\frac{1-\eta}{1-\eta}(\frac{\lambda}{\lambda+t_{wd}})^{-\eta}(-\frac{1}{(\lambda+t_{wd})^2}) \\ \frac{\partial h}{\partial t} &= \frac{(\lambda+t_{wd})^{\eta-2}}{\lambda^\eta} \\ \frac{\partial^2 h}{\partial t^2} &= (\eta-2)\frac{(\lambda+t_{wd})^{\eta-3}}{\lambda^\eta} \end{aligned}$$

Within this family of IR models, it is possible to get concave and convex models. The concavity condition implies that η must be inferior than 2. The case where $\eta = 2$ is the case where the model is linear:

$$h = \frac{\lambda + t_{wd}}{r} - 1$$

Whenever $\eta < 1$, the weight function h is bounded:

$$h = \frac{1}{1-\eta}(1-(\frac{\lambda}{\lambda+t_{wd}})^{1-\eta}) \leq \frac{1}{1-\eta}$$

A first series of experiments was carried out to assess these new models. It turns out that concave models do not always get better performance than convex models. Typically for values of $\eta < 1$ or below 0.5, the weighting function saturates too quickly. We may

suppose that the IDF effect is penalized by such functions. Experiments also showed that the best values are obtained with $\eta = 1.1$ or $\eta = 1.2$. We compared this model to the log-logistic one in table 5.15. 10 random splits were used in order to optimize (c,η) $\eta \in \{0.8, 0.9, 0.95, 1, 1.1, 1.2, 1.5\}$ for both models. Table 5.16 compares the performance to Okapi trained with two varying parameters (k_1 and b), with the following settings:

QLN $c \in \{0.5, 2, 4, 7, 9\}$ and $\eta \in \{1, 1.1, 1.2\}$

BM25 $k_1 \in \{0.8, 1.0, 1.2, 1.5\}$ and $b \in \{0.25, 0.5, 0.75, 0.85\}$

Table 5.15: Q-Log versus Log-Logistic after 10 splits: bold indicates statistical significance

MAP	ROB-d	ROB-t	GIRT	CLEF-d	CLEF-t	TREC-3
QLN	**28.8**	**25.0**	**41.7**	49.6	34.9	**27.0**
LGD	28.2	24.7	41.0	**50.3**	**35.6**	26.2
P10	ROB-d	ROB-t	GIRT	CLEF-d	CLEF-t	TREC-3
QLN	46.2	43.5	**67.9**	35.6	27.1	53.9
LGD	45.8	43.2	66.2	36.0	**27.6**	52.4

Table 5.16: Q-Log versus BM25 after 10 splits: bold indicates statistical significance

MAP	ROB-d	ROB-t	GIRT	CLEF-d	CLEF-t	TREC-3
BM25	26.9	23.8	40.8	51.6	33.4	**27.5**
QLN	**28.3**	**24.7**	**42.5**	50.8	33.0	26.9

So, changing the curvature of the models allows one to obtain significant improvements over the log-logistic model and BM25. It shows that analytical properties of a model are very significant features.

5.6 Conclusion

We have presented the family of information models for adhoc IR. These models draw their inspiration from a long standing idea in information retrieval, namely the one that a word in a document may not behave statistically as expected on the collection. Shannon information can be used to capture whenever a word deviates from its average behavior, and we showed how to design IR models based on this information.

Information models are a simplification of DFR models in the light of retrieval constraints and burstiness The choice of the distribution to be used in such models was crucial for obtaining valid retrieval models: we showed how burstiness relates to heuristic retrieval constraints, and how it can be captured through power-law distributions.

We have defined two effective IR model within this family: the Log-Logistic IR model and the Smoothed Power Law IR model. The Log-Logistic model can be related to the Jelinek-Mercer language model whereas the Smoothed Power law can be seen as an approximation of the INL2 DFR model. These information-based models satisfy the main retrieval conditions as they rely on bursty distributions and use a valid term frequency normalization. The experiments we have conducted on different collections illustrate the good behavior of these models. Overall, our models yield similar performances to state-of the art models. Moreover, we have discussed the impact of term frequency normalization on these models and found that the TF3 normalization was the best for the log-logistic

model. We also show that the use of $q-$logarithm to measure information was beneficial to IR tasks.

One could ponder on the benefit of modeling burstiness if no extra-gain is obtained, even if the strength of information models is to yield at the same time a valid model of word frequencies and a valid IR model. These comparable performances could be explained by the fact that information-based model and state-of-the-art models describe very similar weighting functions albeit their different underlying hypothesis. This is actually what the axiomatic theory tells about retrieval models: they all satisfy the Term Frequency, Concavity, IDF and Document Length conditions and this is why many retrieval models can be cast in a single theoretical framework.

The axiomatic theory provides valuable insight into IR models but it is not sufficient enough to understand the subtility and precise behavior of certain IR models. For example, the experiments comparing the log-logistic to the smooth power law and the extension of information model with the η logarithm suggest that the performance of a weighting function is related to its shape and analytical properties. Except the concavity constraint, there is no constraint or analytical properties yet, able to explain these differences of performance. It is tempting to say that *analytical properties of a function influence the capacity of a retrieval model*. But the *capacity* of a retrieval model is not formally defined on the contrary to classifiers in machine learning. Similarly, analytical properties are a vague concept. Is it the behavior of the second or third derivatives wrt to t which impacts the early precision ? These are still open questions that remain to be investigated for a better understanding of retrieval constraints and weighting functions.

Nevertheless, modeling burstiness will turn out to be beneficial with pseudo relevance feedback. The information models significantly outperform all the other models with pseudo relevance feedback. We will discuss pseudo relevance feedback and explain this good-behavior in the next chapter.

Chapter 6

An Axiomatic Analysis of Pseudo-Relevance Feedback

Contents

6.1 Introduction

Pseudo Relevance Feedback (PRF) aims at automatically expanding the initial query with terms found in the top retrieved documents and we first show how to extend information-based models to suggest new query terms.

As we will see later in this chapter, our preliminary analysis shows that the extension of information-model to PRF outperfom other models and the initial motivation of this chapter was to better understand the good performance of information models in pseudo relevance feedback. However, we would like to better understand the reasons of these improvements and this is why we conducted an empirical analysis of PRF models.

We then link our empirical observations to the properties of PRF models so as to understand from a theoretical standpoint the performance of PRF models. In a nutshell, we extend the Axiomatic Theory for PRF. In particular, we formulate heuristic constraints for PRF similar to the *TF, Concave, Document Length and IDF effects* reviewed in section

4 with an additional constraints refer as *DF effect*, which is experimentally validated. The theoretical study we conduct reveals that several standard PRF models either fail to enforce the IDF effect or the DF effect whereas the log-logistic and the smoothed power law models satisfy all the PRF properties. Our theoretical analysis thus provide an explanation on why the information-based models perform better than other models in PRF settings.

The remainder of the chapter is organized as follows. First, we give a short introduction to PRF in section 6.2. We show in Section 6.3 how to extend information-based model for PRF and give in Section 6.4 some basic statistics on three PRF models, which reveal global trends of PRF models. We then introduce in section 6.5 constraints PRF models should satisfy, prior to reviewing standard PRF models according to their behavior wrt these constraints in section 6.6.

6.2 Pseudo Relevance Feedback

The goal of PRF models is to expand queries to improve performances so as to improve the user formulation of its information need. PRF methods can be seen as a semantic enrichment process and usually consists in 4 steps:

1. Retain the top retrieved n document after an initial search.
2. Select the 'best' tc word from this set of documents
3. Weight the words and add them to the new query
4. Do a search with the new query

The notations used for PRF are given in table 6.1. We note n the number of pseudo relevant document used, F the feedback set and tc the number of term for pseudo relevance feedback. An important change of notations concerns TF and DF which are in this chapter *related to the pseudo feedback set* F.

Notation	Description
n	# of docs retained for PRF
$\mathbf{F}$	Set of documents retained for PRF: $\mathbf{F} = (d_1, \ldots, d_n)$
tc	*TermCount*: # of terms in $\mathbf{F}$ added to query
$TF(w)$	$=\sum_{d \in F} x_{wd}$
$DF(w)$	$=\sum_{d \in F} I(x_{wd} > 0)$

Table 6.1: PRF Notations

We want to give here an example of PRF model before proceeding to further considerations. We briefly describe one popular PRF model within the language modeling approach to IR. This model is known as *Simple Mixture Model*, or mixture model in short.

Following the language model principle, Zhai and Lafferty [84] proposed a generative model for the set $\mathbf{F}$. All documents in the feedback set are supposed to be i.i.d and each document is generated from a mixture of a feedback model and the corpus language model:

$$\begin{aligned} F &\sim Multinomial((1-\lambda)\theta_F + \lambda P(w|C)) \\ P(\mathbf{F}|\theta_F, \beta, \lambda) &= \prod_{w=1}^{V} ((1-\lambda)P(w|\theta_F) + \lambda P(w|C))^{TF(w)} \end{aligned} \tag{6.1}$$

where $\theta_{Fw} = P(w\theta_F)$, λ is a "background" parameter set to some constant and $TF(w)$ is given in table 6.1 and corresponds to the total number of occurences of word w in the set F . The idea underlying this model is that interesting words for the query would be captured by the multinomial parameter $P(w|\theta_F)$ which needs to be learned. θ_F is learned by optimizing the data log-likelihood with an Expectation-Maximization (EM) algorithm, leading to the following E and M steps at iteration (i):

$$\begin{aligned} E-step \quad & E(w)^{(i)} = \frac{(1-\lambda)P^{(i}(w|\theta_F)}{(1-\lambda)P^{(i)}(w|\theta_F)+\lambda P^{(i)}(w|C)} \\ M-step \quad & P^{(i+1)}(w|\theta_F) = \frac{\sum_{d\in\mathbf{F}} x_{wd})E(w)^{(i)}}{\sum_w \sum_{d\in\mathbf{F}} x_{wd}E(w)^{(i)}} \end{aligned}$$

Once θ_F has been estimated, the best tc words are retain and the new query is obtained by interpolating the (original) query language model with the feedback query model θ_F:

$$\theta_{q'} \quad = \quad \alpha\theta_q + (1-\alpha)\theta_F \tag{6.2}$$

Note that this interpolation is controlled by the parameter α. PRF models have at least 3 parameters: n the number of top-retrieved document, tc the number of expansion terms and a parameter to control the interpolation between the first and second query.

We now show an example of words chosen by this mixture model. We consider the query 303 of the robust collection defined as 'Hubble Telescope Achievements'. The preprocessed query is actually 'achiev telescop hubbl' after stemming. If we choose to do PRF with 10 documents and expand the query with 10 words, the mixture model find that the most 'relevant' words are:

test space nasa scientist mirror flaw optic shuttl telescop hubbl

We will review later several PRF models in section 6.6 when we will examine them with axiomatic constraints since the goal of the above development is to briefly introduce important concepts in PRF.

6.3 PRF with Information Models

After having introduced the main ideas of PRF, we want to show to perform PRF with information models The key idea of information models is to measure the significance of a word thanks to its informative content. There is natural and simple extension of this principle to PRF where we measure the importance of a term in a set of documents with the mean information of a word in this set. The average information this set brings on a given term w can directly be computed as:

$$\text{Info}_{\mathbf{F}}(w) = \frac{1}{n}\sum_{d\in\mathbf{F}} -\log P(T_w > t_{wd}|\lambda_w) \tag{6.3}$$

where the mean is taken over all the documents in $\mathbf{F}$. The original query is then modified, following standard approaches to PRF, to take into account the words appearing in the initial query as:

$$q'_w = \frac{q_w}{\max_w\ q_w} + \beta\frac{\text{Info}_{\mathbf{F}}(w)}{\max_w\ \text{Info}_{\mathbf{F}}(w)} \tag{6.4}$$

where β is a parameter controlling the modification brought by $\mathbf{F}$ to the original query and q'_w denotes the updated weight of w in the query

6.3.1 Evaluation

There are many parameters for pseudo-relevance feedback algorithms: the number of document to consider n, the number of terms to add the query tc and the weight to give to those new query terms (parameter β in equation 6.4). Optimizing all these parameters and smoothing ones at the same time would be very costly. Many studies choose a fixed parameter strategy either to compare PRF models or when submitting runs to evaluation campaigns.

For each collection[1], we choose the optimal smoothing parameters for each model (c,μ,k_1) on all queries. The results obtained in this case are given in table 6.2, where LM+MIX corresponds here to the Dirichlet language model. They show, for example, that on the ROBUST collection there is no difference between the baseline systems we will use for pseudo-relevance feedback in terms of MAP. Overall, the precision at 10 is very similar for the different systems, so that there is no bias, with the setting chosen, towards a particular system.

We compare here the results obtained with the information models to two state-of-the-art pseudo-relevance feedback models: Bo2, associated with DFR models ([1]) (cf section 6.6.3) , and the mixture model associated with language models ([84]) (described aboved and in section 6.6.1). The experimental schema is the following:

1. Divide each collection in 10 splits training/test
2. Learn best interpolation weight (β, α) to optimize MAP on training set
3. Measure MAP on the 10 splits and test difference with a t-test
4. Change $|\mathbf{F}| = n$ and termCount tc to add to the queries
5. Go back to 2

For each collection, we average the results obtained over 10 random splits, the variation of $|F|$ and tc being made on each split so as to be able to compare the results of the different settings.

For each setting, we optimize the weight to give to new terms: β (within {0.1, 0.25, 0.5, 0.75, 1, 1.5, 2}) in information and Bo2 models, α (within $\{0.1, 0.2, \ldots, 0.9\}$) in the mixture-model for feedback in language models. In this latter case, we set the feedback mixture noise to its default value (0.5). As before, we used Lemur to carry our experiments and optimize here only the mean average precision.

Table 6.3 displays the results for the different models (as before, a two-sided t-test at the 0.05 level is used to assess whether the difference is statistically significant, which is indicated by a $*$). As one can note, the information models significantly outperform the pseudo-relevance feedback versions of both language models and DFR models. The SPL model is the best one for $n = 5$ and $tc = 5$, while the LGD model yields the best performance in most other cases. Although DFR and information models perform similarly when no feedback is used, their pseudo-relevance feedback versions do present differences, information models outperforming significantly both language and DFR models in this latter case.

We also performed a 5-fold cross-validation to learn all the feedback parameters at the same time : the goal of these experiments is to answer the critics of a fixed parameter strategy to compare PRF models, where one could argue that PRF models need different parameter setting. We choose the parameters that provide the best performance from

[1] the same collection used to validate information models in the previous chapter

Table 6.2: Performances of baseline setting for PRF ($n = 0$, $tc = 0$): bold indicates significant difference

MAP	ROB-t	GIRT	T3-t	CLEF-t
LM+MIX	25.4	41.1	**28.3**	37.0
LGD	25.4	**42.4**	27.1	**37.5**
P10	ROB-t	GIRT	T3-t	CLEF-t
LM+MIX	44.6	68.3	**56.3**	**27.5**
LGD	44.1	68.7	55.3	27.2

Table 6.3: Mean average precision of PRF experiments; bold indicates best performance, * significant difference over LM and Bo2 models

Model	n	tc	ROB-t	GIR	T3-t	CL-t
LM+MIX	5	5	27.5	44.4	30.7	36.6
INL+Bo2	5	5	26.5	42.0	30.6	37.6
LGD	5	5	28.3*	44.3	**32.9***	37.6
SPL	5	5	**28.9***	**45.6***	**32.9***	**39.0***
LM+MIX	5	10	28.3	45.7*	33.6	37.4
INL+Bo2	5	10	27.5	42.7	32.6	37.5
LGD	5	10	29.4*	44.9	**35.0***	**40.2***
SPL	5	10	**29.6***	**47.0***	34.6*	39.5*
LM+MIX	10	10	28.4	45.5	31.8	37.6
INL+Bo2	10	10	27.2	43.0	32.3	37.4
LGD	10	10	**30.0***	46.8*	**35.5***	38.9
SPL	10	10	**30.0***	**48.9***	33.8*	**39.1***
LM+MIX	10	20	29.0	46.2	33.7	38.2
INL+Bo2	10	20	27.7	43.5	33.8	37.7
LGD	10	20	**30.3***	47.6*	**37.4***	38.6
SPL	10	20	29.9*	**50.2***	34.3	**39.7***
LM+MIX	20	20	28.6	47.9	32.9	37.8
INL+Bo2	20	20	27.4	44.3	33.5	36.8
LGD	20	20	**29.5***	48.9*	**37.2***	**41.0***
SPL	20	20	28.8	**50.3***	33.9	39.0*

this set of ranges:

$$\begin{aligned} n &\in \{10, 20\} \\ tc &\in \{10, 20, 50, 75, 100\} \\ \alpha &\in \{0.1, ..., 0.9\} \\ \lambda &\in \{0.1, ..., 0.9\} \\ \beta &\in \{0.01, 0.1, 0.25, 0.5, 0.8, 1, 1.2\} \end{aligned}$$

where α and β are interpolation parameter with the new query (cf section 6.6) and λ is an additional noise parameter for the mixture model. It turns out that the log-logistic model outperforms the mixture model with an average MAP of 29.6 againt 28.8 on ROBUST and 28.8 against 27.9 on TREC 1&2. Even if these differences in performance may not be statistically significant, the difference in the number of terms (tc) used is in fact significant. The Log-logistic model need only 20 news terms whereas the mixture

model achieved its best performance with 100 new terms. This fact leads to us to study the influence of the number of added terms c for different PRF models.

So, we decide to compare several PRF models and their performance for different values of tc:

- the mixture model [84]
- the divergence minimization model [84]
- our log-logistic feedback model
- the recently proposed Geometric Relevance Model (GRM) [75]

These models are reviewed later in section 6.6, and their exact formulation is not necessary here. For all models, the different parameter values (including the interpolation weight) were optimized on all queries.

Figure 6.1 compares the best performance of these 4 different models when the number of feedback terms, tc, varies. The plots clearly indicate that the log-logistic model outperforms the other models and that it does so with fewer terms.

To sum up, we have shown that the extension of information models for PRF yield better performances. We have compared our models to several PRF baselines and through different experiment methodologies such as full cross-validation or a fixed parameter strategy. In both cases, our models seem to be more robust and are able to deliver significant improvements.

Having observed that the information-based models outperforms state of the art model with pseudo relevance feedack, we now want to better understand the reasons of these improvements.

We could argue that these improvements are due to the bursty nature of the probability distributions we used. In fact, state-of-the-art PRF models rely on non-bursty distributions. This is a theoretical argument and we should be able to explain empirically the reason of these improvements. This is why we will analyze PRF models results in order to better understand the different aspects involved in PRF and to answer the following questions:

1. *Why does the log-logistic model perform well with few terms ?*
2. *Why do other models fail in the same cases ?*

6.4 PRF Result Analysis

In order to better understand the behavior of PRF models, we raise the following questions that will guide the experimental analysis of PRF results:

1. Do PRF models agree on the words to select ?
2. What is the profile of words extracted by a given PRF model ?

To answer these questions, we analyze the terms chosen by the previously mentioned models when few terms are used, through two settings:

- setting A, with $n = 10$ and $tc = 10$
- setting B, with $n = 20$ and $tc = 20$

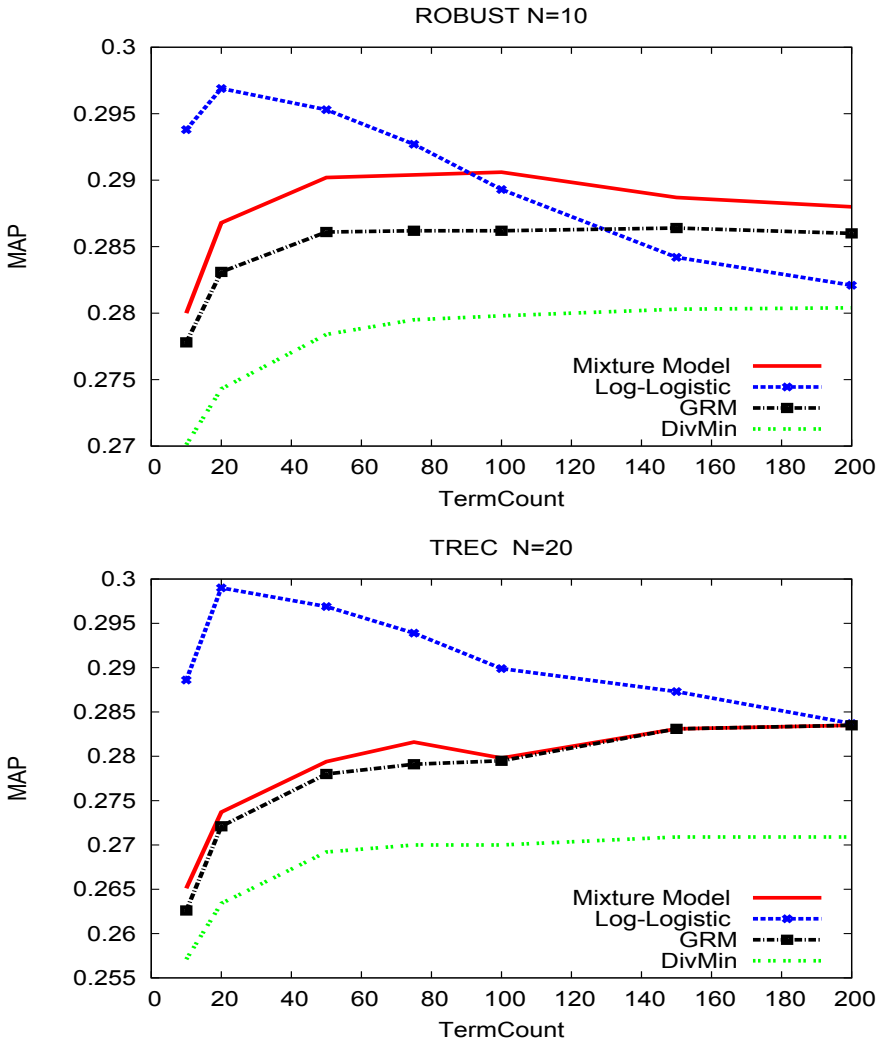


Figure 6.1: MAP on all queries with $tc \in \{10, 20, 50, 75, 100, 150, 200\}$ with best parameters, ROBUST $n = 10$ top, TREC-1&2 $n = 20$ bottom

Table 6.4: Statistics of the size of the Intersection

Collection	n	tc	Mean	Median	Std
robust	10	10	5.58	6.0	1.60
trec-12	10	10	5.29	5.0	1.74
robust	20	20	12	12	3.05
trec-12	20	20	11.8	13	3.14

As the log-logistic feedback model outperform the other feedback models with fewer feedback terms, we focus here on settings A and B, which will allow us to reveal several interesting properties of the different feedback models.

In order to have an unbiased comparison, we use the same IR engine for the retrieval step. Thus, all PRF algorithms are computed on the *same set of documents*. Once new queries are constructed, we use either the Dirichlet language model (for the new queries obtained with language model methods) or the log-logistic model (for the new queries obtained with the mean log-logistic information model) for the second retrieval step, thus allowing one to compare the performance obtained by different methods on the same initial set of PRF documents.

We first focus on a direct comparison between the mixture model and the mean log-logistic information model, by comparing the terms common to both feedback methods, i.e. the terms in the intersection of the two selected sets. Table 6.4 displays the mean, median and standard deviation of the size of the intersection, over all queries, for the collections considered. As one can note, the two methods agree on a little more than half of the terms (ratio mean by tc), showing that the two models select different terms.

We want to understand which 'kind' of words each PRF model choose. For example, do PRF models choose common words or rare words ? This is why we want to find the typical profile of words that a method tends to choose. To have a closer look at the terms selected by both methods, we first compute, for each query, the total frequency of a word in the feedback set (i.e. $TF(w)$) and the document frequency of this word in the feedback set (i.e. $DF(w)$). Then, for each query we can compute the mean frequency of the selected terms in the feedback set as well as its mean document frequency, i.e. $q(tf)$ and $q(df)$:

$$q(tf) = \sum_{i=1}^{tc} \frac{tf(w_i)}{tc} \text{ and } q(df) = \sum_{i=1}^{tc} \frac{df(w_i)}{tc}$$

We then compute the mean of the quantities over all queries.

$$\mu(tf) = \sum_{q} \frac{q(tf)}{|Q|} \text{ and } \mu(df) = \sum_{q} \frac{q(df)}{|Q|}$$

An average idf can be computed in exactly the same way. Table 6.5 displays the above statistics for the three feedback methods: mixture model (MIX), mean log-logistic(LL) information model and divergence minimization model (DIV). Regarding the mixture and log-logistic models, on all collections, the mixture model chooses in average words that have a *higher TF*, and a smaller *DF*. The mixture model also chooses words that are *more frequent in the collection* since the mean IDF values are smaller. On the other hand, the statistics of the divergence model and the geometric relevance models shows that these models extracts very common terms, with low IDF and high DF, which, as we will see later, is one of the main drawback of these models.

In addition to the term statistics, the performance of each PRF algorithm can also be assessed with different settings.

Table 6.5: Statistics of terms extracted by. Suffix A means $n = 10$ and $tc = 10$ while suffix B means $n = 20$ and $tc = 20$

Settings	Statistics	MIX	LL	DIV	GRM
robust-A	$\mu(tf)$	62.9	46.7	53.9	52.33
	$\mu(df)$	6.4	7.21	8.55	8.4
	μ(idf)	4.33	5.095	2.20	2.40
trec-1&2-A	$\mu(tf)$	114 .0	79.12	92.6	92.27
	$\mu(df)$	7.1	7.8	8.77	8.72
	μ(idf)	3.84	4.82	2.51	2.56
robust-B	$\mu(tf)$	68.6	59.9	65.27	64.57
	$\mu(df)$	9.9	11.9	14.7	14.38
	μ(idf)	4.36	4.37	1.66	1.93
trec-1&2-B	$\mu(tf)$	137.8	100.0	114.9	114.8
	$\mu(df)$	12.0	13.43	15.17	15.23
	μ(idf)	3.82	4.29	2.10	2.25

raw We first examine the performance of the feedback terms *without* mixing them with the original queries, a setting we refer to as *raw*.

interse For each query, we keep only (new) terms that belong to the intersection of the mixture (or divergence) and log-logistic models, but keep their weight predicted by each feedback method. We call this setting *interse*.

diff A third setting, *diff*, consists in keeping terms which do not belong to the intersection.

interpo The last setting, *interpo* for interpolation, measures the performance when new terms are mixed with the original query. This corresponds to the standard setting of pseudo-relevance feedback

Table 6.6 displays the results obtained. As one can note, the log-logistic model performs better than the mixture model. What our analysis reveals is that it does so because it chooses better feedback terms, as shown by the performance of the *diff* setting. For the terms in the intersection, method *interse*, the weights assigned by the log-logistic model seem more appropriate than the weights assigned by the other feedback models.

Summary of Analysis

We have analyzed the words selected by several PRF models. This analysis has demonstrated that the log-logistic model agrees with the mixture model on approximately 50% of the words to choose. Furthermore, we have shown that the words only selected by the log-logistic model tends to perform better than the words only selected by the mixture model. The experiments have also demonstrated that:

1. The mixture, GRM and divergence models choose terms with a *higher* TF
2. GRM and the Divergence model select terms with a smaller IDF

Hence, the log-logistic model perform better because it chooses better words and these words tends to have a lower TF and bigger DF in a feedback set

6.5 An Axiomatic Approach to PRF

Our previous experiments provide empirical explanations of the behavior of PRF models. However, it would be interesting to link these observations to the properties of PRF models

Table 6.6: Performance of different methods. Suffix A means $n = 10$ and $tc = 10$ while suffix B means $n = 20$ and $tc = 20$

Settings	FB Model	MIX	LL	DIV
robust-A	raw	23.8	26.9	24.3
	interse	24.6	25.7	24.1
	diff	3	11.0	0.9
	interpo	28.0	29.2	26.3
trec-1&2-A	raw	23.6	25.7	24.1
	interse	24.2	24.5	23.4
	diff	3	9	0.9
	interpo	26.3	28.4	25.4
robust-B	raw	23.7	25.7	22.8
	interse	25.3	26.2	22.6
	diff	3.0	10.0	0.15
	interpo	28.2	28.5	25.9
trec-1&2-B	raw	25.1	27.0	24.9
	interse	26.1	26.5	24.7
	diff	2.1	11.2	0.5
	interpo	27.3	29.4	25.7

so as to understand from a theoretical standpoint the performance of PRF models.

Furthermore, several recently proposed PRF models seem to outperform the mixture model as well, as models based on bagging, models based on a mixture of Dirichlet compound multinomial distributions [24, 82]. The performance of models nevertheless varies from one study to another, as different collections and different ways of tuning model parameters are often used. It is thus very diffcult to draw conclusions on the characteristics of models. What is lacking is a theoretical framework which would allow one to directly compare PRF models, independently of any collection. This is the goal we pursue in this section, where we want to build an axiomatic theory for PRF models.

Axiomatic methods were pioneered by Fang et al [33] and followed by many works including [34, 26, 17] and we gave an description of these methods in chapter 4. In a nutshell, axiomatic methods describe IR functions by constraints they should satisfy. We build on the main conditions an IR function should satisty (cf chapter 4): *the TF, Concave, Doc Length and IDF effect.*

In the context of PRF, the first two constraints relate to the fact that terms frequent in the feedback set are more likely to be effective for feedback, but that the difference in frequencies should be less important in high frequency ranges. The IDF effect is also relevant in feedback, as one generally avoids selecting terms with a low IDF, as such terms are scored poorly by IR system

Let $FW(w; \mathbf{F}, \mathbf{P}_w)$ denote the feedback weight for term w, with $\mathbf{P}_w$ a set of parameters dependent on w[2]. We use as shorthand $FW(w)$ but it important to keep in mind that this function depends on a feedback set and some parameters. We can formalize the above considerations as follows:

[TF effect] *FW increases the normalized term frequency t_{wd}; in analytical terms, this constraint translates as:*

$$\frac{\partial FW(w)}{\partial t_{wd}} > 0$$

[2]The definition of $\mathbf{P}_w$ depends on the PRF model considered. It minimally contains $TF(w)$, but other elements, as IDF(w), are also usually present. We use here this notation for convenience.

[Concavity effect] *The above increase should be less marked in high frequency ranges, which can be formulated as:*

$$\forall d \in \mathbf{F},\ \frac{\partial^2 FW(w)}{\partial t_{wd}^2} < 0$$

[IDF effect] *Let w_a and w_b two words such that $idf(w_b) > idf(w_a)$ and $\forall d \in \mathbf{F},\ t(w_a, d) = t(w_b, d)$. Then*

$$FW(w_b) > FW(w_a).$$

[Document length effect] *The number of occurrences of feedback terms should be normalized by the length of docs they appear in.*

$$\frac{\partial FW(w)}{\partial l_d} < 0$$

The form of the IDF effect retained here is dictated by the particular setting we place ourselves in, namely the one of Pseudo-Relevance Feedback. In this setting, we want to study the increase of the feedback weight wrt IDF, *all other things being equal*. This forces the introduction of the condition on the distribution of frequencies over the feedback documents.

We now introduce a PRF property which is based on the results reported in the previous section. Indeed, as we have seen, the best PRF results were obtained with the log-logistic models which favor feedback terms with a high *document frequency* ($DF(w)$) in the feedback set, which suggests that, *all things being equal*, terms with a higher DF should receive a higher score:

[DF effect] *Let $\epsilon > 0$, and w_a and w_b two words such that:*
(i) IDF(a) = IDF(b)
(ii) The distribution of the frequencies of w_a and w_b in the feedback set are given by:

$$\begin{aligned} T(w_a) &= (t_1, t_2, ..., t_j, 0, ..., 0) \\ T(w_b) &= (t_1, t_2, ..., t_j - \epsilon, \epsilon, 0, ..., 0) \end{aligned}$$

with $\forall i,\ t_i > 0$ and $t_j - \epsilon > 0$ (hence, $TF(w_a) = TF(w_b)$ and $DF(w_b) = DF(w_a) + 1$).

Then: $FW(w_a; \mathbf{F}, \mathbf{P}_{w_a}) < FW(w_b; \mathbf{F}, \mathbf{P}_{w_b})$

In other words, FW is *locally* increasing with $DF(w)$. The following theorem is useful to establish whether a PRF model, which can be decomposed in the documents of $\mathbf{F}$, enforces the DF effect:

Theorem 6. *Suppose FW can be written as:*

$$FW(w; \mathbf{F}, \mathbf{P}_w) = \sum_{d=1}^{n} f(t_{wd}; \mathbf{P'}_w) \tag{6.5}$$

with $\mathbf{P'}_w = \mathbf{P}_w \setminus t_{w,d}$ and $f(0; \mathbf{P'}_w) \geq 0$. Then:

1. *If the function f is strictly concave, then FW enforces the DF effect.*
2. *If the function f is strictly convex, then FW does not enforce the DF effect.*

Proof If f is strictly concave then, the function f is subadditive ($f(a+b) < f(a)+f(b)$). Let a and b be two words such as in the definition of the DF effect.

$$FW(a) = FW(\underbrace{t_1, ..., t_j}_{DF}, \underbrace{0, ..., 0}_{n-DF}) \tag{6.6}$$

$$FW(b) - FW(a) = f(t_j - \epsilon) + f(\epsilon) - f(t_j) > 0 \tag{6.7}$$

which hold as the function f is subadditive. If f is convex, then f is superadditive as $f(0) = 0$, which shows that $FW(b) - FW(a) < 0$. □

As one can note, as the sum of concave functions is concave, feedback functions of the form given by equation 6.5 enforce both the concavity and the DF effects. However, as we will see, there exist models which enforce the DF effect but not the concavity effect.

Prior to assess the validity of the DF constraint, we want to mention a last constraint, which is introduced in [51] and which we will refer to as *Document Score constraint*. This constraint, implemented in relevance models [48] and in the Rocchio algorithm [39], can be formulated as follows:

PRF Constraint 1. [Document Score - DS]
When $FW(w; \mathbf{F}, \mathbf{P}_w)$ explicitly depends on the documents of $\mathbf{F}$ in which w occurs, then documents with a higher score (defined by $RSV(q, d)$) should be given more weight.

The importance of this constraint is however not fully clear, and the models which explicitly integrate it do not count among the best PRF models. For example, in the study conducted in [51], a simple mixture model outperforms models integrating document scores. Furthermore, our modifications of the log-logistic model to take this constraint into account did not lead to any significant improvement, so that we are not sure of the status one should give to this constraint. As we will discuss in section 6.7, the strategies consisting of resampling feedback documents, and proposed for example in [24] and [49], may be effective and model-independent ways of integrating this constraint.

6.5.1 Validation of the DF Constraint

One way to assess the validity of the DF constraint is to determine whether DF values are related with MAP scores in relevance feedback settings. Indeed, the DF constraint states that, all other parameters being equal, terms with higher DF should be preferred. Thus, in average, one should observe that terms with high DF scores yield larger increase in MAP values. To see whether this is the case, we computed the impact on the MAP of different terms selected from true relevance judgments, and plotted this impact against both TF and DF values. Our relying on true relevant documents and not documents obtained from pseudo-relevance feedback is based on (a) the fact that pseudo-relevance feedback aims at approximating relevance feedback, an (b) the fact that it is more difficult to observe clear trends in pseudo-relevance sets where the precision (e.g. P@10) and MAP of each query have large variances. The framework associated with true relevance judgments is thus cleaner and allows easier interpretation. In order to assess the impact of DF scores on the MAP values independently of any IR model, we make use of the following experimental setting:

- Start with a first retrieval with a Dirichlet language model;
- Let R_q denote the set of relevant documents for query q: Select the first 10 relevant documents if possible, else select the top $|R_q|(|R_q| < 10)$ relevant documents;
- Construct a new query (50 words) with the mixture model;

- Construct a new query (50 words) with the log-logistic model;
- Compute statistics for each word in the new queries.

Statistics include a normalized DF, equal to $DF(w)/|R_q|$, and a normalized TF, first using a document length normalization, then using the transformation $\log(1+TF(w))/|R_q|$ to avoid too important a dispersion in plots. Each word w is added independently with weights predicted by the retained PRF model. For each word w, we measure the MAP of the initial query augmented with this word. The difference in performance with the initial query is then computed as: $\Delta(\text{MAP}) = \text{MAP}(q + FW(w)w) - \text{MAP}(q)$. We thus obtain, for each term, the following statistics:

- Δ(MAP)
- $\log(1 + TF(w))/|R_q|$
- $DF(w)/|R_q|$

Figures 6.2 and 6.3 display a 3D view of these statistics for all queries, based on Gnuplot and two collections: TREC1&2 and ROBUST. In order to have a better view of the patterns obtained, we have used a 30x30 grid, and two kernel smoothers. The use of a kernel smoother K amounts to smooth the value at any point by the values obtained at neighboring points:

$$\Delta(\text{MAP})(x) = \sum_{i=1}^{p} \Delta(\text{MAP})(x_i) K(x, x_i)$$

where $i, 1 \leq i \leq p$ indexes the set of feedback terms. We use here both an asymmetric, exponential kernel and a symmetric, gaussian kernel. As mentioned before, the TF statistics was normalized to account for different lengths. The shape of the plots obtained remains however consistent without any normalization or when standard normalizations are used. The plots displayed in Figures 6.2 and 6.3 are based on the DFR normalization:

$$TF(w) = \sum_{d \in R_q} x_{w,d} \log(1 + c\frac{avgl}{l_d})$$

As one can note, on all plots of Figures 6.2 and 6.3, the best performing regions in the (TF,DF) space correspond to large DFs. Furthermore, for all TF values, the increase in MAP parallels the increase in DF (or, in other words, Δ(MAP) increases with DF for fixed TF), for both exponential and gaussian kernels. This validates the DF constraint and shows the importance of retaining terms with high DF in relevance feedback. Interestingly, the reverse is not true for TF. Indeed, for fixed DF, Δ(MAP) does not increase nicely with TF, when using the exponential kernel (it does however on the two collections with the gaussian kernel). This implies that if terms with large TF are interesting, they should not be given too much weight. The results displayed in Table 6.5 suggest that the mixture model [84] suffers from this problem.

6.5.2 Validation of IDF Effect

We follow the previous approach to asses the IDF effect. From true relevance feedback, we extract the following statistics:

- Δ(MAP)
- $TF(w)/|R_q|$
- $IDF(w)$

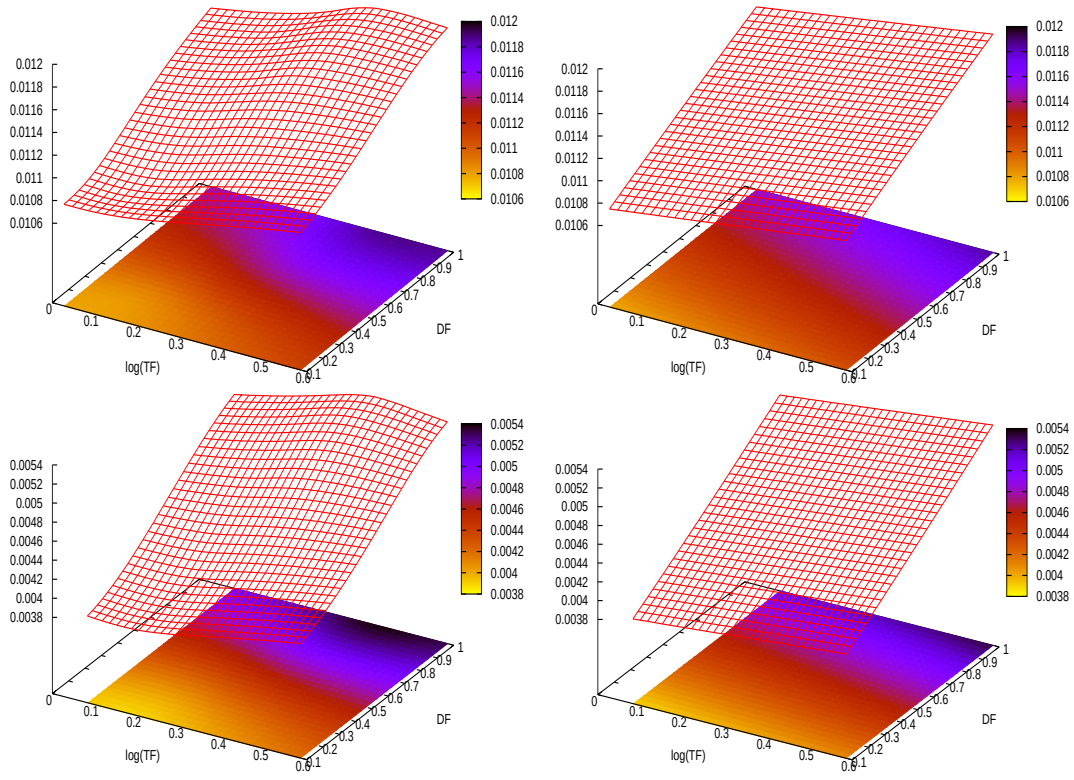


Figure 6.2: (log(TF),DF) vs Δ MAP on TREC-12; true relevant documents are used with $n = 10$, $tc = 50$ and exponential (left) and Gaussian (right) kernel grids (15×15). Top row: log-logistic model; bottom row: language model

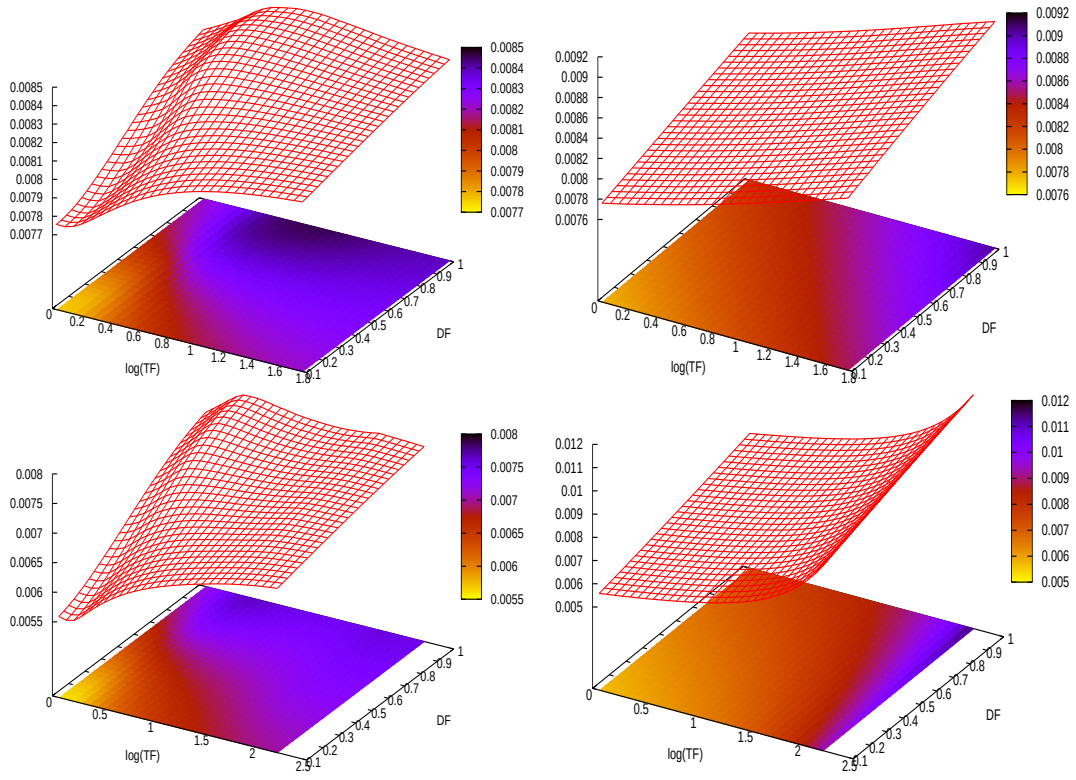


Figure 6.3: (log(TF),DF) vs Δ MAP on ROBUST; true relevant documents are used with $n = 10$, $tc = 50$ and exponential (left) and Gaussian (right) kernel grids (15×15). Top row: log-logistic model; bottom row: language model

Figures 6.4 and 6.5 show the 3D-view of these statistic for the Log-Logistic model and the mixture model for both Gaussian and Exponential kernels. First of all, the plots on the ROBUST collection look different from the figures for the TREC collection. Second, the log-logistic plots are very different from the mixture model on the ROBUST collection but very similar on the TREC collection.

On the ROBUST collection, the log-logistic statistic seem to support the IDF condition. Note that for the Gaussian Kernel figure, there is a low density region for high idf. This can be explain by the fact it is not possible in general to have both a high idf and a high TF, thus explaining why there is no word selected by the log-logistic model in this area. The situation concerning the mixture model is puzzling at first sight and does not support the IDF condition. This behavior can be explained by the fact that the mixture model selects words with a high TF. Hence, among the 50 words selected by query, there may not be many rare words, which explain why certain regions of the (TF,IDF) space have a low density.

On the other hand, the figures for the other TREC collection suggest that it is important to penalize common words (low idf) and rare words (high idf). Then, the IDF condition does not appear completely valid from these observations. It does capture the fact that common words should be penalized but the data shows that rare words should be penalized as well. However, the TF and DF effect do penalize rare words, which compensate the deficiency of the IDF condition for rare words. Overall, the figures on the two collections support the idea that common words should be penalized, thus justifying the IDF condition. Even if the IDF condition is not fully valid alone, it becomes more adequate with the TF and DF condition, which will be able to filter rare terms.

6.5.3 Validation of the different conditions with a TF-IDF family

In order to further validate the constraints, let us introduce the family of feedback functions defined by:

$$t_{wd} = x_{wd} \log(1 + c\frac{avg\_l}{l_d})$$
$$FW(w) = \sum_{d \in F} t_{wd}^k \text{IDF}(w) \tag{6.8}$$

This equation amounts to a standard *tf-idf* weighting, with an exponent k which allows one to control the convexity/concavity of the feedback model.

Because of the form of t_{wd}, and the way IDF(w) is taken into account, the above family of functions satisfies the first, second and fourth CIR constraints. If $k > 1$ then the function is strictly convex and, according to Theorem 6, does not satisfy the DF constraint. Furthermore, $\forall d \in \mathbf{F}$, $\frac{\partial^2 FW(w)}{\partial t^2} > 0$ so that the second CIR constraint is not satisfied either. On the contrary, if $k < 1$, then the function is strictly concave and satisfies the DF constraint as well as all the CIR constraints. The linear case, being both concave and convex, is *in-between*.

One can then build PRF models from equation 6.8 with varying k, and see whether the results agree with the theoretical findings implied by Theorem 6. We used the reweighting scheme of equation 6.8 with equation 6.4.

Table 6.7 displays the term statistics ($\mu(tf)$,$\mu(df)$, mean IDF) for different values of k. As one can note, the smaller k, the bigger $\mu(df)$ is. In other words, the slower the function grows, the more terms with large DF are preferred.

Table 6.8 displays the MAP for different values of k. At least two important points arise from the results obtained. First, convex functions ($k > 1$) have lower performance than concave functions for all datasets, and the more a model violates the constraints,

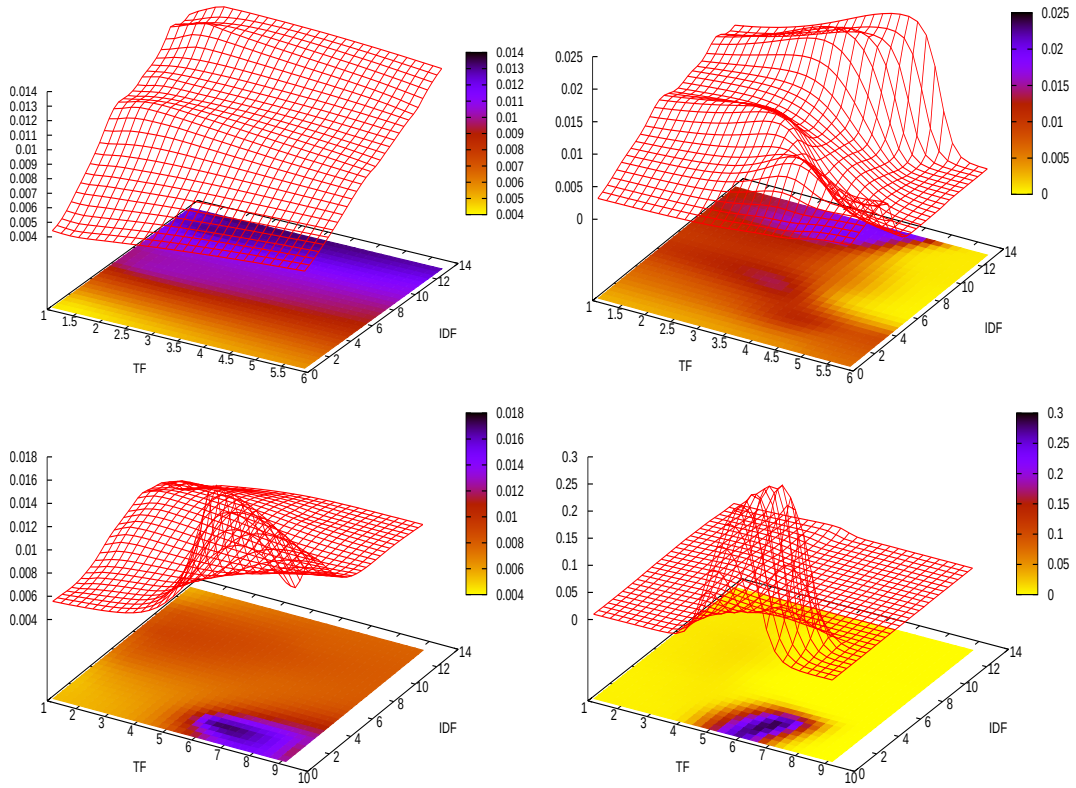


Figure 6.4: True Oracle (TF,IDF) vs Δ MAP sur ROBUST $n = 10$ $tc = 50$. Exponential (left) and Gaussian (right) Kernel Grid 15×15. LG(top row) LM(bottom)

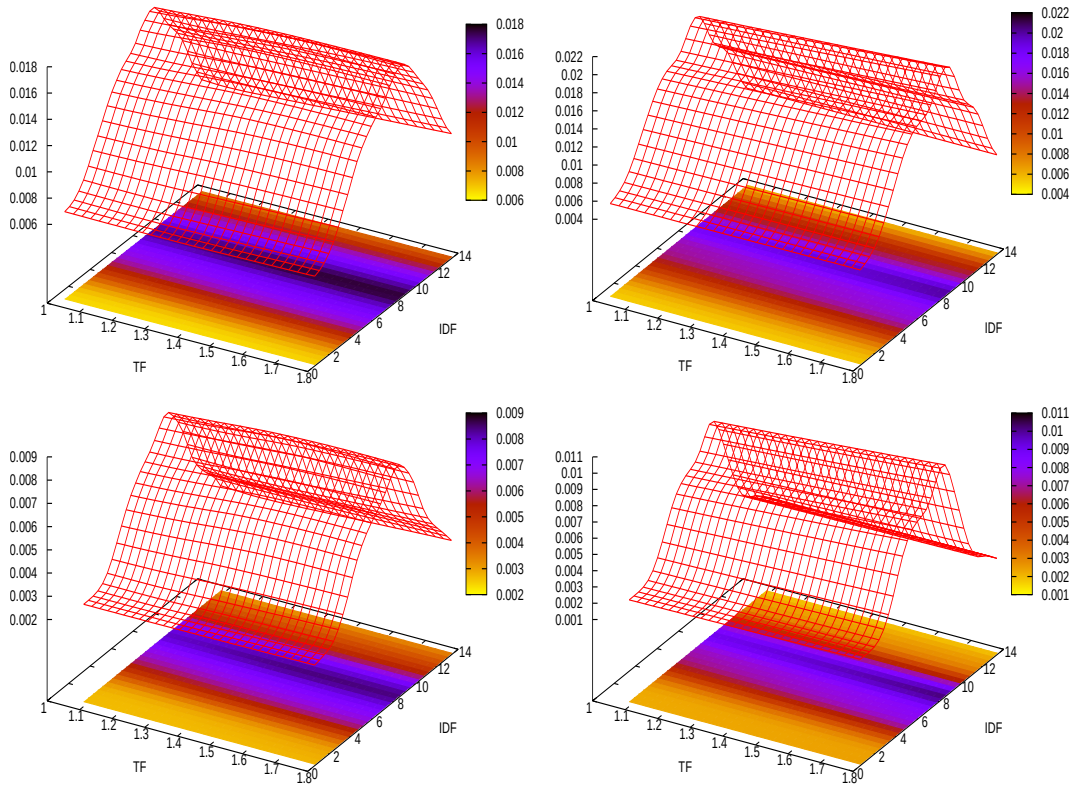


Figure 6.5: True Oracle (TF,IDF) vs Δ MAP sur TREC-1&2 $n = 10$ $tc = 50$. Exponential (left) and Gaussian (right) Kernel Grid 15×15. LG(top row) LM(bottom)

Table 6.7: Statistics on TREC-12-A

Power k	$\mu(tf)$	$\mu(df)$	Mean IDF
0.2	70.46	7.4	5.21
0.5	85.70	7.1	5.09
0.8	88.56	6.82	5.14
1	89.7	6.6	5.1
1.2	91.0	6.35	5.1
1.5	90.3	6.1	5.0
2	89.2	5.8	4.9

Table 6.8: MAP for different power function. Suffix A means $n = 10$ and $tc = 10$ while suffix B means $n = 20$ and $tc = 20$

Power k	robust-A	trec-12-A	robust-B	trec-12-B
0.2	29.3	28.7	28.7	30.0
0.5	**30.1**	**29.5**	**29.4**	**30.5**
0.8	29.6	29.3	29.4	30.3
1	29.2	28.9	29.1	29.9
1.2	28.9	28.6	28.6	29.6
1.5	28.6	28.1	28.3	28.9
2	28.1	27.2	27.4	28.0
log-logistic	29.4	28.7	28.5	29.9

the worse it is. This confirms the validity of the constraints we have reviewed. Second, the square root function ($k = 0.5$) has the best performance on all collections: it also outperforms the standard log-logistic model. When the function grows slowly (k equals to 0.2), the DF statistics is somehow preferred compared to TF. The square root function achieves a different and better trade-off between the TF and DF information. This is an interesting finding as it shows that the TF information is still useful and should not be too downweighted wrt the DF one.

6.6 Review of PRF Models

We review in this section different PRF models according to their behavior wrt the characterizations we have defined. We start with language models, then review the recent model introduced in [82] which borrows from both generative approaches *à la* language model and approaches related to the *Probability Ranking Principle* (PRP), prior to review Divergence from Randomness (DFR) and Information-based models.

6.6.1 PRF for Language Models

PRF models within the language modeling (LM) approach to information retrieval assume that words in the feedback document set are distributed according to a Multinomial distribution, θ_F (the notation θ_F summarizes the set of parameters $P(w|\theta_F)$). Once the parameters have been estimated, PRF models in the LM approach proceed by interpolating the (original) query language model with the feedback query model θ_F:

$$\theta_{q'} \quad = \quad \alpha\theta_q + (1-\alpha)\theta_F \tag{6.9}$$

In practice, one restricts θ_F to the top tc words, setting all other values to 0. The different feedback models then differ in the way θ_F is estimated. We review the main LM based feedback models below.

Mixture Model

Zhai and Lafferty [84] propose a generative model for the set **F**. All documents are i.i.d and each document is generated from a mixture of the feedback query model and the corpus language model:

$$P(\mathbf{F}|\theta_F, \beta, \lambda) = \prod_{w=1}^{V} ((1-\lambda)P(w|\theta_F) + \lambda P(w|C))^{TF(w)} \tag{6.10}$$

where λ is a "background" noise set to some constant. Finally θ_F is learned by optimizing the data log-likelihood with an Expectation-Maximization (EM) algorithm, leading to the following E and M steps at iteration (i):

$$E-step \quad E(w)^{(i)} = \frac{(1-\lambda)P^{(i}(w|\theta_F)}{(1-\lambda)P^{(i)}(w|\theta_F)+\lambda P^{(i)}(w|C)}$$
$$M-step \quad P^{(i+1)}(w|\theta_F) = \frac{\sum_{d\in\mathbf{F}} x_{wd})E(w)^{(i)}}{\sum_w \sum_{d\in\mathbf{F}} x_{wd}E(w)^{(i)}}$$

where $E(w)^{(i)}$ denotes the expectation of observing w in the feedback set; furthermore, $FW(w) = P(w|\theta_F)$. As one can note, none of the above formulas involve $DF(w)$, neither directly nor indirectly. The mixture model is thus agnostic wrt to DF, and thus does not enforce the DF effect.

Regarding the other properties, one can note that the weight of the feedback terms ($P(w|\theta_F)$) increases with $TF(w)$ (which is $\sum_{d\in\mathbf{F}} x_{w,d}$), decreases with IDF($w$) (the argument for this is the same as the one developed in [33], a study to which we refer readers). Thus, both TF and IDF effects are enforced.

Furthermore, even though counts are normalized by the length (in fact an approximation of it) of the feedback documents, all these documents are merged together, so that the Document length effect is not fully enforced.

The situation wrt the Concavity effect is even less clear. In particular, if one approximates the denominator with the length of the feedback documents (such an approximation being based on the fact that $E(w)^{(i)}$ corresponds to the expectation of w in the feedback set), then the second partial derivative of $P(w|\theta_F)$ wrt to $t(w,d)$ is 0. This suggests that this model does not fully enforce the Concavity effect, and thus that it gives too much weight to high frequency words. This is indeed what we have observed in table 6.5: the mixture model selects terms with a mean TF which is significantly higher than the mean TF of the other models.

Divergence Minimization

In addition to the mixture model, a divergence minimization model:

$$D(\theta_q|RF) = \frac{1}{|n|}\sum_{i=1}^{n} D(\theta_F \parallel \theta_{d_i}) - \delta D(\theta_F||p(.\parallel C))$$

is also proposed in [84], where θ_{d_i} denotes the empirical distribution of words in document d_i. Minimizing this divergence gives the following solution:

$$P(w|\theta_F) \propto \exp\Big(\frac{1}{(1-\delta)}\frac{1}{n}\sum_{i=1}^{n}\log(p(w|\theta_{d_i})) - \frac{\delta}{1-\delta}\log(p(w|C))\Big)$$

Here again, $FW(w) = P(w|\theta_F)$. This equation corresponds to the form given in equation 6.5 with a strictly concave function (log). Thus, by Theorem 6, this model enforces the DF effect. It also enforces the TF, Concavity and Document length effects.

Our previous experiments, reported in table 6.6, as well as those reported in [51], show that this model does not perform as well as other ones. Indeed, as shown in table 6.5, the IDF effect is not sufficiently enforced, and the model fails to downweight common words.

Let us consider two terms w_a and w_b such that $\forall d \in \mathbf{F}\ t(w_a, d) = t(w_b, d) = t_d$, and let $z_b = p(w_b|C)$ and $z_a = p(w_a|C)$ such that $z_a < z_b$ (we use z to ease notations). The IDF effect stipulates that, in this case, $FW(a)$ should be greater than $FW(b)$. Using Jelinek Mercer smoothing, $FW(w_a) - FW(w_b)$ has the sign of:

$$\sum_{d \in \mathbf{F}} \{ \overbrace{\log(\frac{(1-\lambda)t(d) + \lambda z_a}{(1-\lambda)t(d) + \lambda z_b})}^{>0} - \delta \overbrace{\log(\frac{z_a}{z_b})}^{<0} \} \tag{6.11}$$

As one can note, the above quantity is not necessarily negative, especially when δ is small, which can happen in practice as δ is optimized on an independent training set. This shows that the divergence minimization model is not guaranteed to enforce the IDF effect. Furthermore, if one considers two words w_a and w_b which occur in only 1 document out of n, a necessary condition for the divergence minimization model to enforce the IDF effect is $\frac{n-1}{n} < \delta$, which shows again that δ values close to 1 are required (in practice, typical values obtained are in the range of 0.1). This explains the small values displayed in table 6.5 for the IDF statistic.

Other PRF Methods fo Language Models

A regularized version of the mixture model, known as the regularized mixture model (RMM) and making use of latent topics, is proposed in [78] to correct some of the deficiencies of the simple mixture model. RMM has the advantage of providing a joint estimation of the document relevance weights and the topic conditional word probabilities, yielding a robust setting of the feedback parameters. However, the experiments reported in [51] show that this model is less effective than the simple mixture model in terms of retrieval performance. We will thus not study it further here, but want to mention, nevertheless, an interesting re-interpretation of this model in the context of the concave-convex procedure framework [29] which we will discuss in section 6.7.

Another PRF model proposed in the framework of the language modeling approach is the so-called relevance model, proposed by Lavrenko *et al.* [48], and defined by:

$$FW(w) \quad \propto \quad \sum_{d \in \mathbf{F}} P_{LM}(w|\theta_d) P(d|q) \tag{6.12}$$

where P_{LM} denotes the standard language model. Furthermore, it corresponds to the form of equation 6.5 of Theorem 6, with a linear function, which is neither strictly concave nor strictly convex. This model is neutral wrt the DF effect.

Regarding the IDF effect, Let w_a and w_b two words such that $p(w_a|C) > p(w_b|C)$ and $\forall d \in \mathbf{F}\ t(w_a, d) = t(w_b, d) = t_d$. Using Jelinek-Mercer smoothing, we have

$$\begin{aligned} FW(a) - FW(b) &= \sum_{d \in \mathbf{F}} P(d|q)\big((1-\lambda)t_d + \lambda p(w_a|C) - (1-\lambda)t_d - \lambda p(w_b|C)\big) \\ &= \sum_{d \in \mathbf{F}} P(d|q)\lambda(p(w_a|C) - p(w_b|C)) \quad > 0 \end{aligned}$$

which shows that relevance models do not satisfy the IDF condition.

The relevance model has recently been refined in the study presented in [75] through a geometric variant, referred to as GRM, and defined by:

$$FW(w) \propto \prod_{d \in \mathbf{F}} P_{LM}(w|\theta_d)^{P(d|q)} \tag{6.13}$$

Let us consider this model with Jelinek-Mercer smoothing [86]: $P_{LM}(w|\theta_d) = (1-\lambda)\frac{x_{wd}}{l_d} + \lambda\frac{F_w}{L}$, . Let w_a and w_b be two words as defined in the DF effect, and let us further assume that feedback documents are of the same length l and equiprobable given q. Then $FW(w_a)$ and $FW(w_b)$ respectively differ on the two quantities:

$$(i) \qquad \overbrace{((1-\lambda)\frac{x_{w_a,d_j}}{l} + \lambda\frac{F_{w_a}}{L})}^{A}\overbrace{(\lambda\frac{F_{w_b}}{L})}^{B}$$

$$(ii) \quad ((1-\lambda)\frac{x_{w_a,d_j}-\epsilon}{l} + \lambda\frac{F_{w_a}}{L})(\underbrace{(1-\lambda)\frac{\epsilon}{l}}_{\epsilon'} + \lambda\frac{F_{w_b}}{L})$$

The second quantity amounts to:

$$(A-\epsilon')(B+\epsilon') = AB + \epsilon'(A-B) - (\epsilon')^2$$

But $A - B = (1-\lambda)\frac{x_{w_a,d_j}}{l}$, a quantity which is strictly greater than $(1-\lambda)\frac{\epsilon}{l} = \epsilon'$ by the assumptions of the DF effect. Thus, the GRM model enforces the DF effect when Jelinek-Mercer is used. However, this model fails to enforce the IDF effect. Let w_a and w_b two words such that $p(w_a|C) > p(w_b|C)$ and $\forall d \in \mathbf{F}\ t(w_a,d) = t(w_b,d) = t_d$.

$$FW(w_a) - FW(w_b) =_{\text{sign}} \sum_d P(d|q) \log \frac{\lambda t_d + (1-\lambda)p(w_a|C)}{\lambda t_d + (1-\lambda)p(w_b|C)} > 0$$

which is strictly positive.This explains the results displayed in table 6.5, showing that the GRM model selects terms with low IDF.

6.6.2 PRF under the PRP

In [70], the *offer weight* is proposed to perform PRF under the PRP:

$$FW(w) = \frac{DF(w)}{n}\overbrace{\log\frac{DF(w)+0.5}{N_w - DF(w)+0.5}\frac{n-DF(w)+0.5}{N-N_w-n+DF(w)+0.5}}^{RSJ}$$

The offer weight is simply the product of the Document frequency times the Robertson Sparck Jones (RSJ) weight. Roberston explain in that the RSJ weight tends to favor too much rare words and that the correction by the DF factor correct this problem and thus improve performances. Checking the DF condition is not straightforward for this model and is left for future work. If the RSJ weight can be assimilated to an IDF weight, the above PRF model agrees with the IDF condition and the DF condition. However, it does not consider the TF effect.

Xu and Akella [82] propose an instanciation of the Probability Ranking Principle (PRP) in which relevant documents are assumed to be generated from a Dirichlet Compound Multinomial (DCM) distribution, or an approximation of it, called eDCM and introduced in [31]. The PRF version of this model simply assumes that the feedback documents are relevant. Terms are then generated according to two latent generative models based on the (e)DCM distribution and associated with two variables, relevant z_{FR} and non-relevant z_N. The variable z_N is intended to capture general English words occurring frequently in the whole collection, whereas z_{FR} is used to represent terms occurring in the feedback documents and pertinent to the user's information need. The parameters of the two components are estimated through rather time-consuming and complex estimation procedures, typically based on gradient descent or the EM algorithm. [82]

furthermore proposes two modifications of the EM algorithm to estimate the parameters of the relevant component, in a way similar to the one followed by [78]. Disregarding the non-relevant component for the moment, the weight assigned to feedback terms by the relevant component is given by (M-step of the EM algorithm):

$$P(w|z_{FR}) \propto \sum_{d \in \mathbf{F}} I(c(w,d) > 0) P(z_{FR}|d,w) + \lambda c(w,q)$$

This formula, being based on the presence/absence of terms in the feedback documents, enforces the DF effect. We need to check the situation of the TF condition as it is involved in E-step of the EM algorithm. That said, the higher the DF of a term, the higher its TF is likely to be, so that can nevertheless indirectly select high frequency terms by selecting terms with high DF.

We conjecture that this is the case with the (e)DCM model, which seems to behave well in practice. Finally, the EM steps also suggest that this model satisfy the IDF condition, as much as the mixture model does.

6.6.3 PRF in DFR and Information Models

In DFR and information models, the original query is modified to take into account the words appearing in $\mathbf{F}$ according to the following scheme:

$$q'_w = \frac{q_w}{\max_w q_w} + \beta \frac{\text{Info}_{\mathbf{F}}(w)}{\max_w \text{Info}_{\mathbf{F}}(w)} \tag{6.14}$$

where β is a parameter controlling the modification brought by $\mathbf{F}$ to the original query and q'_w denotes the updated weight of w in the query. In this case: $FW(w) = \text{Info}(w, \mathbf{F})$.

Bo Models

Standard PRF models in the DFR family are the Bo models [1], which are defined by:

$$\text{Info}(w, \mathbf{F}) = \log_2(1 + g_w) + TF(w) \log_2(\frac{1 + g_w}{g_w})$$

where $g_w = \frac{N_w}{N}$ in $Bo1$ model and $g_w = P(w|C)(\sum_{d \in F} l_d)$ in $Bo2$ model. In other words, documents in $\mathbf{F}$ are merged together and a geometric probability model (or a different distribution, the choice of the distribution being irrelevant for our argument) is used to measure the informative content of a word.

First, Bo models do account for the TF effect and IDF effect. Second, as this model is DF agnostic it does not enforce the DF effect. Furthermore, when using the geometric distribution, the Concavity effect is not enforced as the second derivative of $FW(w)$ wrt to $TF(w)$ is null. Neither does it enforce the Document length effect, as feedback documents are merged together.

Log-logistic Model

In information-based models, the average information brought by the feedback documents on given term w is used as a criterion to rank terms, which amounts to:

$$FW(w) = \text{Info}(w, \mathbf{F}) = \frac{1}{n} \sum_{d \in \mathbf{F}} -\log P(X_w > t(w,d)) | \lambda_w)$$

where $t(w,d)$ is the normalized number of occurrences of w in d, and λ_w a parameter associated to w and set to: $\lambda_w = \frac{N_w}{N}$. Two instanciations of the general information-based family are considered , respectively based on the log-logistic distribution and a

smoothed power law (SPL). The log-logistic model for pseudo relevance feedback is thus defined by:

$$t_{wd} = x_{wd} \log(1 + c\frac{avg\_l}{l_d})$$
$$FW(w) = \frac{1}{n} \sum_{d \in F} [\log(\frac{N_w}{N} + t_{wd}) + \text{IDF}(w)]$$

As the logarithm is a concave function, the log-logistic model enforces the DF effect by Theorem 6. Furthermore, it is compliant with all the other properties as it based on the general information formulation with a bursty distribution (as shown in [17]). Let w_a and w_b, two words such as in the IDF condition. Let $r_a = \frac{N_{wa}}{N}$ and $r_b = \frac{N_{wb}}{N}$ such that $r_b < r_a$, then:

$$\begin{aligned} FW(w_a) - FW(w_b) &= \sum_d \log\left(\frac{r_a + t_d}{r_a} \frac{r_b}{r_b + t_d}\right) \\ &= \sum_d \log \frac{r_a r_b + r_b t_d}{r_a r_b + r_a t_d} \quad < 0 \end{aligned}$$

which is unconditionally negative.

The smoothed power law model (SPL) satisfy the TF, DF and Document Length effect. The IDF effect is not straightforward to verify. The good performance obtained previoulsy with few terms and the fact that the SPL aims at approximating the DFR model $InL2$, where the IDF effect is clear, suggests that the SPL model satisfy the IDF condition. Nevertheless, we checked the IDF statistic of the SPL model as in table 6.5. The mean idf is 4.5 on robust-A and and 4.2 for trec-1&2-A, which suggest that the model do penalize common words.

6.6.4 Summary

The above theoretical study has revealed the following elements for the PRF models we have reviewed:

1. In the language modeling approach, the simple mixture model enforces neither the DF effect nor the Document length effect. The divergence minimization model does not unconditionally enforce the IDF effect. More surprisingly, the RM and GRM models do not enforce the IDF effect.

2. Considering models related to the *Probability Ranking Principle*, the relevance model proposed in [82] on the basis of the Dirichlet Compound Multinomial satisfies most properties, including the DF effect.

3. In the *Divergence from Randomness* approach, *Bo* models fail to enforce the DF effect, as well as the Concavity and Document length effects. In the family of information-based models, both the log-logistic and SPL models satisfy all the PRF properties.

Table 6.9 summarizes the previous analysis. These theoretical results provide a good explanation of the statistics collected on two large collections and displayed in table 6.5. They also provide an explanation for the good behavior of the log-logistic and SPL models developed in the framework of the information-based family. These two models outperform the other models in PRF settings. The log-logistic model enforces all the PRF effects we have reviewed, for all the admissible values of its parameter. The SPL model seems to satisfy all conditions: even if the situation with respec to IDF effect is not proven, there are good indications that this model do penalize common words.

PRF Models vs Effects	TF	Concave	Doc Length	IDF	Doc Score	DF
Mixture	✓			✓		
DivMin	✓	✓	✓			✓
EDCM	?			✓		✓
Bo Models	✓			✓		
Relevance Model	✓		✓		✓	
Geom Rel. Model	✓	✓	✓		✓	✓
Log-logistic	✓	✓	✓	✓		✓

Table 6.9: PRF Models versus the conditions they verify

6.7 Discussion

We have studied here the main characteristics of PRF reweighting schemes through several constraints reweighting functions should satisfy. There are however a certain number of additional elements that can be used to improve performance of PRF systems. The study presented in [50], for example, proposes a learning approach to determine the value of the parameter mixing the original query with the feedback terms. Interestingly, such a parameter can be set on a query-dependent manner for improved performance. The study presented in [52] focuses on the use of positional and proximity information in the relevance model for PRF, where position and proximity are relative to query terms. Again, this information leads to improved performance. It is not clear yet how one can integrate such an information in the other PRF models we have reviewed, in particular in the LL and SPL models, and this is an aspect one will have to investigate further. Another kind of information that can successfully be exploited in PRF is the one related to query aspects.

The study presented in [25] for example proposes an algorithm to identify query aspects and automatically expand queries in a way such that all aspects are well covered. A similar strategy can be deployed on top of any PRF reweighting function, so as to guarantee a certain aspect coverage in the newly formed query. Another comprehensive, and related, study is the one presented in [23, 29]. In this study, a unified optimization framework is retained for robust PRF. The constraints considered however differ from the constraints we have defined, as they aim at capturing diversity through aspect coverage. The general framework of concave-convex optimization (fully detailed in [22]) is nevertheless interesting and bridges several different models [29]. Lastly, several studies have recently put forward the problem of uncertainty when estimating PRF weights [24, 49]. These studies show that resampling feedback documents is beneficial as it allows a better estimate of the weights of the terms to be considered for feedback. Interestingly, these approaches can be deployed with any PRF reweighting model and allow a simple and neat integration of the DS constraint in any PRF model, as the resampling procedure is based on the score of the document obtained in the first retrieval step.

6.8 Conclusion

This chapter has introduced conditions PRF models should satisfy. These conditions are based on standard IR constraints, with the addition of a *Document Frequency* (DF) constraint which we have experimentally validated. We have partially validated the IDF condition and we have then investigated standard PRF models wrt to these constraints. This theoretical study has revealed several important points.

First, the simple mixture and the divergence minimization models, are deficient as one does not satisfy the DF constraint while the other does not sufficiently enforce the IDF

effect. Second, relevance models and their geometric variant do not satisfy either the IDF condition. Note that smoothing language models do enforce the IDF effect for *ad-hoc* retrieval and we showed here that this is not always the case for PRF. Only two models satisfy all the PRF constraints but the one related to the document score (DS). These models are the log-logistic and the smoothed power law models of the information-based family. Moreover, all these theoretical findings were experimentally illustrated. Overall, this chapter provides an *explanation on why the information-based models perform better than other models in PRF settings.*

Chapter 7

Conclusion

We have studied probabilistic models for word frequencies and for information retrieval. Our goal was to link these probabilitistic models in order to have a good model of word frequencies and a good IR model at the same time.

First of all, we have studied the problem of modeling word frequencies in a collection with a major emphasis on the burstiness phenomenon and state of the art probabilistic model such as the 2-Poisson mixture model, the Negative Binomial, the Beta Binomial and its multivariate extensions DCM and EDCM were reviewed.

Even if the burstiness phenomenon is often mentionned in studies dealing with word frequencies probabilistic models, its precise meaning is vague and is often poorly defined. This is why we returned to the roots of the burstiness characterization as proposed by Katz [45]. We then summarized the significant studies of Church [13, 12], who introduced and validated empirically the notion of adaptation for word frequencies.

Overall, burstiness addresses the fact that for given word, its occurrences in a document, are far from being independent from each other. Even if burstiness has been studied extensively in several studies, each of which proposed a different probabilistic model, our approach differentiate from others by tackling this phenomenon with a *formal* definition of burstiness. This definition translates as a property of probability distributions, related to the log-convexity of the survival function $P(X > x)$. The benefits of this formal definition enable to test whether a distribution can account or not for burstiness and can guide the design of new probability distributions.

This formal definition of burstiness leads us to consider the family of power law distributions as candidates for modeling word frequencies. We introduced two novel models of word frequencies: the *Beta Negative Binomial* [14] distribution, a discrete model, and the *Log-Logistic* distribution [16, 15], a continuous one. The Beta Negative Binomial build on the Negative Binomial proposed by Church [13] as it can be viewed as an infinite mixture of Negative Binomial distributions, which itself is an infinite mixture of Poisson distributions. We then showed how particular instances of the Log-Logistic distributions can be viewed as a continuous counterpart of the Beta Negative Binomial. For both distributions, we provided constant time estimation procedure on the contrary to the DCM and EDCM models, either with a generalized method of moments that can be approximated with the mean document frequency.

The Beta Negative Binomial and Log-Logistic were compared to Poisson distributions and Katz-Mixture on several IR collections. We stressed that the burstiness phenomenon is related to a variance problem. Experiments have validated the BNB and Log-Logistic ability to capture word burstiness which besides suggest that the definition of burstiness we proposed is indeed appropriate. In a nutshell, the Beta Negative Binomial and the Log-Logistic distributions are sound models of word frequencies: they enjoy good theoretical

properties, as bursty distributions, and they fit well word frequencies empirically.

Our second goal was to design IR model compatible with word burstiness as state-of-the-art IR model do not rely on bursty distribution according to our definition. We then have studied probabilistic information retrieval models and review the Probability Ranking framework, the language modeling approach to IR and the Divergence From Randomness models. Although these three probabilistic IR models differentiate from each other, either by a underlying theorical framework or by a distinct choice of a word frequency distribution, all these well performing models share common properties that allow one to describe these models in a single framework. This is the approach advocated by Fang [33] referred to as the axiomatic approach to IR. The main conditions that an IR model should satisfy are the TF, the Concave, the Document Length and IDF conditions. We gave an analytical version of these axiomatic constraints in order to ease their applications to the analysis of IR models.

Among the three main families of probabilistic IR models, the Divergence From Randomness framework seemed to be the best fit to the Beta Negative Binomial and Log-Logistic requirements and this is why we analyzed Divergence From Randomness models with the axiomatic conditions [15, 19]. Above all, this analysis revealed a link between the first normalization principle of DFR model and a particular property of IR models, namely the concavity in word frequencies. Overall, it seems that there is no direct alignment between the concavity of IR models and the burstiness property of the underlying probability distributions. Most state of the art models are concave functions with term frequency but most of the distributions used are not bursty. Therefore, burstiness and the IR model concavity in term frequency seem to be *two sides of the same coin* and this suggested that current paradigms for IR models are not fully compatible with the bursty distributions we wanted to use. This is what seeded the idea of a novel IR family: *information-based models*.

Information models [17, 19, 18] draw their inspiration from a long-standing hypothesis in IR, namely the fact that the difference in the behaviors of a word at the document and collection levels brings information on the significance of the word for the document. Shannon information can be used to capture whenever a word deviates from its average behavior, and we showed how to design IR models based on this information. Above all, these models have a remarkable property: a direct relationship between the burstiness property of the probability distributions used and the concavity of the resulting IR model. In addition, information-based models enjoy good theoretical properties: they satisfy most retrieval heuristic constraints when these models rely on bursty distributions and they can be understood as a simplification of DFR models in the light of retrieval constraints and burstiness.

We then have proposed two effective IR models within this family: *the log-logistic and the smoothed power law models.* The experiments we have conducted on different collections illustrate the good behavior of these models. These models yield state of the art performance, without pseudo relevance feedback, and significantly outperforms state of the art models with pseudo relevance feedback. We have tested these models with different term frequency normalizations and extended them with the beneficial use of the q-logarithm.

Furthermore, the good performance of information models for pseudo relevance feedback lead us to analyze theoretically and empirically several pseudo relevance feedback algorithms [20, 21]. As a result, a list of pseudo feedback constraints was drawn up to better characterize valid pseudo feedback algorithms. In particular, we formulate heuristic constraints for PRF similar to the TF, Concave, Document Length and IDF effects for IR models with an additional constraint refered to as *Document Frequency effect*. We have analyzed the terms chosen by several PRF models, validated experimentally the DF condition and we reviewed several models according to the PRF conditions. The theo-

retical study we have conducted reveals that several standard PRF models either fail to enforce the IDF effect or the DF effect. For example, the simple mixture in the language modelling family is deficient as it does not satisfy the DF constraint and relevance models do not satisfy the IDF condition. On the contrary, the log-logistic and the smoothed power law models satisfy all the PRF properties . This theoretical analysis thus provide an explanation on why the information-based models perform better than other models in PRF settings. All in all, we have proposed:

1. new probabilistic models of word frequencies: the Beta Negative Binomial and the Log-logistic distribution.

2. new probabilistic models for IR: the information based models.

3. a theoretical analysis of PRF models

These new models were analyzed thoroughly and were proved to be theoretically and empirically sound.

Future Work

There are several interesting questions, directions in order to pursue our study on word frequencies and IR models. Are there other properties than burstiness characterizing word frequencies that could be relevant to capture and to use in an IR model ? In others words, as word frequencies data and IR models have distinct and common properties, which are the general properties that should be preserved in an IR setting ? For example, the grouped word frequencies exhibits a Zipfian distribution [4] and the Heap Law [5] characterizes the relation between the size of the vocabulary and the collection lengh. Does the LogLogistic IR model fully account for these phenomenon ? Should these features be taken into account in an IR setting ?

Another question deals with the axiomatic approach to IR. The axiomatic constraints describe general conditions which are not sufficient alone, as it is possible to design an IR model meeting this constraints and which perform poorly empirically. Overall, the three mainstream family of IR models performs similarly and this is why ideally we would like to have a mathematical notion of equivalence between IR models or a notion of capacity of a ranking function. For example, it would be interesting to formalize that the InL2 DFR model performs similarly to the log-logistic model on a given neighborhood of query and document, which would be written as $InL2 \sim_{V(q,d)} LGD$. The axiomatic constraint could be understood as a weak form of equivalence between IR models.

The last direction we will mention would extend our study in a supervised setting with learning to rank methods. Could we infer new axiomatic conditions from training data which could be use more generally for several IR tasks ? Respectively, could/should we learn an IR model compatible with some axiomatic constraints ?

Appendices

A1 Preprocessing

Many applications require a preprocessing step in order to extract the relevant information from the raw content of a document. For most IR scenarios, the preprocessing of documents consists in filtering words too frequent words (stopwords) like articles, and if required standardize the surface form of the observed word (remove conjugations, plurals) and to count for each term its number of occurrences in a document. For example, let's examine the following (famous) French verses from La Fontaine:

Maître Corbeau, sur un arbre perché,
Tenait en son bec un fromage.
Maître Renard, par l'odeur alléché,
Lui tint à peu près ce langage :
Hé ! bonjour, Monsieur du Corbeau.
Que vous êtes joli ! que vous me semblez beau !

Initially, the preprocessing filter stopwords like "ce", "un", ... Then, the occurrences of the words are counted: the term *Corbeau* has a number of occurrences of 2 in this document. In the same way, *fromage* occurs 1 times. One can thus represent a document by a vector whose each dimension contains the frequency of a particular term. For example, a preprocessing of the previous text can lead to the following vectorial representation:

$$\vec{d} = \begin{pmatrix} maitre & 2 \\ corbeau & 2 \\ arbre & 1 \\ perche & 1 \\ tenait & 1 \\ bec & 1 \\ fromage & 1 \\ tint & 1 \\ langage & 1 \\ bonjour & 1 \\ joli & 1 \\ semblez & 1 \\ beau & 1 \\ \dots & \dots \\ cigale & 0 \\ fourmi & 0 \\ \dots & 0 \end{pmatrix}$$

Each dimension in the vector corresponds to a given index term, and the coordinates corresponds here to the number of occurrences of the word in the document. Then, all non-occurring terms have 0 occurences and are not explicitly represented in the previous vector. The frequencies of the *different words are supposed to be statistically independent.* For example, it will be supposed that the random variable for occurrences of *fromage* is independent of the random variable for *Corbeau.* After preprocessing, the corpus of documents can then be represented as a matrix: $\mathbf{X} = (x_{wd})$ where rows stand for words and columns for documents.

A2 Estimation of the 2-Poisson Model

The moment generating function for the 2-Poisson is given by:

$$\begin{aligned} m(t) &= E(e^{tx}) = \alpha \sum_{x=0}^{+\infty} \frac{e^{-\lambda_E} \lambda_E^x}{x!} e^{tx} + (1-\alpha) \sum_{x=0}^{+\infty} \frac{e^{-\lambda_G} {\lambda_G}^x}{x!} e^{tx} \\ &= \alpha e^{-\lambda_E} \sum_{x=0}^{+\infty} \frac{e^{tx} \lambda_E^x}{x!} + (1-\alpha) e^{-\lambda_G} \sum_{x=0}^{+\infty} \frac{e^{tx} {\lambda_G}^x}{x!} \\ &= \alpha e^{\lambda_E (e^t - 1)} + (1-\alpha) e^{\lambda_G (e^t - 1} \end{aligned}$$

The first three derivatives of $m(t)$ are computed and then t is set to 0 which gives the 3 first moments of the distribution:

$$R_1 = \alpha\lambda_E + (1-\alpha)\lambda_G \quad (1)$$

$$R_2 = \alpha(\lambda_E^2 + \lambda_E) + (1-\alpha)(\lambda_G^2 + \lambda_G) \quad (2)$$

$$R_3 = \alpha(\lambda_E^3 + 3\lambda_E^2 + \lambda_E) + (1-\alpha)(\lambda_G^3 + 3\lambda_G^2 + \lambda_G) \quad (3)$$

Let $M = R_1$, $L = R_2$,$K = R_3 + 2R_1 - 3R_2$.

$$M = \alpha\lambda_E + (1-\alpha)\lambda_G \quad (4)$$

$$L = \alpha\lambda_E^2 + (1-\alpha)\lambda_G^2 \quad (5)$$

$$K = \alpha\lambda_E^3 + (1-\alpha)\lambda_G^3 \quad (6)$$

These equations shows that λ_E and λ_G are the roots of :

$$(M^2 - L)\lambda^2 + (K - LM)\lambda + (L^2 - MK) = 0 \quad (7)$$

Finally, α can be estimated with:

$$\alpha = \frac{M - \lambda_G}{\lambda_E * \lambda_G} \quad (8)$$

Publications

Here is the list of my publications where the symbol $\star$ refers to publications prior to this thesis.

International Journal Papers

- S. Clinchant and E. Gaussier. Retrieval constraints and word frequency distributions: a log-logistic model for IR. Information Retrieval, 14(1), 2010.
- $\star$ J. Ah-Pine, M. Bressan, S. Clinchant, G. Csurka, Y. Hoppenot, and J.-M. Renders. Crossing textual and visual content in different application scenarios. Multimedia Tools Appl., 42(1): 31–56, 2009.

Book Chapter

- J. Ah-Pine, S. Clinchant, G. Csurka, F. Perronnin, and JM Renders, Leveraging Image, Text and Cross–media Similarities for Diversity–focused Multimedia Retrieval in ImageCLEF , Experimental Evaluation in Visual Information Retrieval Springer Series: The Information Retrieval, 2010

National Journal Papers

- S. Clinchant and E. Gaussier. Modèle de RI fondés sur l'information. Document Numérique, Juin 2011, to appear

International Conferences

- S. Clinchant and E. Gaussier. Is document frequency important for PRF? In Proceeding of International Conference on the Theory of Information Retrieval, ICTIR 2011, to appear
- S. Clinchant, J. Ah-Pine, and G. Csurka. Semantic combination of textual and visual information in multimedia retrieval. In International Conference on Multimedia Retrieval, 2011.
- S. Clinchant and E. Gaussier. Information-based models for ad hoc IR. In Proceeding of the 33rd international ACM SIGIR conference on Research and development in information retrieval, SIGIR '10, pages 234–241, New York, NY, USA, 2010. ACM. *Nominated for Best Paper award.*
- S. Clinchant and E. Gaussier. Bridging language modeling and divergence from randomness models: A log-logistic model for ir. In In Proceeding of International Conference on the Theory of Information Retrieval, pages 54–65, 2009.

- S. Clinchant and E. Gaussier. The BNB distribution for text modeling. In C. Mac- donald, I. Ounis, V. Plachouras, I. Ruthven, and R. W. White, editors, European Conference in Information Retrieval ECIR, volume 4956 of Lecture Notes in Computer Science, pages 150–161. Springer, 2008.
- ⋆ S. Clinchant, C. Goutte, and E. Gaussier. Lexical entailment for information re- trieval. In ECIR, pages 217–228, 2006.

National Conferences

- S. Clinchant and E.Gaussier . A document frequency constraint for pseudo-relevance feedback models. In CORIA, pages 73-88. Editions Universitaires d'Avignon, 2011
- S. Clinchant and E. Gaussier. Modèles de RI fondé sur l'information. In CORIA, pages 99–114, 2010. *Prix du meilleur article*

Posters

- S. Clinchant and E. Gaussier. Do IR models satisfy the TDC constraint ?, In Proceeding of the 33rd international ACM SIGIR conference on Research and development in information retrieval , SIGIR'11 to appear
- G.Csurka, S. Clinchant , G.Jacquet, Medical Image Modality Classification and Retrieval. In 9th Internatioanl Workshop on Content-Based Multimedia Indexing, 2011
- S. Clinchant and E. Gaussier. Retrieval constraints and word frequency distributions: a log-logistic model for IR. In CIKM, pages 1975–1978, 2009.

Index

Bibliography

[1] Giambattista Amati, Claudio Carpineto, Giovanni Romano, and Fondazione Ugo Bordoni. Fondazione Ugo Bordoni at TREC 2003: robust and web track, 2003.

[2] Gianni Amati and Cornelis Joost Van Rijsbergen. Probabilistic models of information retrieval based on measuring the divergence from randomness. *ACM Trans. Inf. Syst.*, 20(4):357–389, 2002.

[3] Arthur Asuncion, Max Welling, Padhraic Smyth, and Yee Whye Teh. On smoothing and inference for topic models. In *In Proceedings of the 25th Conference on Uncertainty in Artificial Intelligence*, 2009.

[4] R.H. Baayen. *Word Frequency Distributions*. Kluwer Academic, 2001.

[5] Ricardo Baeza-Yates and Berthier Ribeiro-Neto. *Modern Information Retrieval*. Addison-Wesley Publishing Company, USA, 2nd edition, 2008.

[6] A. L. Barabasi and R. Albert. Emergence of scaling in random networks. *Science*, 286(5439):509–512, October 1999.

[7] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

[8] Wray Buntine and Aleks Jakulin. Applying discrete PCA in data analysis. In *AUAI '04: Proceedings of the 20th conference on Uncertainty in artificial intelligence*, pages 59–66, Arlington, Virginia, United States, 2004. AUAI Press.

[9] Chris Burges, Tal Shaked, Erin Renshaw, Matt Deeds, Nicole Hamilton, and Greg Hullender. Learning to rank using gradient descent. In *In ICML*, pages 89–96, 2005.

[10] Deepayan Chakrabarti and Christos Faloutsos. Graph mining: Laws, generators, and algorithms. *ACM Comput. Surv.*, 38(1):2, 2006.

[11] Kenneth Church and William A. Gale. Inverse document frequency (idf): A measure of deviations from Poisson. In *Proceedings of the Third Workshop on Very Large Corpora*, pages 121–130, 1995.

[12] Kenneth W. Church. Empirical estimates of adaptation: the chance of two noriegas is closer to p/2 than p2. In *Proceedings of the 18th conference on Computational linguistics*, pages 180–186, Morristown, NJ, USA, 2000. Association for Computational Linguistics.

[13] Kenneth W. Church and William A. Gale. Poisson mixtures. *Natural Language Engineering*, 1:163–190, 1995.

[14] Stéphane Clinchant and Éric Gaussier. The BNB distribution for text modeling. In Craig Macdonald, Iadh Ounis, Vassilis Plachouras, Ian Ruthven, and Ryen W. White, editors, *ECIR*, volume 4956 of *Lecture Notes in Computer Science*, pages 150–161. Springer, 2008.

[15] Stéphane Clinchant and Éric Gaussier. Bridging language modeling and divergence from randomness models: A Log-Logistic model for IR. In *ICTIR*, pages 54–65, 2009.

[16] Stéphane Clinchant and Éric Gaussier. Retrieval constraints and word frequency distributions: a log-logistic model for ir. In *CIKM*, pages 1975–1978, 2009.

[17] Stéphane Clinchant and Eric Gaussier. Information-based models for *ad hoc* IR. In *Proceeding of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '10, pages 234–241, New York, NY, USA, 2010. ACM.

[18] Stéphane Clinchant and Éric Gaussier. Modèles de RI fondés sur l'information. In *CORIA*, pages 99–114, 2010.

[19] Stéphane Clinchant and Éric Gaussier. Retrieval constraints and word frequency distributions: a log-logistic model for ir. *Information Retrieval*, 14(1), 2010.

[20] Stéphane Clinchant and Éric Gaussier. A document frequency constraint for pseudo-relevance feedback models. In *CORIA*, pages 73–88. Éditions Universitaires d'Avignon, 2011.

[21] Stéphane Clinchant and Éric Gaussier. Is document frequency important for prf? In *ICTIR, to appear*, 2011.

[22] K. Collins-Thompson. Estimating robust query models with convex optimization. In *NIPS*, pages 329–336, 2008.

[23] K. Collins-Thompson. Reducing the risk of query expansion via robust constrained optimization. In *Proceeding of the 18th ACM conference on Information and knowledge management*, CIKM '09, pages 837–846, New York, NY, USA, 2009. ACM.

[24] Kevyn Collins-Thompson and Jamie Callan. Estimation and use of uncertainty in pseudo-relevance feedback. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '07, pages 303–310, New York, NY, USA, 2007. ACM.

[25] D. W. Crabtree, P. Andreae, and X. Gao. Exploiting underrepresented query aspects for automatic query expansion. In *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '07, pages 191–200, New York, NY, USA, 2007. ACM.

[26] Ronan Cummins and Colm O'Riordan. An axiomatic comparison of learned term-weighting schemes in information retrieval: clarifications and extensions. *Artif. Intell. Rev.*, 28:51–68, June 2007.

[27] Scott Deerwester. Improving Information Retrieval with Latent Semantic Indexing. In Christine L. Borgman and Edward Y. H. Pai, editors, *Proceedings of the 51st ASIS Annual Meeting (ASIS '88)*, volume 25, Atlanta, Georgia, October 1988. American Society for Information Science.

[28] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *JOURNAL OF THE ROYAL STATISTICAL SOCIETY, SERIES B*, 39(1):1–38, 1977.

[29] Joshua V. Dillon and Kevyn Collins-Thompson. A unified optimization framework for robust pseudo-relevance feedback algorithms. In *CIKM*, pages 1069–1078, 2010.

[30] C. Durot. Testing convexity or concavity of a cumulated hazard rate. *IEEE Transactions on Reliability*, 57(3):465–473, 2008.

[31] Charles Elkan. Clustering documents with an exponential-family approximation of the dirichlet compound multinomial distribution. In William W. Cohen and Andrew Moore, editors, *ICML*, volume 148 of *ACM International Conference Proceeding Series*, pages 289–296. ACM, 2006.

[32] S.E. Fienberg E.M. Airoldi, W.W. Cohen. Statistical models for frequent terms in text. In *CMU-CLAD Technical Report - http://reports-archive.adm.cs.cmu.edu/cald2005.html*, 2004.

[33] Hui Fang, Tao Tao, and ChengXiang Zhai. A formal study of information retrieval heuristics. In *SIGIR '04: Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, 2004.

[34] Hui Fang and ChengXiang Zhai. Semantic term matching in axiomatic approaches to information retrieval. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '06, pages 115–122, New York, NY, USA, 2006. ACM.

[35] Yoav Freund, Raj Iyer, Robert E. Schapire, and Yoram Singer. An efficient boosting algorithm for combining preferences. *J. Mach. Learn. Res.*, 4:933–969, December 2003.

[36] Michael D. Gordon and Peter Lenk. When is the probability ranking principle suboptimal? *JASIS*, 43(1):1–14, 1992.

[37] S. P. Harter. A probabilistic approach to automatic keyword indexing. *Journal of the American Society for Information Science*, 26, 1975.

[38] Djoerd Hiemstra, Stephen Robertson, and Hugo Zaragoza. Parsimonious language models for information retrieval. In *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '04, pages 178–185, New York, NY, USA, 2004. ACM.

[39] Keiichiro Hoashi, Kazunori Matsumoto, Naomi Inoue, and Kazuo Hashimoto. Query expansion based on predictive algorithms for collaborative filtering. In *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '01, pages 414–415, New York, NY, USA, 2001. ACM.

[40] Thomas Hofmann. Probabilistic latent semantic indexing. In *SIGIR*, pages 50–57. ACM, 1999.

[41] Martin Jansche. Parametric models of linguistic count data. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - Volume 1*, ACL '03, pages 288–295, Stroudsburg, PA, USA, 2003. Association for Computational Linguistics.

[42] Hervé Jégou, Matthijs Douze, and Cordelia Schmid. On the burstiness of visual elements. In *Conference on Computer Vision & Pattern Recognition*, jun 2009.

[43] N. Johnson, A. Kemp, and S. Kotz. *Univariate Discrete Distributions*. John Wiley & Sons, Inc., 1993.

[44] N. L. Johnson and S. Kotz. Distributions in statistics: continuous multivariate distributions. 1972.

[45] Slava M. Katz. Distribution of content words and phrases in text and language modelling. *Nat. Lang. Eng.*, 2(1):15–59, 1996.

[46] John Lafferty and Chengxiang Zhai. Document language models, query models, and risk minimization for information retrieval. In *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '01, pages 111–119, New York, NY, USA, 2001. ACM.

[47] Victor Lavrenko, Martin Choquette, and W. Bruce Croft. Cross-lingual relevance models. In *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '02, pages 175–182, New York, NY, USA, 2002. ACM.

[48] Victor Lavrenko and W. Bruce Croft. Relevance based language models. In *SIGIR '01: Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 120–127, New York, NY, USA, 2001. ACM.

[49] Kyung Soon Lee, W. Bruce Croft, and James Allan. A cluster-based resampling method for pseudo-relevance feedback. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '08, pages 235–242, New York, NY, USA, 2008. ACM.

[50] Yuanhua Lv and ChengXiang Zhai. Adaptive relevance feedback in information retrieval. In *Proceeding of the 18th ACM conference on Information and knowledge management*, CIKM '09, pages 255–264, New York, NY, USA, 2009. ACM.

[51] Yuanhua Lv and ChengXiang Zhai. A comparative study of methods for estimating query language models with pseudo feedback. In *CIKM '09: Proceeding of the 18th ACM conference on Information and knowledge management*, pages 1895–1898, New York, NY, USA, 2009. ACM.

[52] Yuanhua Lv and ChengXiang Zhai. Positional relevance model for pseudo-relevance feedback. In *Proceeding of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '10, pages 579–586, New York, NY, USA, 2010. ACM.

[53] Rasmus Elsborg Madsen, David Kauchak, and Charles Elkan. Modeling word burstiness using the dirichlet distribution. In Luc De Raedt and Stefan Wrobel, editors, *ICML*, volume 119 of *ACM International Conference Proceeding Series*, pages 545–552. ACM, 2005.

[54] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. *Introduction to Information Retrieval*. Cambridge University Press, New York, NY, USA, 2008.

[55] Christopher D. Manning and Hinrich Schütze. *Foundations of statistical natural language processing*. MIT Press, Cambridge, MA, USA, 1999.

[56] Eugene L. Margulis. N-Poisson document modelling. In *SIGIR '92: Proceedings of the 15th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 177–189, New York, NY, USA, 1992. ACM.

[57] A. Mccallum and K. Nigam. A comparison of event models for naive bayes text classification. In *The Fifteenth National Conference on Artificial Intelligence (AAAI*, 1998.

[58] Qiaozhu Mei, Hui Fang, and ChengXiang Zhai. A study of poisson query generation model for information retrieval. In *SIGIR '07: Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 319–326, New York, NY, USA, 2007. ACM.

[59] T. Minka. *Estimating a Dirichlet Distribution*. PhD thesis, Unpublished paper available at http://research.microsoft.com/m̃inka, 2003.

[60] R. Nallapati, T. Minka, and S. Robertson. The smoothed-dirichlet distribution: a new building block for generative models. In *CIIR Technical Report - http://www.cs.cmu.edu/ nmramesh/sd_tc.pdf*, 2006.

[61] Ramesh Nallapati. Discriminative models for information retrieval. In *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '04, pages 64–71, New York, NY, USA, 2004. ACM.

[62] Jan Naudts. The q -exponential family in statistical physics. *Journal of Physics: Conference Series*, 201(1):012003, 2010.

[63] Jian-Yun Nie. *Cross-Language Information Retrieval*. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers, 2010.

[64] Kamal Nigam, Andrew Kachites Mccallum, Sebastian Thrun, and Tom Mitchell. Text classification from labeled and unlabeled documents using em. In *Machine Learning*, pages 103–134, 1999.

[65] S. Kotz N.L. Johnson and A.W. Kemp. *Univariate Discrete Distributions : N.L. Johnson, S. Kotz and A.W. Kemp (1992): 2nd Edition. New York: John Wiley, ISBN 0-471-54897-9*, volume 17. February 1994.

[66] Paul Ogilvie and James P. Callan. Experiments Using the Lemur Toolkit. In *Text REtrieval Conference*, 2001.

[67] Jay M. Ponte and W. Bruce Croft. A language modeling approach to information retrieval. In *SIGIR*, pages 275–281. ACM, 1998.

[68] Jason D. M. Rennie. The log-log term frequency distribution, 2005.

[69] L. Rigouste. *Modéthodes probabilistes pour l'analyse exploratoire de données textuelles*. PhD thesis, Thèse de l'ENST, Télécom Paris, 2006.

[70] C S. Robertson, H. Zaragoza, Stephen Robertson, and Hugo Zaragoza. The probabilistic relevance framework: Bm25 and beyond.

[71] S. E. Robertson. The Probability Ranking Principle in IR. *Journal of Documentation*, 33(4):294–304, 1977.

[72] S. E. Robertson and S. Walker. Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval. In *SIGIR '94: Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 232–241, New York, NY, USA, 1994. Springer-Verlag New York, Inc.

[73] G. Salton and M. J. McGill. *Introduction to Modern Information Retrieval*. McGraw-Hill, Inc., New York, NY, USA, 1983.

[74] Avik Sarkar, Paul H. Garthwaite, and Anne De Roeck. A bayesian mixture model for term re-occurrence and burstiness. In *Proceedings of the Ninth Conference on Computational Natural Language Learning*, CONLL '05, pages 48–55, Stroudsburg, PA, USA, 2005. Association for Computational Linguistics.

[75] Jangwon Seo and W. Bruce Croft. Geometric representations for multiple documents. In *SIGIR '10: Proceeding of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, pages 251–258, New York, NY, USA, 2010. ACM.

[76] Claude E. Shannon. A mathematical theory of communication. *The Bell system technical journal*, 27:379–423, 1948.

[77] Amit Singhal, Chris Buckley, Mandar Mitra, and Ar Mitra. Pivoted document length normalization. pages 21–29. ACM Press, 1996.

[78] Tao Tao and ChengXiang Zhai. Regularized estimation of mixture models for robust pseudo-relevance feedback. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '06, pages 162–169, New York, NY, USA, 2006. ACM.

[79] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. Hierarchical Dirichlet processes. Technical Report 653, Department of Statistics, University of California at Berkeley, 2004.

[80] I. Ounis V. Plachouras, B. He. University of Glasgow at TREC 2004: Experiments in web, robust and terabyte tracks with terrier, 2004.

[81] Jun Xu. A boosting algorithm for information retrieval. In *In Proceedings of SIGIR'07*, 2007.

[82] Zuobing Xu and Ram Akella. A new probabilistic retrieval model based on the dirichlet compound multinomial distribution. In *SIGIR '08: Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 427–434, New York, NY, USA, 2008. ACM.

[83] Yisong Yue and Thomas Finley. A support vector method for optimizing average precision. In *In Proceedings of SIGIR'07*, pages 271–278. ACM, 2007.

[84] Chengxiang Zhai and John Lafferty. Model-based feedback in the language modeling approach to information retrieval. In *CIKM '01: Proceedings of the tenth international conference on Information and knowledge management*, pages 403–410, New York, NY, USA, 2001. ACM.

[85] Chengxiang Zhai and John Lafferty. A study of smoothing methods for language models applied to ad hoc information retrieval. In *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 334–342, New York, NY, USA, 2001. ACM.

[86] Chengxiang Zhai and John Lafferty. A study of smoothing methods for language models applied to information retrieval. *ACM Trans. Inf. Syst.*, 22(2):179–214, 2004.