COMPUTER SCIENCE DEPARTMENT

Relevance-Based Language Models: New Estimations and Applications

PHD THESIS

Javier Parapar López

2013

UNIVERSIDADE DA CORUÑA

PhD Thesis

Relevance-Based Language Models: New Estimations and Applications

Javier Parapar López

PhD supervisor:
Prof. Dr. Álvaro Barreiro García

Thesis committee:

Prof. Dr. Fabio Crestani

Prof. Dr. José Luis Freire Nistal

Dr. Leif Azzopardi

Dr. David E. Losada Carril

Dr. Pablo Castells Azpilicueta

D. Álvaro Barreiro García, Catedrático de Universidad en el área de Ciencias de la Computación e Inteligencia Artificial de la Universidad de A Coruña

CERTIFICA

Que la presente memoria titulada ***Relevance-Based Language Models: New Estimations and Applications*** ha sido realizada bajo su dirección y constituye la Tesis que presenta **D. Javier Parapar López** para optar al grado de Doctor con mención Internacional por la Universidad de A Coruña.

A Coruña, febrero de 2013

Firmado: Dr. Álvaro Barreiro García
Director de la Tesis

Firmado: Javier Parapar López
El doctorando

A Papá, Mamá, Jorge y David

All things are difficult before they are easy
Thomas Fuller

Abstract

Relevance-Based Language Models introduced in the Language Modelling framework the concept of relevance, which is explicit in other retrieval models such as the Probabilistic models. Relevance Models have been mainly used for a specific task within Information Retrieval called Pseudo-Relevance Feedback, a kind of local query expansion technique where relevance is assumed over a top of documents from the initial retrieval and where those documents are used to select expansion terms for the original query and produce a, hopefully more effective, second retrieval.

In this thesis we investigate some new estimations for Relevance Models for both Pseudo-Relevance Feedback and other tasks beyond retrieval, particularly, constrained text clustering and item recommendation in Recommender Systems. We study the benefits of our proposals for those tasks in comparison with existing estimations. This new modellings are able not only to improve the effectiveness of the existing estimations and methods but also to outperform their robustness, a critical factor when dealing with Pseudo-Relevance Feedback methods. These objectives are pursued by different means: promoting divergent terms in the estimation of the Relevance Models, presenting new cluster-based retrieval models, introducing new methods for automatically determine the size of the pseudo-relevant set on a query-basis, and originally producing new modellings under the Relevance-Based Language Modelling framework for the constrained text clustering and the item recommendation problems.

Acknowledgments

This work will be not finished without the help of many. First of all, I must thank my advisor Prof. Álvaro Barreiro for his great supervision and guidance and for letting me to be part of his research group and writing this thesis under his advice. Of course, I also wish to acknowledge the support of every member of the Information Retrieval Lab at University of A Coruña, and particularly Roi, Chema, Edu, Isma, Xose, Martín and Pedro for their support all through these years. I have also to acknowledge the help of a former member, Dr. David Losada, for his help and support during all the stages of this dissertation.

During my period as Ph.D. student I had the pleasure of visiting the University of Lugano, I shall not forget to thank Prof. Fabio Crestani, Dr. Mark Carman and the whole USI IR Group for their warm reception. I also visited Prof. Ricardo Baeza at Yahoo!, again I must thank him and the whole Yahoo! Research Labs Barcelona for their friendly welcome and great support.

I want also acknowledge the input of those reviewers who contributed to improve this and other works. Specially, I must thank the members of this thesis' committee for their effort. I shall acknowledge the financial support of *Xunta de Galicia* and the Government of Spain for my Ph.D. grant, the travel grants, and the funding through research projects.

Finally, I can not forget my friends and family, whose support during my whole life made the path to accomplish the objective pursued with this work easier. I can assure you that *the best is yet to come*.

Index

List of Tables

List of Figures

Chapter 1

Thesis Outline

1.1 Introduction

Since the first formal statements of how the search process should be carried out in the libraries' archives during the fifties, Information Retrieval (IR) techniques have become essential for the daily activity of most of the human beings. Nowadays the homepage of almost every web browser installed in the personal computers point to a web search engine such as Google, Yahoo! or Bing, this is not only for marketing purposes, but also, and more importantly, it is because today the search engines are vital to access information. And those search engines would not be possible without the research efforts made on the Information Retrieval field. Information Retrieval is in fact the *science of searching*, or maybe a better description could be the *science of finding*. Several definitions have being proposed to characterise this, still young, research area, in our opinion one of the most accurate ones was stated in Manning et al. (2008):

> *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)*

The aforementioned search engines are complex information systems with many components such as crawlers, parsers, tokenisers, indexers, searchers, classifiers or interaction interfaces. But over all them, the retrieval models are in the core of any search engine. Retrieval models allow to the search engines to provide the user with effective and efficient documents rankings in answer

to his information need. Several retrieval models have been proposed along the history of Information Retrieval. Particularly, this thesis work is framed under the well-known and highly effective Statistical Language Models. Ponte and Croft (1998) introduced the use of the Language Models (LM) in Information Retrieval, models which closest roots have to be found in the field of automatic speech recognition. Specifically, in this work we will deal with a later expansion for LM called Relevance-Based Language Models (Lavrenko and Croft, 2001). Relevance Models (RM) introduced in the LM framework the concept of relevance, which is explicit in other retrieval models such as the Probabilistic models. RM have been mainly used for a specific task within IR called Pseudo-Relevance Feedback (PRF). PRF is a kind of local query expansion technique where relevance is assumed over a top of documents from the initial retrieval; these documents are used to select expansion terms for the original query and produce a, hopefully more effective, second retrieval.

In this thesis we investigate some new estimations for RM for both PRF and other tasks beyond retrieval, particularly, constrained text clustering and item recommendation in recommender systems. We study the benefits of our proposals for those tasks in comparison with existing estimations. We also play attention to some practical aspects in RM such as parameter tuning, and specially a new proposal for automatically determine the size of the top of documents selected for PRF.

In the remaining of this chapter we discuss the motivation of this Thesis (1.2), the contributions of this work (1.3) and finally a general overview of the work presented in this thesis (1.4)

1.2 Motivation

Producing a good quality document ranking with respect to a user need is still not a closed issue. Although for some kind of queries produced by common users in web search the results are satisfactory in terms of effectiveness, there is still a lot of room for improvement in many retrieval scenarios and other tasks beyond document retrieval. Several techniques have been explored to pursue this objective. Among those techniques, PRF seems to be one of most promising when talking about improvements in effectiveness. It is commonly acknowledged that RM are one of the best performing PRF techniques; despite this, it was only recently when detailed studies about the different es-

timations of RM have been carried out (Lv and Zhai, 2009a). That is, in fact, the main motivation of this thesis; producing new RM estimations for the retrieval task comparing their performance against state-of-the-art baselines. We do not only aim to produce more effective models but also to derive more robust estimations that, at some point, tackle the endemic topic drift problem of the PRF techniques. In line with this we want to further explore mainly two different paths: promotion of divergence and cluster based methods. Referring to the former, the promotion of divergence at term level in PRF has been previously explored in other retrieval frameworks with success (Carpineto et al., 2001), meanwhile in RM the attempts to do this were not conclusive (Zhai and Lafferty, 2001) and only at model level rather than at term level. Regarding cluster based techniques to boost up RM, research efforts have been centred on apply cluster based retrieval approaches to improve the selection of the pseudo-relevant set (Lee et al., 2008) and this is the idea that we want to further explore not only with the aim of improving effectiveness but also robustness.

As result of the very good behaviour of RM in terms of effectiveness for document retrieval, RM have been tested in other retrieval situations such as sentence retrieval (Balasubramanian et al., 2007), passage retrieval (Li and Zhu, 2008) or sentiment retrieval (Eguchi and Lavrenko, 2006). Furthermore, the same concept of relevance models was successfully applied for other non-retrieval tasks such as automatic image annotation (Jeon et al., 2003) or social-tagging (Lavrenko et al., 2002). This inspired us in going a step further and, in the same way that LM were reinterpreted for the retrieval task coming from the field of automatic speech recognition, we want to produce further estimations of the RM for other tasks where the relevance models are not anymore associated between queries and documents. Particularly, we will model the recommendation problem and the constrained clustering tasks, two problems very related with document retrieval in which we have worked before in other research lines. RM are a principled way of doing massive query expansion; we will use RM for other non-query expansions. In the recommendation problem the relevance models will be estimated over the user preferences and user's neighbours preferences, for producing user's profile expansion, meanwhile in the constrained text clustering task the relevance model will be computed over the text document and the documents with which this document shares constraints, resulting in the document expansion of the constrained one.

When addressing this research it is mandatory to tackle some practical problems. Specifically, a good election of the values of the parameters it is known to be crucial in the performance of the different formulations of RM. Some studies have been lately published on this issue (Winaver et al., 2007; Huang et al., 2008), but the efficiency of existing approaches compromises their use in query-time in real applications. Because of this, we decided to study this point in detail, trying to produce meaningful automatic methods for determining the values of some of these parameters in a less computationally expensive way.

1.3 Thesis Statement

The main claim of this work is that it is possible to formulate new estimations of the Relevance Models for further improving the effectiveness of the retrieval task and tasks beyond retrieval. This new modellings will be able to not only improve the effectiveness of the existing estimations and methods but also to outperform their robustness, a critical factor when dealing with PRF methods. These objectives are pursued by different means: promoting divergent terms in the estimation of the RM, presenting new cluster-based retrieval models, introducing new methods for automatically determining the size of the pseudo-relevant set on a query-basis, and producing new modellings under the RM framework for the constrained text clustering and recommending problems.

Aligned with the means stated above, the contributions of this thesis are the following. Three different methods are proposed in order to improve the effectiveness of the retrieval task when using RM. First, a new study of how to introduce the promotion of divergent terms in the estimation of the RM is presented, incorporating in the RM framework this idea intrinsic to other PRF frameworks like the Kullback-Leibler Distance (KLD) based query expansion method under the Rocchio's framework (Carpineto et al., 2001) or the whole Divergence From the Randomness (DFR) retrieval model (Amati and Van Rijsbergen, 2002). Second, a new cluster based approach designed with the aim of selecting better document candidates to form the pseudo-relevant set is presented and compared with alternative ways of constructing the pseudo-relevant set for RM. Third, a query-basis method for determining the number of the top documents selected for expansion is formulated, studying the ef-

fect in terms of effectiveness and the consequences in terms of efficiency with respect to existing methods.

Furthermore, two new modellings away from the document retrieval task are presented. On the one hand, a new modelling of the constrained text clustering task is introduced, where the corresponding estimations of the RM in order to introduce the constraints' information directly in the document representation are formulated. On the other hand, the formulation of the item recommendation problem in the context of a recommender system as an RM estimation is formalised. All these techniques are validated on empirical evidence, through several series of thorough experiments, which prove to be robust across different collections.

1.4 Thesis Outline

The main novel contributions in this thesis are presented in chapters 3, 4, 5, 6 and 7. Chapter 2 contains a general introduction to RM that an IR specialist could skip, but any interested reader may find it as a brief introduction to the state of the art. The organisation of the chapters is as follows [1]:

- Chapter 2 is a brief overview of the main concepts of IR, PRF and above all RM. The different steps of how PRF is traditionally addressed in a search process are clarified and particularly how those are produced in RM. A brief review of the literature in terms of PRF techniques and RM estimations is presented with the aim of providing with a clear view of the existing alternatives to the ones proposed in this work.

- Chapter 3 presents the first of our efforts to produce further RM estimations, in this case based on the promotion of divergent terms when computing the relevance model for a given query. A detailed evaluation against the standard formulation for RM and other non RM based PRF approach is produced comparing both effectiveness and robustness values.

- Chapter 4 introduces our proposal for modifying the selection of the documents belonging to the pseudo-relevant set by applying cluster-based retrieval strategies. Our aim is not only to exploit in this process

[1] Following the recommendations by Evans et al. (2012) we decided to keep self-contained chapters exposing in each one the relevant literature and state-of-the-art study particular to the specific chapter's topic. We also follow recommendations in terms of formatting and style.

the information from *good* but also from *bad clusters*. Evaluation against the results of previous efforts in this line is reported comparing again effectiveness and robustness figures.

- Chapter 5 describes our attempts for providing with an automatic method to determine, at query level, the number of documents selected from the top of the initial retrieval over which it is estimated the query relevance model. Our proposal, which is based on the study of the score distributions, tries to improve the effectiveness but above all the robustness of RM. We studied how our method compares with other previous techniques remarking its advantages.

- Chapter 6 begins with our proposals for modelling new tasks under the RM framework. In this case we approach the constrained text clustering task. In this chapter we present how to accommodate the constraints directly in the document representation, avoiding the use of specially tailored constrained clustering algorithms while achieving comparable and even better clustering effectiveness that those ones.

- Chapter 7 examines another problem beyond document retrieval: item recommendation. We propose an alternative formulation of the collaborative filtering methods based in the RM framework, modelling the item recommendation problem as a profile expansion problem and formulating the corresponding estimations for the RM. Detailed comparison with state-of-the-art recommendation methods shows impressive improvements for the different effectiveness measures.

- Finally, chapter 8 presents the conclusions of the thesis and a summary of the future research lines.

Chapter 2

Introduction to Relevance Models

This chapter is devoted to introduce the basic concepts of Information Retrieval (IR), Pseudo-Relevance Feedback (PRF) and Relevance-Based Language Models (RM). In this chapter we pretend to briefly revisit essential aspects for understanding the remaining of this thesis but mainly to illustrate RM for the unfamiliarised reader.

2.1 Information Retrieval

In the beginnings of the Web, the most common information access approach was by direct browsing. Later on, the web directories appeared facilitating the categorized classification of the web pages for access. But the greatest change in how web pages were accessed was produced by the advent of the web searchers and their use of Information Retrieval (IR) methods. Of course, was not the Web where IR and searchers were born, but it was, in our opinion, the most successful application domain, having web searchers a very important responsibility in the spread of the Web usage.

In this context, it is important to distinguish between data and information. Usually for going from the raw data to the actual information a sense-making process has to be made. This process is commonly bound to a *user* and an *information need*. The objective of IR systems and methods is to simulate the behaviour of a user which has a question to be answered and a collection

of documents where the answer to the question is contained by a subset of them. Ideally the user would read the documents of the collection discarding those not interesting for answering the question and keeping those which contain pieces of information that satisfy his need (Van Rijsbergen, 1979). Documents without an actual information need are merely data, becoming information when they increase the knowledge of a user for a given need commonly expressed in form of *question* or *query*. One important problem very recurring in the literature (Belkin et al., 1982) is the difficulty for the user of expressing his actual information need which results bad query formulations. Belkin refers to this problem as the *Anomalous State of Knowledge* hypothesis and justify this problem because the gap in the user knowledge, gap which the user is trying to fill producing a kind of vicious circle.

Once that we have a user, an information need and a collection of data, we can try to define the subjective concept of *relevance*. A simplistic way of defining relevance is at document level, a document is relevant to a user's information need if it totally or partially satisfies the user need. The most satisfies the need the most relevant the document is. IR systems should preferably rank documents in decreasing order of relevance.

Information Retrieval spans beyond traditional document retrieval. Several tasks in the scope of IR do not depend on the expression of the information need as a query such as clustering or topic detection and tracking; or they do not depend on answering that information need with a set of documents such as question answering or expert search. But, for the sake of simplicity, we will explain the search process considering the traditional document retrieval paradigm or ad-hoc retrieval task.

The objective of retrieval models is to compare the representations of the documents in the collection with the information need representation returning a ranked list of documents ordered by their relevance to the user's information need. Three important steps have to been performed: the computation of the documents' representation (document indexing), the computation of user's information need representation (the query formulation problem) and the comparison of both of them, resulting in the document ranking (the document retrieval).

Document indexing involves several processes which lead to the creation of the inverted file. Indexing is out of the scope of this thesis so we are not going to go in detail about its different aspects. It is an off-line process which comprises steps such as term tokenising, stemming, stop-word removal, index

compression, etc. (Blanco, 2008).

The query formulation problem comprises the process of coding the user's information need from the user's mind to the IR system representation. The most common way in which this task is accomplished is through the use of search boxes where the users introduce natural language queries. These queries are commonly processed in the same way as documents had been treated in order to allow an easy query-document comparison. Natural language words are therefore on-line converted in query terms applying techniques such as stemming or stop-word removal.

Once that the document representation and the query formulation process are cleared, the next step is the matching phase. In this stage, the IR systems must return a rank of documents ordered from most relevant to less relevant according to the user's information need. This objective is achieved by using specific ranking algorithms commonly referred as retrieval models. Retrieval models are formal mathematical models which exploit information, commonly, from the document collection and the user query to provide with an effective document ranking. Several retrieval models were presented along the IR history: the Boolean Retrieval Model (Lancaster and Fayen, 1973), the Vector Space Model (Salton et al., 1975), the Probabilistic Models (Fuhr, 1992), etc. An alternative probabilistic retrieval framework is the Language Modelling Framework to which this thesis is devoted to. Language Models (Ponte and Croft, 1998) were presented with the aim of predicting the likelihood of observing the query terms once observed the probability distributions of the documents in the collection.

Another important intermediate step between the initial query formulation and the final document ranking is the query reformulation phase. In this step, which is not mandatory, the original query formulation is further modified with the aim of improving the retrieval effectiveness. This is the topic were this thesis is framed. When this process is done with real judgements from the user this process is commonly named as *relevance feedback*, otherwise, when assumptions are done in that sense these methods are referred as *Pseudo-Relevance Feedback* (PRF). PRF is a type of blind query expansion (Carpineto and Romano, 2012) where usually a certain top of documents from an initial retrieval with the original formulation of the query are assumed to be relevant for the user's information need. This set of documents is called pseudo-relevant set. PRF methods exploit statistics about the original query, the pseudo-relevant set and the document collection to expand the

original query representation with more terms that are ideally correlated, in terms of relevance, with the original query (Salton and Buckley, 1990).

Among the most successful PRF methods that were presented in the context of the Language Models retrieval framework are the Relevance-Based Language Models (Lavrenko and Croft, 2001). To those models this thesis is devoted and we will explain those in detail in the next sections. So in the remaining of this chapter, we will briefly review the most commonly used retrieval models, particularly explaining the Language Modelling framework (2.2), then we will review existing approaches for Pseudo-Relevance Feedback (2.3) and comment in detail the basis of the Relevance Models (2.4).

2.2 Language Models

Statistical Language Models (Zhai, 2008) were originally introduced in Ponte and Croft (1998). Ponte and Croft proposed a new document scoring strategy coined as *query likelihood* scoring. In this framework, first, a probabilistic model of the documents in the collection is estimated based on its textual contents. Second, each document is scored according to the likelihood of generating the query with its estimated model. The probabilistic distributions estimated over the sequences of words are the so called Language Models (LM).

As previously commented, Language Models had been previously tested for assessing the language usages on other research areas such speech recognition or machine translation where it is very important to infer the likelihood of words to appear. In speech recognition, for instance, the main use of LM is to estimate the likelihood of a word to be said, considering a sequence of previous words already recognised, and so to determine which of the candidate words is chosen as output of the speech recogniser. When translating the use of the LM to the retrieval (document scoring) task, each document plays a similar role to the already recognised sequence of words meanwhile the query plays the role of a candidate word. The ranking is constructed by sorting the documents by the probability of the query to be generated by the corresponding document language model.

Formally, the scoring method is defined as follows. Let q be a query, d a document and Θ_d the language model estimated with the words of document

d. Then the score of document d given the query q is defined as

$$score(q, d) = P(q|\Theta_d) \tag{2.1}$$

Two main problems remain to be determined: how to define the language model Θ_d and how to estimate Θ_d based on the document contents. Depending on the alternatives used to address those two problems different variations of the LM have been proposed. It is not the aim of this chapter to review that precise topic but for defining the LM we can mention the original multiple Bernoulli model (Ponte and Croft, 1998), the multinomial model (Song and Croft, 1999) or the Poisson model (Mei et al., 2007).

The event generation model based on the original multiple Bernoulli (Ponte and Croft, 1998) assumes that the presence or absence of each word is independent of each other. Every word w_j in the lexicon V has a random variable $X_j \in \{0, 1\}$ indicating $X_j = 1$ the presence of the word w_j in the query and $X_j = 0$ its absence. Thus, the multiple Bernoulli language model is estimated as:

$$\Theta_d = \{P(X_j = 1|\Theta_d)\}_{j=1}^{|V|} \tag{2.2}$$

Resulting in the following query likelihood

$$P(q|\Theta_d) = \prod_{w_i \in q} P(X_i = 1|\Theta_d) \prod_{w_j \notin q} (1 - P(X_i = 1|\Theta_d)) \tag{2.3}$$

In this thesis, we will mainly use a multinomial language model. In the multinomial distribution model, also called uni-gram language model, the sequence of words is obtained by generating each word independently having the multinomial distribution the same number of parameters as the size of the lexicon:

$$\Theta_d = \{P(w_j|\Theta_d)\}_{j=1}^{|V|} \;\; s.t. \sum_{j=1}^{|V|} P(w_j|\Theta_d) = 1 \tag{2.4}$$

where $P(w_j|\Theta_d)$ corresponds with the probability of the term w_j according to the estimated distribution and V is the lexicon of the collection. Assuming $q = q_1 \dots q_n$ the query likelihood for a document d would be computed as:

$$P(q|\Theta_d) = \prod_{i=1}^{n} P(q_i|\Theta_d) = \prod_{w \in V} P(w|\Theta_d)^{c(w,q)} \tag{2.5}$$

where $c(w, q)$ is the count of word w in the query q and n the number of terms in the query.

In order to rank the documents in the collection, the probability of a document given a query, $P(d|q)$, is estimated using the Bayes' rule as presented in Eq. 2.6.

$$P(d|q) = \frac{P(q|d) \cdot P(d)}{P(q)} \stackrel{rank}{=} \log P(q|d) + \log P(d) \tag{2.6}$$

In practice $P(q)$ is dropped for document ranking purposes. The prior $P(d)$ encodes a-priori information on documents and the query likelihood, $P(q|d)$ is estimated with Eq. 2.3 when using multiple Bernoulli or 2.5 when considering a multinomial distribution. Usually, $P(d)$ is assumed to be uniform, not affecting to the document ranking, although it also can be used to capture query-independent evidences such as document lengths (Blanco and Barreiro, 2008; Losada and Azzopardi, 2008; Parapar et al., 2009), page linkage in web retrieval (Thijs Westerveld et al., 2002) or other document features.

In every modelling alternative, an important topic is the probability assigned to the unseen query words in the documents. This problem generally affects the performance of different language modelling approaches and is tackled by the *smoothing methods* (Zhai and Lafferty, 2004) such as Jelinek-Mercer, Dirichlet, Absolute Discount or Laplace Smoothing. The smoothing process is the adjustment of the maximum likelihood estimator used in the language model in order to produce more accurate estimations. In this thesis we will mainly work with Dirichlet smoothing as it has been demonstrated together with multinomial modelling as one of the best performing LM alternatives.

In the *Bayesian smoothing* method using Dirichlet Prior (a.k.a. Dirichlet smoothing) (MacKay and Peto, 1994) the language model is a multinomial distribution where the conjugate prior for Bayesian analysis is the Dirichlet distribution with the following parameters: $(\mu P(w_1|\mathcal{C}), \mu P(w_2|\mathcal{C}), \ldots, \mu P(w_{|V|}|\mathcal{C}))$. Thus, the final estimation for the query likelihood using the multinomial model and Dirichlet smoothing is obtained as in Eq. 2.7.

$$P(q|d) = \prod_{i=1}^{n} P(q_i|d) = \prod_{i=1}^{n} \frac{tf(q_i, d) + \mu \cdot P(q_i|\mathcal{C})}{|d| + \mu} \tag{2.7}$$

where n is the number of query terms, $tf(q_i, d)$ is the raw term frequency of

q_i in d, $|d|$ is the document length expressed in number of terms, and μ is a parameter for adjusting the amount of smoothing applied. $P(q_i|\mathcal{C})$ is the probability of the term q_i occurring in the collection $\mathcal{C}$ that is usually obtained with the maximum likelihood estimator computed using the collection of documents.

2.3 Pseudo-Relevance Feedback

Traditionally, retrieval models only considered the user's query as the input against which compare the collection of documents to produce the results ranking. This was a valid strategy in early stage search engines. In the past, the systems were designed to search over library catalogues where the queries were structured, very refined and the user tend to be an expert in the task. Unfortunately, with the popularisation of the search engines the situation changed dramatically. Nowadays, the most successful use of search engines is considered to be the web searchers. In the case of web search, there is no assumption on the expertise of the users neither with respect to the query's topic nor with the search process itself. Furthermore, the average length of user's queries in web search, although it increased lately, is still under three words (Hitwise, 2011; Arampatzis and Kamps, 2008). This short nature of the queries can produce a poor matching between of user's query terms and collections terms, resulting in low retrieval effectiveness.

For dealing with the issues explained above, the most successful attempts were made by producing the expansion of the original query by different means. Automatic Query Expansion (QE) has its roots in early works in the sixties (Maron and Kuhns, 1960). Since then, several ways to approach the expansion of the original query have been explored. We can classify the QE approaches in: linguistic analysis methods, corpus-specific techniques, query-specific techniques, methods exploiting the analysis of search logs or methods exploiting other external resources (Carpineto and Romano, 2012). All of them share some common sub-sequential steps: data acquisition, data preprocessing, candidate feature generation and ranking, feature selection, query reformulation and second retrieval. Depending on how those steps are carried out we will deal with one or other type of QE method.

Nevertheless, there are still some problems affecting QE which impede its generalised use in some application domains. Initially, works on automatic

QE tend to show improvements in recall but affecting the precision figures. Although, later on the improvements were achieved for both precision and recall these improvements were in average across queries, in fact, some queries were greatly improved but some other were negatively affected. This problem, which is commonly known as the robustness problem, is still nowadays a known issue of QE methods. It motivated a lot of research in terms in for which queries QE is appropriated (selective QE) and to what extent (adaptive QE). Also, methods for better selection of the sources in the data acquisition step have been explored.

A lot of factors affect to the effectiveness and robustness of the QE methods, and we will address in this thesis some of them in a particular a QE technique called Pseudo-Relevance Feedback. Pseudo-Relevance Feedback (PRF) is a blind and local method for QE where, without the existence of explicit relevance judgements, a certain set of documents are considered to be relevant for the original query. This set of documents is called the Pseudo-Relevant Set (RS). Using the RS information and the original query the PRF methods compute an expanded version of the original query. So the data acquisition step is based on recollection of the RS and the features involved are simply the terms with highest importance in the RS. Both Relevance Feedback and PRF reinforce the retrieval model original decision over relevance, the difference is that in Relevance Feedback there is an explicit assess by the user over the relevance of the initial ranking documents, meanwhile in PRF that is assumed.

Several different methods have been used for the PRF task. Rocchio's framework for relevance feedback (Rocchio, 1971) was one of the very early successful methods presented in the context of the Vector Space Model. This framework allowed different ways of computing the terms weight needed for carrying out the feature selection step. Different weighting functions were tested: the Binary Independence Model (Robertson and Sparck Jones, 1976), the Chi-square method (Doszkocs, 1978), the Robertson Selection Value (Robertson, 1991), and finally the Kullback-Leibler distance method by Carpineto et al. (2001) which turned out to be one of the best performing. But in line with the main topic of this thesis we have to analyse in detail one of the most successful PRF methods presented to the date in the LM framework: the Relevance-Based Language Models (Lavrenko and Croft, 2001), to them is devoted the next section.

2.4 Relevance Models

Relevance-Based Language Models (Popularly, Relevance Models or RM for short) (Lavrenko and Croft, 2001) are among the best-performing ranking techniques in text retrieval. They were devised with the aim of explicitly introducing the concept of relevance, intrinsic to the probabilistic model of IR, in statistical Language Models. In fact, both LM and Probabilistic models have been directly connected by assuming in the Probabilistic framework that $P(d, q|\mathcal{R} = \bar{r}) = P(d|\mathcal{R} = \bar{r})P(q|\mathcal{R} = \bar{r})$ and $P(d, \mathcal{R}) = P(d)P(\mathcal{R})$ (Lafferty and Zhai, 2002). Relevance Models achieve state-of-the-art performance in terms of effectiveness for the pseudo-relevance feedback task. RM have been established as high-performance PRF approaches showing great improvements over the results obtained with the initial ranking. Since this approach was originally presented by Lavrenko and Croft (2001) it has been used in combination with other methods such as the employment of query variants (Collins-Thompson and Callan, 2007), cluster based retrieval (Lee et al., 2008), passage retrieval (Li and Zhu, 2008) or sentence retrieval (Balasubramanian et al., 2007).

The RM approach builds better query models using the information given by the pseudo-relevant documents. A formal definition of relevance model could be a mechanism that determines the probability $P(w|R)$ of observing a word w in the documents relevant to a particular information need (Lavrenko and Croft, 2001). Given an accurate model of relevance R, if we want to rank a set of documents to be presented to the user according to the Probability Ranking Principle (PRP)(Robertson, 1997) the best rank would be constructed by sorting the documents according to the posterior probability of their belonging to the relevant class R. This is equivalent to rank the documents by the odds of being observed in the relevant class: $P(d|R)/P(d|\bar{R})$. Under the word independence assumption the rank can be computed as:

$$\frac{P(d|R)}{P(d|\bar{R})} \sim \prod_{w \in d} \frac{P(w|R)}{P(w|\bar{R})} \tag{2.8}$$

Only one question remains to be answered, how to learn the relevance model R. This is equivalent to answer the following question: given an unknown process R from which we have sampled every query word $q_1 \dots q_n$ after n (the query length) times, what is the probability that the next word we sample will

be w?.

$$P(w|R) \approx P(w|q_1 \dots q_n) = \frac{P(w, q1 \dots q_n)}{P(q_1 \dots q_n)} \tag{2.9}$$

The objective now is to estimate the joint probability of observing the word w and the query terms together (the numerator of Eq. 2.9). The denominator of Eq.2.9 can be computed as $P(q_1 \dots q_n) = \sum_w P(w, q_1 \dots q_n)$.

Two estimations were originally presented (Lavrenko and Croft, 2001). RM1 assumes that the words in the relevant documents and the query words are sampled identically and independently from the relevance model (i.i.d. sampling). The steps of the derivation can be observed in the original paper and the result is an estimation where the query likelihood for every document is used as the weight for the document and the probability of a word is averaged over every document language model. In contrast, RM2 assumes that the query words are independent of each other, but they are dependent of the words of the relevant documents (conditional sampling). The result is that relevant documents containing query words can be used for computing the association of their words with the query terms. A quite detailed explanation of the RM for PRF is given in the Chapter 7 of the book Croft et al. (2009).

In RM the original query is considered a very short sample of words obtained from the relevance model (R). If more words from R are desired then it is reasonable to choose those words with highest estimated probability when considering the words for the distribution already seen. So the terms in the lexicon of the collection are sorted according to that estimated probability, which after doing the assumptions using the RM1 method, is estimated as in Eq. 2.10.

$$P(w|R) \propto \sum_{d \in \mathcal{C}} P(d) \cdot P(w|d) \cdot \prod_{i=1}^{n} P(q_i|d) \tag{2.10}$$

Usually $P(d)$ is assumed to be uniform. $\prod_{i=1}^{n} P(q_i|d)$ is the query likelihood given the document model, which is traditionally computed using Dirichlet smoothing (see Eq. 2.7). Then for assigning a probability to the terms in the relevance model we have to estimate $P(w|d)$; in order to do so it is also common to use Dirichlet smoothing. The final retrieval is obtained by four steps:

1. Initially the documents in the collection $\mathcal{C}$ are ranked using their query likelihood. This query likelihood is usually estimated with some kind of

smoothing, commonly Dirichlet smoothing as in Eq. 2.7.

2. A certain top r documents from the initial retrieval are taken for the estimation instead of the whole collection $\mathcal{C}$, let us call this pseudo-relevance set RS.

3. The relevance model probabilities $P(w|R)$ are calculated using the estimate presented in Eq. 2.10, with RS instead of $\mathcal{C}$.

4. To build the expanded query the e terms with highest estimated $P(w|R)$ are selected. The expanded query is used to produce a second document ranking using negative cross entropy as in Eq. 2.11. In this second retrieval Dirichlet smoothing is commonly used.

$$\sum_{i=1}^{e} P(w_i|R) \cdot \log P(w_i|d) \tag{2.11}$$

RM3 (Abdul-jaleel et al., 2004) is a later extension of RM that performs better than RM1 in terms of effectiveness. RM3 interpolates the terms selected by RM1 with the original query as in Eq. 2.12 instead of using them directly. The final query is used in the same way as in RM1 to produce a second ranking using negative cross entropy.

$$P(w|q') = (1 - \lambda) \cdot P(w|q) + \lambda \cdot P(w|R) \tag{2.12}$$

As it has been demonstrated (Lv and Zhai, 2009a) as the best performing estimation of RM to the date, we will centre the work in this thesis on RM3. Although for some task we will also considered other estimations and other PRF methods for comparison, we firmly believe that RM3 is a very effective and quite robust starting point.

Chapter 3

Promoting Divergence in Relevance Models

In this chapter we present an alternative estimation of RM promoting terms that being present in the pseudo-relevant set are also distant from the language model of the collection. We compared this approach with the RM3 estimation and with an adaptation to the Language Modelling framework of the Rocchio's KLD-based term ranking function. The evaluation showed that this alternative estimation of RM reports consistently better results than RM3, showing in average to be the most stable across collections in terms of robustness.

3.1 Introduction and Motivation

As previously commented, QE approaches can be classified between global techniques which produce a query rewriting without considering the original rank produced by the query, and local techniques in which the expanded query is generated using the information of the initial retrieval list. In the later set, the PRF methods are framed. In Salton (1971) the initial efforts on exploiting the local information to improve the query formulation were presented. When Rocchio's framework for relevance feedback Rocchio (1971) was suggested, different term ranking functions were proposed. It was in this framework, designed to work under the umbrella of the Vector Space Model, where the idea of the divergence was initially introduced in the relevance

feedback task. Carpineto et al. (2001) presented a term ranking function which is based on computing Kullback Leibler divergence between the collection and RS distributions.

On the other hand, Relevance Models (RM) (Lavrenko and Croft, 2001), have been established as a high-performance PRF approach showing great improvements over the results obtained with the initial ranking. But, despite the success of RM, it was only recently when the researchers tackled the necessity of comparing different estimations for the RM. Lv and Zhai (2009a) compared five methods to estimate the query language models: RM3 and RM4 (Abdul-jaleel et al., 2004); a divergence minimization model (DMM) and a simple mixture model (SMM) (Zhai and Lafferty, 2001); and a regularized mixture model (RMM) (Tao and Zhai, 2006). The main finding of Lv and Zhai was that, in general, RM3 is the best and most stable method among the others. As commented before, RM3 and RM4 Abdul-jaleel et al. (2004) are extensions of the originally formulated RM1 and RM2 approximations, respectively. These extensions linearly interpolate the original query with the terms selected for expansion using RM1 or RM2.

As the introduction of the divergence in the Rocchio's framework resulted in important improvements in effectiveness (Carpineto et al., 2001) we wanted to test how the use of divergence could be incorporate in the LM framework and particularly in the RM framework. For that reason, in this chapter we will devise a new RM estimation which incorporates the idea of divergence taken from other relevance feedback frameworks (Parapar and Barreiro, 2011b). So, the contributions of this chapter are two techniques that promote divergent terms comparing them with the traditional RM3 estimation. First we adapt the discriminational model to score candidate expansion terms in the Rocchio's framework based on the Kullback-Leibler Distance (KLD) to work under the LM framework, improving also the performance of the original method by interpolating the selected expansion terms with the original query as in RM3. In our second contribution we present a new RM estimation that promotes divergent terms for expansion, i.e., terms that are far from the collection language model. We adopted the evaluation methodology from Lv and Zhai (2009a) and the results showed that the new estimated relevance model performs better than RM3 and that its behaviour, in terms of robustness across collections, is more stable than the other methods.

The rest of the chapter is as follows. Section 3.2 presents some background to the methods. Section 3.3 explains the proposed methods for PRF

with promotion of divergence. In Section 3.4 the evaluation and its results are reported. Section 3.5 describes the related work and, finally, conclusions are reported in Section 3.6.

3.2 Background

In this section we will introduce some theoretical basis for this chapter not reviewed in the general background chapter.

3.2.1 Kullback Leibler Divergence for Pseudo Relevance Feedback

Rocchio (1971) introduced the first experiments in query modification combining query expansion with term reweighting. Rocchio's framework was presented over the Vector Space Model (VSM) (Salton et al., 1975; Salton, 1989). In this framework the expanded query (q') is obtained as:

$$q' = q + \frac{1}{r}\sum_{i=1}^{r} R_i - \frac{1}{nr}\sum_{i=1}^{nr} S_i \tag{3.1}$$

where r is the number of relevant documents, nr is the number of non-relevant documents, R_i is the vector representation of the i$^{\text{th}}$ relevant document and S_i is the vector representation of the i$^{\text{th}}$ non-relevant document.

Thus, the original idea was to combine the vectors of the original query with an average of the vectors of the relevant documents and subtracting the average vector of the non-relevant documents. Although initial results where positive they were soon surpassed by other works (Ide, 1971) incorporating other ideas such as the normalization for the number of relevant and non-relevant documents, the use of only relevant documents (positive relevance feedback), and limiting the number of terms which are going to be present in the expanded query.

Furthermore, when using Rocchio's framework for blind relevance feedback, instead of being used as is, some improvements were applied. First, assumption of relevance over a certain top of documents from the initial retrieval is done. Then a score is assigned to each term in the top retrieved documents, score which is computed with a weighting function computed over the whole collection. Although this was a quite direct application it has

an important weakness: the terms are weighted with respect to their usefulness in the whole collection rather than their importance with respect to the query.

With the intention of solving this weakness, studies which compare the differences in the term distribution over the pseudo-relevant set and the whole collection were carried out. Particularly, Carpineto et al. (2001) presented a method for term scoring in the context of Rocchio's framework for PRF. The authors tried to maximize the divergence between the probability distributions of the terms estimated in the pseudo-relevant set (p_{RS}) and the distribution estimated over the whole collection ($p_{\mathcal{C}}$). In order to do so they used the Kullback Leibler Divergence (KLD) calculated as in Eq. 3.2.

$$KLD(p_{RS}, p_{\mathcal{C}}) = \sum_{w \in V} p_{RS}(w) \cdot \log \frac{p_{RS}(w)}{p_{\mathcal{C}}(w)} \tag{3.2}$$

This use of KLD captures the relative entropy between the collection and the RS distributions. To build the expanded query they selected the terms which most contribute to the divergence of both distributions (higher KLD score). In that work they compared the KLD term ranking function with Rocchio's weights, Robertson's Selection Value (Robertson, 1991), Chi-squared and Doszkoc's variant of Chi-squared (Doszkocs, 1978). The results showed that the presented KLD term scoring function performed the best.

This work inspired us in two ways: first we wanted to adapt this KLD scoring method to work under the LM framework, which is known to perform much better than the traditional VSM. Second we wanted to explore the incorporation the divergence from the collection idea into the RM framework. These work lines are reported in the next section.

3.3 Promoting the Divergence in Pseudo Relevance Feedback in the Language Modelling Framework

In this section we describe two approaches presented under the Language Modelling framework to promote divergence in the PRF context.

3.3.1 Kullback Leibler Divergence Based Query Expansion in the Language Modelling Framework

Unfortunately, although the KLD method outperformed the other term ranking methods in the Rocchio's framework, it was not compared with RM. This was mainly motivated because they were published at the same time. One of the objectives of this thesis is to devise new RM estimations which take into account the divergence idea. In order to establish a fair comparison with the state of the art we considered a priority to migrate the KLD idea to a better performing framework such as LM. With this adaptation we could fairly compare RM3, the KLD based approach and our new estimation.

The adaptation of the KLD to the LM framework was quite direct, the KLD scoring function was computed as in Eq. 3.3

$$\begin{aligned} kld_{score}(w) = p_{RS}(w) \cdot \log \frac{p_{RS}(w)}{p_{\mathcal{C}}(w)} \approx \\ \approx \frac{tf(w,RS)}{\sum_{v \in V(RS)} tf(v,RS)} \cdot \log \frac{tf(w,RS) \cdot \sum_{v \in V} tf(v,\mathcal{C})}{\sum_{v \in V(RS)} tf(v,RS) \cdot tf(w,\mathcal{C})} \end{aligned} \tag{3.3}$$

where $tf(w, RS)$ is the raw term frequency of w in the pseudo-relevant set, $V(RS)$ is the lexicon of the pseudo-relevant set RS, V is the lexicon of the collection and $tf(w, \mathcal{C})$ is the raw term frequency of w in the whole collection.

One of the key success factors of this PRF approach was to select only a certain number of terms (e) for expansion. The terms selected are those with the highest scores assigned by the weighting function. Working under the LM framework, it is desirable that those weights could be considered as probability values. To obtain a probability value for each of those e terms selected for expansion we decided to re-normalize the scores obtained with Eq. 3.3 as in Eq. 3.4.

$$KLD(w) = \frac{kld_{score}(w)}{\sum_{i=1}^{e} kld_{score}(w_i)} \tag{3.4}$$

In RM3 it was demonstrated that the interpolation of the original query and the expanded query produces a more effective second retrieval than when using the expansion terms in isolation. So we incorporated this idea in the KLD-based model interpolating the e terms selected as result of the KLD scor-

ing formula with the original query. Therefore, the second retrieval is processed with an interpolated query as presented in Eq. in 3.5:

$$P(w|q') = (1 - \lambda) \cdot P(w|q) + \lambda \cdot KLD(w) \tag{3.5}$$

3.3.2 Relevance Models with Promotion of Divergent Terms

The KLD-based introduction of divergence in the Language Modelling framework presented above was made as a plug-in in the Language Modelling framework. According to the analysis presented in Lv and Zhai (2009a), the advantage in terms of stability of RM3 was attributable to the use of the query likelihood scores in the estimation made by RM1, which is not present in the KLD approach. Thus, to take advantage of this, we present a new RM estimation that promotes divergent terms maintaining the benefits from the original RM estimations, i.e., the use of the query likelihood scores. This new estimation arises naturally when the objective is to select expansion terms that, having high estimated probability in the RS, diverge from the collection distribution, i. e. they are more discriminative terms.

Based on the original RM1 estimation presented in Eq. 2.10, the most straightforward way of introducing such idea is by replacing the $P(w|d)$ by $P(w|d) - P(w|\mathcal{C})$. In this way, those terms whose density is higher in RS than in the collection are promoted, meanwhile those with low density in the RS are demoted.

Another important point in order to reinforce the promotion of divergent terms is how $P(w|d)$ is smoothed. Usually in RM this is done using Dirichlet smoothing choosing as background distribution the collection distribution. In the presented method, we decided to apply the smoothing but instead of using the collection distribution as background distribution we chose to use the distribution in the relevance set. Therefore, the objective is to get for expansion the best terms that describe the documents taking into account both the RS and the divergence from the collection distribution. The computation was performed as in Eq. 3.6.

$$P(w|d) - P(w|\mathcal{C}) \propto \frac{tf(w,d) + \frac{\mu \cdot tf(w,RS)}{\sum_{v \in V(RS)} tf(v,RS)}}{|d| + \mu} - \frac{tf(w,\mathcal{C})}{|\mathcal{C}|_w} \tag{3.6}$$

where $|\mathcal{C}|_w$ is the number of tokens in the collection.

Note that $P(w|d) - P(w|\mathcal{C})$ could provide negative scores for those terms with less estimated probability in the documents of the relevant set than in the whole collection. In order to avoid this, a re-normalization of such subtraction is done, let us call the re-normalized term $P_{\mathcal{C}^-}(w|d)$. With these considerations the final estimation is computed as in Eq. 3.7.

$$P(w|R) \propto \sum_{d \in RS} P(d) \cdot P_{\mathcal{C}^-}(w|d) \cdot \prod_{i=1}^{n} P(q_i|d) \tag{3.7}$$

After this, the second retrieval was performed as in RM3 (interpolating with the original query) as indicated in Section 2.4.

Another way of introducing the divergence idea would be the use of a document prior to promote documents that are far away from the collections' distribution, acting at document level rather than at term level. Nevertheless, no improvements were achieved with our experiments applying that approach.

Now we have to remark an important point that, to the best of our knowledge, was never discussed properly in the context of RM: the different roles of the smoothing parameters in the distinct steps of the process. In RM3 smoothing is applied up to four times (see Section 2.4), and Dirichlet is commonly used in every occasion, so we can distinguish:

1. μ_1, the smoothing parameter in the initial retrieval (Eq. 2.7, step 1).

2. μ_2, the smoothing parameter in $P(w|d)$ (Eq. 2.10, step 3).

3. μ_3, the smoothing parameter in $\prod_{i=1}^{n} P(q_i|d)$ (Eq. 2.10, step 3).

4. μ_4, the smoothing parameter in the second retrieval (Eq. 2.11, step 4).

Usually in the literature all the four parameters are considered to be only one and the parameter is even not trained taking default values as for example in Lv and Zhai (2009a) ($\mu = 1000$). Although this may produce good values, being a very good property of the method, the roles of the different μ parameters are quite different. Meanwhile μ_1 and μ_3 parameters are clearly affecting the same query likelihood and should be kept equal, for the other two parameters this is not so clear. The parameters μ_1 and μ_4 control the smoothing in the document language model when calculating the query likelihood in order to produce a ranking, but the nature of the queries of both retrieval processes is quite different: shorter queries against longer queries.

Nevertheless it is demonstrated in Zhai and Lafferty (2004) that the optimal μ values in both scenarios are quite similar, so we can fix $\mu_1 = \mu_3 = \mu_4$.

On the contrary, the smoothing parameter μ_2 is used to control the smoothing when estimating the probability of the terms under the relevance model in order to select them to do the expansion. Although it is still the language model of a document, here the document is not involved in the computation of a query likelihood, therefore, it can be considered a different parameter. For this reason, it does not seem reasonable a-priori to fix the same values for the μ parameters used for retrieval as for the μ parameter used in the estimation of $P(w|d)$. This intuition was confirmed later in the experimentation, being the trained values quite different for both smoothing parameters. In fact, the optimal values trained in the evaluation process of both parameters in RM3 never matched.

3.4 Experiments and Results

In this section we describe the evaluation methodology, including collection and metric election, and we carefully analyse the results comparing the behaviour of our proposals with respect to the baselines. We will reuse the most part of these experimental conditions in the remaining of the thesis.

3.4.1 Collections

Table 3.1: Collections and topics for training and test used in the document retrieval evaluation

Col.	*# of Docs*	*Topics*	
		Training	Test
AP88-89	164,597	51-100	151-200
WT2G	247,491	401-450	–
TREC-678	528,155	301-350	351-400
WT10G	1,692,096	451-500	501-550

To evaluate the different approaches we chose the same collections used in previous works about RM estimations (Lv and Zhai, 2009a): a subset of the Associated Press collection corresponding to the 1988 and 1989 years (AP88-89), the Small Web Collection WT2G and the disk 4 and 5 from TREC

(TREC-678). Additionally, we decide to use the WT10G collection, which was not used in Lv and Zhai (2009a), to report test values in a web collection. In AP88-89, TREC-678 and WT10G we used training and test evaluation: we performed training for Mean Average Precision in a set of topics and testing over another set. For WT2G we report well-tuned values over the trained topics, as it was done in Lv and Zhai (2009a). Short queries (title only) were used because they are the most suitable to be expanded. All the collections were preprocessed with standard stop-word removal and Porter stemmer, as it has been demonstrated the best performing scenario for this task (Lv and Zhai, 2009a). In Table 3.1 the evaluation settings are summarized.

3.4.2 Compared Methods

We compared four methods:

- **LM**: the baseline Language Modelling retrieval model with Dirichlet smoothing as in Eq. 2.7
- **RM3**: the standard formulation of RM3, as explained in Section 2.4.
- **KLD3**: the KLD based PRF method adapted to the LM framework as detailed in Section 3.3.1.
- **RM3DT**: the proposed formulation of RM with estimations promoting divergent terms as described in Section 3.3.2.

3.4.3 Training and Evaluation

The two basic metrics in IR evaluation are precision and recall. The precision P_r of a ranking produced by a retrieval method at some cut-off point r is the fraction of the top r documents that are relevant to the query. On the other side, the recall R_r of a method at a value r is the proportion of the total number of known relevant documents retrieved at that point. Average Precision (AP) was designed to provide a fair comparison across multiple precision levels and is considered as a standard evaluation metric in IR. AP is defined as the arithmetic mean of the precision at all the levels where a relevant document occurs. When averaging AP across a set of topics the resulting evaluation metric is what it is called Mean Average Precision (MAP). In order to follow with the traditional evaluation procedure for this task and report effectiveness results for MAP.

As discussed before, we performed a training and test strategy, more precisely we performed training for MAP and test for AP88-89, TREC-678 and WT10G meanwhile well-tuned values are reported for WT2G as in Lv and Zhai (2009a).

The parameters tuned were: the smoothing parameter of the initial retrieval μ_1 ($\mu_1 \in \{10, 100, 1000, 2000, 3000, 4000, 5000, 6000\}$), which trained value was also used for μ_3 and μ_4 and which was tuned for LM, KLD3, RM3 and RM3DT. The number of documents in the pseudo-relevant set $r = |RS|$ ($r \in \{5, 10, 25, 50, 75, 100\}$) was tuned for KLD3, RM3 and RM3DT. The number of terms selected for expansion e ($e \in \{5, 10, 25, 50, 75, 100\}$) was tuned for KLD3, RM3 and RM3DT. The interpolation weight λ ($\lambda \in \{0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1\}$) was tuned for KLD3, RM3, RM3DT. The smoothing parameter μ_2 ($\mu_2 \in \{10, 100, 1000, 2000, 3000, 4000, 5000, 6000\}$) was tuned for RM3 and RM3DT.

Finally, test values are reported for MAP and Robustness Index (RI) over the initial retrieval (LM). One of the aims of this thesis is to devise more robust estimations of Relevance Models. In order to assess the robustness of the methods we decided to use this RI measure specially designed to evaluate the behaviour of PRF methods. The Robustness Index ($-1 \leq RI(Q) \leq 1$), also called Reliability of Improvement Index, of a model with respect to a baseline was formulated in Sakai et al. (2005) as in Eq. 3.8:

$$RI(Q) = \frac{n_+ - n_-}{|Q|} \tag{3.8}$$

where Q is the set of queries over the RI has to be calculated, n_+ is the number of improved queries, n_- the number of degraded queries and $|Q|$ the total number of queries in Q.

For statistically analyse the differences between methods' effectiveness we used Wilcoxon Signed Rank Test (Wackerly et al., 2008). The null hypothesis of the Wilcoxon signed rank is that two methods share the same distribution. The Wilcoxon test statistic takes the paired score differences and ranks them in ascending order by the absolute value. The sign of each difference is given to its rank as a label (*negative* or *positive*). For a two-sided test, the minimum of the sums of the two sets of ranks is the test statistic. With the statistic value a table can be consulted to determine the $p-value$. When the size of the sample is greater than 25 the table can be skipped and a normal approximation to the distribution exists.

Table 3.2: Values for Mean Average Precision (MAP) on the test topics. Statistical significant improvements (Wilcoxon $p < 0.1$, and Wilcoxon $p < 0.05$ underlined) with respect to LM, RM3, KLD3, and RM3DT are superscripted with l, r, k, and d respectively. Best values are bolded.

	MAP			
Col.	*LM*	*RM3*	*KLD3*	*RM3DT*
AP88-89	.2775	.3606$^{\underline{L}}$ (+30%)	**.3667**$^{\underline{l}}$ (+32%)	.3625$^{\underline{L}}$ (+31%)
WT2G	.3115	.3445^{lk} (+10%)	.3352$^{\underline{l}}$ (+07%)	**.3467**^{lk} (+11%)
TREC-678	.1915	.2295$^{\underline{l}}$ (+20%)	.2293$^{\underline{l}}$ (+20%)	**.2412**$^{\underline{lrk}}$(+26%)
WT10G	.2182	.2468$^{\underline{l}}$ (+13%)	.2238 (+02%)	**.2478**$^{\underline{lrk}}$(+13%)

Table 3.3: Values for Robustness Index (RI) with respect to the LM baseline model for every collection. Best values are bolded.

	RI		
Col.	*RM3*	*KLD3*	*RM3DT*
AP88-89	.38	**.56**	**.56**
WT2G	**.44**	.38	.40
TREC-678	.04	**.47**	.21
WT10G	.28	-.04	**.36**

3.4.4 Results

Analysing the MAP values for the test topics (see Table 3.2) it has to be noted that the three PRF methods always outperform the baseline LM as expected. The adaptation of the KLD method to the LM framework using query interpolation performs quite well, obtaining improvements up to the 32% in the AP88-89; this is a very interesting point considering that KLD3 has fewer parameters to tune. Nevertheless the other PRF methods achieve statistically significant improvements over the KLD3 in four occasions.

The RM3 method performs also quite well in terms of effectiveness with great improvements over the baseline as expected, as it is the state-of-the art in PRF. RM3 performs better than KLD3 in three collections, achieving in one case statistically significance. In the AP88-89 collection the differences across the three PRF methods are negligible, not being ever statistically significant.

The proposed RM3DT estimation achieves statistically significant improvements over the KLD3 method in three occasions and over the RM3 in two, being always better than the later in terms of MAP. Another important point to analyse is the robustness of the methods, and how this is maintained across collections. Considering the values presented in Table 3.3 we can conclude that the RI numbers of the KLD3 method are quite acceptable and similar

across collections, except in the WT10G collection. RM3 values are still acceptable (always bigger than zero) but are considerable lower than the other methods in the AP88-89 and TREC-678 collections. Contrarily RM3 performs slightly better than the other methods in the WT2G collection. This fact may be explained because the values on the WT2G collection are well-tuned, suggesting that a good parameter setting affects more to the robustness of the RM3 method than to the other ones. Comparing both RM methods RM3DT seems to be more stable in terms of RI across collections.

The differences in robustness between RM3 and RM3DT can be analysed observing the queries penalized by RM3 and improved by RM3DT. Let us take as example the query `Parkinson's disease`, for this query LM obtained an average precision of 0.3231, RM3 damaged the query to 0.2927, while RM3DT improved it to 0.5083. Observing the top 25 expansion terms selected in both approaches we can view that many good terms are selected by both methods (for example `patient`, `brain` or `alzheimer`) but the RM3 method introduces terms that are so common that, although being very present in the RS, they introduce a lot of noise in the retrieval such as `page`, `can`, `year`, `will`, `new`, `say`, `may` or `home`, meanwhile those terms are not present in the top 25 RM3DT expansion terms because they were penalized for being so common in the collection.

3.5 Related Work

Zhai and Lafferty (2001) explored the divergence idea proposing a Divergence Minimization Model (DMM). The DMM approach tries to minimize the divergence between the query model and the model of the feedback documents. The DMM objective is to build a feedback model that is close to every pseudo-relevant document language model and far away from the collection language model, which is assumed as the non-relevance model. This was stated as an optimization problem. The DMM approach was already compared in Lv and Zhai (2009a) with Relevance Models showing that DMM performs worse than RM3. Li (2008) proposed a new robust relevance model which combines three different aspects: common word discounting, non-uniform document priors and the modification of the traditional pseudo feedback paradigm by considering the original query as a pseudo feedback document rather than combining it with the expanded query. With the in-

troduction of three additional parameters in the model, for adjusting these new aspects in the estimation, evaluation showed some improvements over RM3. Particularly the method seems to be more robust to the variation in the number of feedback documents.

This thesis is framed under the Language Modelling framework but it is necessary to say that the idea of using divergence to improve the retrieval performance has been already deeply studied under other retrieval models, to the point of existing whole models based on it. The Divergence From the Randomness (DFR) model (Amati and Van Rijsbergen, 2002) is based on a similar idea: the more the terms occurrences in the documents diverge from their expected occurrences considering a random distribution the more information carried by the terms. In the DFR model the QE process is done based on a generalization of the Rocchio's framework (He and Ounis, 2007). Different weighting schemes, including the aforementioned KLD, were tested being the Bose-Einstein Bo1 model the best in terms of effectiveness, which also select those terms that diverge most from the randomness, using for those estimations the collections' statistics. In another paper (Ye et al., 2010) the Rocchio's classical feedback method was integrated in the DFR framework for PRF.

In other IR tasks such as adaptive filtering this divergence idea has also been used. Lv and Zhai (2009b) presented different discriminative features for queries and documents to be used in a technique which learns for each query the interpolation weight of the original query with the expansion terms. Particularly the entropy of the feedback documents and the document clarity are used. With the entropy of the feedback documents basically they capture at term level how heterogeneous is the term distribution in the RS. With the clarity of the feedback documents they try to "explain away" common terms present in the RS.

3.6 Conclusions

In this chapter we have presented two different methods for PRF based on the idea of promoting the divergent terms in the RS. KLD3 is an adaptation to the LM framework of a KLD based method including the linear interpolation with the original query. RM3DT is a new estimation for the RM that computes the probability of a term given a feedback document by the subtracting to

the terms' probability in the document its probability in the collection and applying the smoothing over the RS.

Also, it was analysed the role of the different smoothing parameters involved in the RM methods, showing the different roles that those smoothing parameters play. We compared the new methods with the LM baseline and the RM3 estimation. Particularly the RM3DT performed, for MAP, better than RM3 in every collection, showing, as the KLD3 method, a very good stability across collections in terms of robustness.

Chapter 4

Cluster Based Retrieval and Relevance Models

In the last years, cluster based retrieval has been demonstrated as an effective tool for both interactive retrieval and pseudo-relevance feedback techniques. In this chapter we propose a new cluster based retrieval function which uses the best and worst clusters of a document in the cluster ranking, to improve the pseudo-relevant set. In this way we approach one vital point for RM and in general for every PRF method which is the RS construction. The evaluation shows improvements in some standard TREC collections over classical non cluster-aided RM and other cluster-based baseline in both effectiveness and robustness.

4.1 Introduction and Motivation

Several strategies were studied in the history of the Information Retrieval in order to improve the retrieval models effectiveness. One technique that has been demonstrated successful is relevance feedback as previously analysed in this thesis. On the other hand, clustering has been considered as a useful tool in the retrieval process since the formulation of the cluster hypothesis in 1979 (Van Rijsbergen, 1979). This hypothesis states that very related documents tend to be relevant to the same query. Since then, clustering has been used in IR for tasks such as distributed retrieval (Xu and Croft, 1999), results presentation (Zamir and Etzioni, 1998), document browsing (Cutting et al., 1992),

novelty detection (Fernandez et al., 2010), etc. Indeed, several experiments (Hearst and Pedersen, 1996; Tombros et al., 2002) have demonstrated that clustering algorithms working at pseudo feedback time can obtain clusters with a high percentage of relevant documents, still the automatic identification of these clusters between the whole set of them is still a challenge.

Although initial experiments using query specific clustering (Liu and Croft, 2004) in order to improve the retrieval effectiveness were not conclusive, after improving the cluster representation (Liu and Croft, 2008) and with the use of clustering algorithms that support overlapping (Kurland and Lee, 2004), finally the quality of the initial ranking was significantly improved with cluster based re-ranking (Liu and Croft, 2008; Kurland, 2009).

It was only recently when a cluster based retrieval approach was used to improve the quality of the pseudo-relevant set, for further use in query expansion methods (Lee et al., 2008). This approach takes advantage of the better initial ranking produced by the cluster based retrieval to select a better pseudo-relevant set, improving in this way the effectiveness, sometimes significantly. But, although this kind of methods tend to improve the effectiveness in average, one known problem of them is the lack of robustness, i.e., still a significant amount of queries are negatively affected by them. One of the main factors of this behaviour is the presence of non-relevant documents in the feedback set.

In this chapter we aim to accomplish one of the thesis objectives which is to produce more robust RM methods. With this objective, we present a new cluster based retrieval method that exploits bad clusters in order to reduce the amount of non-relevant documents in the feedback set. We consider not just if a document is present inside a "good" cluster to update its score, but also the presence of the document in "bad" (low relevance score) clusters. As far as we know the information referring bad clusters has not been exploited yet in the context of pseudo-relevance feedback.

We tested our approach in several TREC Collections and compared with the Language Modelling retrieval approach (Zhai and Lafferty, 2004), traditional RM3 formulation (Abdul-jaleel et al., 2004) and with the Resampling method presented by Lee et al. (2008). The evaluation shows that the results in terms of MAP are better than the Resampling approach; furthermore, our method consistently improves the robustness values in all the collections.

The rest of the chapter is presented as follows. Section 4.2 presents our proposal explaining the different steps of the model. Section 4.3 explains the

evaluation methodology and comments the results. In Section 4.4 we describe the related work and finally conclusions are reported in Section 4.5.

4.2 Cluster Based Relevance Modelling

In order to get a better pseudo-relevant set we formulated a new cluster based re-ranking function. Every re-ranking approach has a set of common-steps: obtaining of the initial retrieval, selection of the set of documents to re-rank, the re-ranking process itself and, optionally, further processing based on the re-ranked results.

For the initial retrieval and aligned with the topic of this thesis we chose the high performance LM framework. More precisely, we performed the initial retrieval assuming a multinomial model with Dirichlet smoothing (see Eq. 2.7). The second step is to fix the top of documents subject to be re-ranked (d_{init}). We will refer to the size of this top hereinafter as the parameter N. The next phase in a cluster based technique is to perform the document clustering. For our proposal, we chose a clustering algorithm with overlapping. Once the top documents are clustered, we calculate the cluster query likelihood for every resulting cluster. Finally, the clusters query likelihoods and the documents query likelihoods are combined by our proposed retrieval formula and the documents are reranked according to the new scores. And lastly, the top documents of the new ranking are used as PRS to feed a query expansion process based on RM.

For the discussion of how to perform the initial retrieval we refer to the reader to Chapter 2. Next, we will address those issues of our proposal not commented before, namely: the clustering algorithm, the estimation of the cluster query-likelihood, the re-ranking process and the application of RM.

4.2.1 Clustering Algorithm

Once that the initial ranking is obtained, clustering is performed over the top N documents. The use of clustering algorithms with overlapping has already been demonstrated successful (Kurland and Lee, 2004) in cluster based retrieval. Indeed, initial approaches to query specific clustering (Liu and Croft, 2004) were not conclusive and it was only after incorporating clustering algorithms with overlapping (Liu and Croft, 2008) when the results were improved. As we explained before, one of the main points of our method is

to use the information provided by bad clusters to avoid non-relevant documents in the pseudo-relevant set. In order to do this we used a clustering algorithm that supports overlapping, i.e. one document can belong to one or more clusters.

The straightforward selection based in previous works could be using a k-nearest neighbours (k-NN) algorithm, but the k-NN forces each document to have exactly k neighbours. This aspect is not desired in our approach because we will exploit that a document belongs to a low scored cluster. If we had used k-NN, a non-relevant document with low query likelihood and no close neighbours could attract other documents that, although they are not close to that document, they are the closest ones.

So we decided to cluster the documents in base to a given threshold t, grouping for each document those neighbours that are more similar than t. Let's call this way of grouping thr-NN. Given a document d_i, its neighbourhood is the set of documents d_j such that $sim(d_i, d_j) \geq t$. The purpose of this algorithm is that non-relevant documents could be isolated in singletons (Lu et al., 1996).

Term Frequency-Inverse Document Frequency ($tf \cdot idf$) was used as document representation in the clustering algorithm. $tf \cdot idf$ measures the importance of a term to describe a document not only based on the number of times that it appears in the document, but also to the number of documents in which it appears. A term which appears very rarely in the collection should be given more weight for describing a document, as it is very specific. So the weight of the term w_i in the document d_j was computed as in Eq. 4.1:

$$weight(w_i, d_j) = tf(w_i, d_j) \cdot \log \frac{|\mathcal{C}|}{df(w_i)} \tag{4.1}$$

where $tf(w_i, d_j)$ is the raw term frequency of the term w_i in the document d_j, $|\mathcal{C}|$ is the number of documents in the collection and $df(w_i)$ the document frequency of the term w_i.

For the similarity measure between documents ($sim(d_i, d_k)$) we choose traditional cosine distance as in Eq. 4.2

$$sim(d_i, d_j) = \frac{\sum_{k=1}^{|V|}(weight(w_k, d_i) \cdot weight(w_k, d_j))}{\sqrt{\sum_{k=1}^{|V|} weight(w_k, d_i)^2}\sqrt{\sum_{k=1}^{|V|} weight(w_k, d_j)^2}} \tag{4.2}$$

where $|V|$ is the size of the lexicon of the collection, i.e., the number of dif-

ferent terms in the collection.

4.2.2 Cluster Query Likelihood

In order to exploit the cluster information in our retrieval approach we need a way of estimating the cluster query likelihood. In the origin, the first approaches to cluster retrieval considered the clusters as meta-documents, i.e. one cluster is represented as the concatenation of the documents that belong to it (Kurland and Lee, 2004; Liu and Croft, 2004; Lee et al., 2008), or the centroid of the cluster (Voorhees, 1985). But these representations suffer from several problems because of the document and cluster sizes. As demonstrated in Liu and Croft (2008), the geometric mean is a better cluster representation in order to calculate the cluster query likelihood, so it was chosen in our approach. The cluster query likelihood based on the geometric mean representation was calculated combining equations 4.3 and 4.4.

$$P(q|C) = \prod_{i=1}^{n} P(q_i|C) \tag{4.3}$$

$$P(w|C) = \prod_{i=1}^{|C|} P(w|d_i)^{\frac{1}{|C|}} \tag{4.4}$$

where n is the number of query terms, $|C|$ is the number of documents in the cluster C, and $P(w|d_i)$ was computed using a Dirichlet estimate. So finally the cluster query likelihood applying logarithmic identities can be calculated as in Eq. 4.5

$$P(q|C) \stackrel{rank}{=} \prod_{i=1}^{n} e^{\frac{\sum_{i=1}^{|C|} \log P(w|d_i)}{|C|}} \tag{4.5}$$

4.2.3 Cluster Based Reranking

Previous approaches to cluster based re-ranking only used the presence of a document in a good cluster as indicator of its relevance. As previous explained these approaches when used to construct pseudo-relevant sets for further processing with query expansion, suffer from the problem that even the good clusters are not one hundred per cent composed of relevant documents. The inclusion of non-relevant documents in the relevance set can

produce a poor performance of the query expansion process resulting in effectiveness degradation for that query.

The final objective of our approach is to reduce the number of non-relevant documents in the pseudo-relevant set. To achieve that point we decided to use the information given by the bad clusters. Our hypothesis is that given two documents d_1 and d_2, and being C_{1max}, C_{1min}, C_{2max} and C_{2min} the clusters with best and worst query likelihood to which d_1 and d_2 respectively belong to, if $P(q|C_{1max}) = P(q|C_{2max})$ and $P(q|d_1) = P(q|d_2)$ then if $P(q|C_{1min}) > P(q|C_{2min})$ should indicate that d_1 is likely to be more relevant than d_2. In other words if a document belongs to low clusters in the cluster ranking, it should be a pseudo negative indicator about its relevance.

So in order to produce a document ranking we decided to combine the document query likelihood, with the pseudo positive information in terms of best cluster, and the negative in terms of the worst cluster to which the document belongs. The query likelihood combination is presented in Eq. 4.6.

$$P'(q|d) = P(q|d) \times \max_{d \in C_i} P(q|C_i) \times \min_{d \in C_i} P(q|C_i) \tag{4.6}$$

where $P(q|d)$ was estimated as in Eq. 2.7 and $P(q|C_i)$ was estimated as in Eq. 4.5. This estimation alleviates to some point the problem of previous approaches that leave the cluster reranking as is, trusting in the relevance of every document inside of high ranked clusters.

Ideally removing all the non-relevant documents from the relevant set would have a great impact in order to get better expanded queries and, as a consequence, to improve the final retrieval effectiveness. Even although some relevant documents could be penalised because they group with other ones which appear low in the ranking, this effect will be extensively compensated by the benefit of removing the non-relevant documents from the relevance set.

Once that we compute the cluster-based reranking of the top N documents in the initial retrieval, we can use this altered ranking to feed traditional pseudo-relevance feedback methods. In this case, we will test RM3 (as explained in Chapter 2) in consonance with the objective of this thesis.

4.3 Experiments and Results

The evaluation of our approach was performed over four TREC collections comparing with a baseline retrieval model, a baseline feedback model and a baseline cluster based feedback model.

4.3.1 Settings and Methodology

Referring to the collections and training and test query-sets, we followed the same experimental settings as in Chapter 3. For evaluation datasets we choose a subset of the Associated Press collection corresponding to the 1988 and 1989 years (AP88-89), the Small Web Collection WT2G, the disk 4 and 5 from TREC (TREC-678) and the WT10G collection. In AP88-89, TREC-678 and WT10G we used training and test evaluation (see Table 3.1) meanwhile for the WT2G well-tuned values are reported. Short queries (title only) were used because they are the most suitable to be expanded. All the collections were preprocessed with standard stop-word removal and Porter stemmer.

4.3.2 Compared Methods

We compared four methods:

- **LM**: the baseline Language Modelling retrieval model with Dirichlet smoothing as in Eq. 2.7. This approach was also used by the other methods for producing the initial retrieval.

- **RM3**: the standard formulation of RM3, as explained in Section 2.4. This model was also used by the cluster based pseudo-relevance feedback methods in the PRF phase.

- **Resampling**: the cluster based resampling method presented by Lee et al. (2008), which is the existing cluster based pseudo-relevance feedback baseline. A brief description of this method is presented in Section 4.4.

- **CBRM3**: the proposed cluster based pseudo feedback described in Section 4.2.

Table 4.1: Values for Mean Average Precision (MAP) on the test topics. Statistical significant improvements (Wilcoxon $p < 0.1$, and Wilcoxon $p < 0.05$ underlined) with respect to LM, RM3, Resampling, and CBRM3 are superscripted with l, r, s, and c respectively. Best values are bolded.

	MAP			
Col.	*LM*	*RM3*	*Resampling*	*CBRM3*
AP88-89	.2775	.3408$^{\underline{l}}$ (+22%)	.3581$^{\underline{l}}$ (+29%)	**.3626**$^{\underline{lr}}$ (+31%)
WT2G	.3115	.3376$^{\underline{l}}$ (+08%)	.3285$^{\underline{l}}$ (+05%)	**.3457**$^{\underline{ls}}$ (+12%)
TREC-678	.1915	.2194$^{\underline{l}}$ (+15%)	.2190$^{\underline{l}}$ (+15%)	**.2220**$^{\underline{ls}}$ (+16%)
WT10G	.2182	.2402^{l} (+10%)	.2316^{l} (+06%)	**.2450**$^{\underline{ls}}$ (+12%)

4.3.3 Training and Evaluation

As commented we performed a training and test strategy for MAP. There are several parameters to train. Namely, the smoothing parameter μ was tuned for LM, RM3, Resampling and CBRM3 ($\mu \in \{10, 100, 1000, 2000, 3000, 4000, 5000, 6000\}$). The parameters $r = |RS|$, the size of the pseudo-relevant set, e, the number of expansion terms, and λ, the interpolation factor, for the pseudo feedback based query expansion were trained in the RM3, Resampling and CBRM3 methods ($r \in \{5, 10, 25, 50, 75, 100\}$, $e \in \{5, 10, 25, 50, 75, 100\}$ and $\lambda \in \{0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1\}$). For both cluster-based approaches, N, the size of the d_{init}, was set to 100. The Resampling method needs, apart from μ, r, e and λ, the parameter k (which is the number of nearest neighbours in the k-NN method) that was set to 5 according to the settings described in Lee et al. (2008). Furthermore, and according to the authors' description, a similarity threshold for the k-NN clustering method was set to 0.25, avoiding in this way that documents with less than that cosine similarity score were grouped together. Four our method, CBRM3, t that was set to 0.10.

We have to remark that the effectiveness values of the PRF methods can be further improved by producing a separated training of the different μ parameters as we did in the previous chapter. Nevertheless, it was not the objective of this chapter to demonstrate the different roles played by the smoothing parameters as it has been already clarified before. So for keeping in reasonable values the training computational costs we decided to follow the traditional approach and only considering one μ parameter.

Table 4.2: Values for Robustness Index (RI) with respect to the LM baseline model for every collection. Best values are bolded.

	RI		
Col.	*RM3*	*Resampling*	*CBRM3*
AP88-89	.23	**.40**	**.40**
WT2G	**.35**	.10	.20
TREC-678	.22	.06	**.31**
WT10G	.12	.16	**.18**

4.3.4 Results

Analysing the MAP values for the test topics (see Table 4.1) it has to be notice that our approach significantly outperforms with $p - value < 0.05$ the LM baseline for every collection, a fact that neither the RM3 nor the Resampling method achieve. Our method also achieves statistically significant improvements either over Resampling or over RM3 for all the collections. The values of the Resampling method did not achieve statistical significant improvements over RM3; moreover, the Resampling method achieves worse MAP values than RM3 in two of the collections. This finding is partially contradictory with the results reported in Lee et al. (2008). Our explanation to this fact is that meanwhile the effectiveness values of the Resampling method are similar or higher in our experiments and the original paper, our RM3 implementation reports much higher MAP values (for instance AP88-89: 0.3408 vs. 0.2803 or WT10G: 0.2402 vs. 0.1966) being much more complicated to beat it.

We have seen that our method is robust across collections. Furthermore, we analysed query robustness values measured with RI over the LM baseline model in every collection, values are reported in Table 4.2. Our method achieves the best RI values of the three query expansion approaches but in the WT2G collection where RM3 is the best method. This last fact also was showed in the previous chapter and it may be explained because the values on the WT2G collection are well-tuned, suggesting that a good parameter setting affects more to the robustness of the RM3 method than to the other ones.

4.4 Related Work

Despite since the formulation of the cluster hypothesis (Van Rijsbergen, 1979) several works tried to explode clustering information to improve information retrieval tasks. It was only recently when conclusive results were presented

improving retrieval effectiveness using query specific clustering. We have to cite the work of Kurland and Domshlak (2008) where several features were aggregated to obtain better high precision in the re-ranking of top documents. Kurland and Domshlak used several features related with cluster information, namely *query faithfulness* that measures the similarity of a cluster with the query, *self faithfulness* that estimates how a cluster used as a query model ranks its own documents, the *initial list faithfulness* than apprises how a ranking induced by a cluster reflects the initial rank induced by the query, and the *peer faithfulness* that quantifies how a cluster is similar to its *peer* clusters. The results using the different features separately seem indicate that *peer faithfulness* it is the better indicator, although the aggregation of all the features reports the best values. An approach based on similar facts is presented in Kurland (2009), in this paper the author present several approaches (aspect models and interpolation models) that also combine information about peer clusters. In this case Kurland estimates the relevance of a document by combining its probability of belonging to a cluster with the cluster query likelihood for every cluster or for every cluster that the document belongs to depending on the approach. High precision is again improved, although the performance is quite dependent of the settings and MAP values are also reported but only in a cut-off of 50 documents.

In Kurland (2008) several cluster based re-ranking approaches were presented, exploiting in this case clusters with high percentage of relevant documents. The objective is to automatically find those clusters and explode them to improve the initial ranking, but still the problem of filtering non-relevant documents from them is not addressed. Again several features are combined resulting in improvements in high precision figures over the initial ranking.

Recently Lee et al. (2008) proposed query specific clustering in order to improve the quality of the pseudo-relevant set used in the query expansion process, in this case also Relevance Models (Lavrenko and Croft, 2001). This method uses as cluster re-ranking method over the initial retrieval, particularly the original cluster query likelihood presented in Liu and Croft (2004) is used but with overlapping clusters (k-NN). The results show significant improvements over the initial LM based rank and the RM rank in several collections. This paper also combines the idea of dominant documents in order to construct the pseudo-relevant set. A document which is repeated in several good clusters should be considered a dominant document and so repeated also in the feedback set. This approach that we used in order to compare

our method still shows some problems with query robustness that we tried to solve with our alternative approach using pseudo negative information. It is also important to remark that in this work Lee et al. use document concatenation for representing the clusters. In our method we decided to adopt the geometric mean representation (Liu and Croft, 2008) as it has been already demonstrated as a more effective representation for cluster based retrieval.

The use of bad clusters can be considered as a kind of negative information. The exploitation of negative information to improve retrieval effectiveness is a challenging task. In particular, it is very important for difficult queries, where there are not relevant documents present in the ranking and so it is not possible to use positive information for relevance feedback. Indeed negative relevance feedback has been revisited recently with effective results. Wang et al. (2007) proposed different methods for re-ranking using a set of known non-relevant documents which acts as the negative feedback set. The methods were tested on specifically modified TREC collections (only difficult queries were used from the TREC settings). Results in those collections show improvements over the initial LM-based ranking and traditional pseudo feedback methods. In Wang et al. (2008d) different strategies for negative feedback for both LM and vector space retrieval models are compared. The best results were achieved when working with LM and a multi-negative queries strategy, as the authors state *"This shows that irrelevant documents may distract in different ways and do not form a coherent cluster"*. This last remark seems to indicate the need of not only selecting good clusters but also to filter from them the non-relevant documents.

Related with the need of considering negative information, in this case associated with clustering processes, we have to remark the analysis already presented in Lu et al. (1996). In this paper it is analysed the fact that after running a cluster algorithm over the top documents of a rank, most of the singletons (clusters with only one document) are non-relevant documents, and should be removed. This analysis suggested us that the clustering algorithms should allow the creation of singletons. Although not every singleton shall contain a non-relevant document, allowing the creation of singletons, the real non-relevant documents will not be promoted in the ranking benefited of the fact that they are clustered with relevant ones. Meanwhile, the truly relevant documents will not be affected negatively.

Also recently several works approached the task of getting a better pseudo-relevant set, in this case to increase the diversity, but none of them show

conclusive results. Collins-Thompson and Callan (2007) presented a sampling approach over the top documents based on query variants. The query variants are simply obtained by removing terms from the query. The improvements achieved seem to be produced by the use of query variants. Although they got improvements in query robustness and high precision the method does not outperform the baseline Relevance Model in terms of MAP. The objective of having a less redundant pseudo-relevant set is also approached in Sakai et al. (2005). In this work the authors introduced a resampling method which is based on clustering. The top documents are clustered based on the common query terms selecting only some documents of each cluster in order to improve diversity in the relevance set. But again the results presented in the evaluation were not conclusive.

4.5 Conclusions

The cluster based pseudo relevance feedback method presented in this chapter introduces the use of bad clusters in order to achieve pseudo feedback sets with less non-relevant documents. The pseudo negative information is obtained from the belonging of the documents to a "bad" cluster in a cluster re-ranking approach. The results show improvements in MAP over the existing cluster based approaches for pseudo relevance feedback, that in some settings are statistical significant. Another good result is the improvement in terms of query robustness: our approach penalise less queries than previous ones.

Our method was developed with the objective of improving the estimation of the Relevance Models by refining the composition of the RS. Nevertheless, this method can also be applied to other PRF methods different from RM, for instance in Parapar and Barreiro (2011a) this approach was applied to the KLD-based query expansion process.

Chapter 5

Estimating the Size of the Pseudo-Relevance Feedback Set

It is known that one of the factors that more affect to the PRF robustness is the selection for some queries of harmful expansion terms. In order to minimise this effect in the PRF methods a crucial point is to reduce the number of non-relevant documents in the relevant set. In the previous chapter, this problem was tackled by the use of cluster reranking methods; an alternative approach to this problem is presented in this chapter. We try to automatically determine for each query how many documents we should select as RS. For achieving this objective we will study the score distributions of the initial retrieval and trying to discern between relevant and non-relevant documents.

5.1 Introduction and Motivation

One crucial aspect of the pseudo-relevance feedback methods is robustness. In this context, robustness is defined as the quality of not hurting the effectiveness values achieved by the retrieval model in the initial rank for every query. Most of existing pseudo-relevance feedback methods outperform the effectiveness of the initial retrieval in average but they tend to harm some of the queries. This is an important point for solving in order to popularise

the use of this methods in the commercial search engines. The most common phenomenon causing the decrease of effectiveness for a query is the *topic drift*. Topic drift refers to the situation where the expansion of the query produced that the topic of the original user need has moved (drifted) away to a different one. For instance, for the TREC topic 101: *Design of the "Star Wars" Anti-missile Defense System*, a very clear example of topic drift would be the returning of documents about the film. The topic drift can be naturally produced by the addition of terms, but this problem can be greatly intensified when the RS has plenty of irrelevant documents

This problem has been exposed very early in the literature (Mitra et al., 1998) and caused lots of works on areas such as query performance prediction (Cronen-Townsend et al., 2002; Carmel et al., 2006) which investigates how to predict the performance of a query anticipating those queries that will be negatively affected by the expansion, selective pseudo-relevance feedback(Sakai and Robertson, 2001; Amati et al., 2004) which tries to decide for which queries PRF should or not be applied, and adaptive pseudo-relevance feedback (Lv and Zhai, 2009b) that is centred on adjust the weight of the expansion terms over the original query automatically depending on the nature of the given query.

The different approaches to decide when of how much apply PRF have considered pre-query processing indicators and initial ranking examination. Several evidences have been considered such as the number of query terms in the pseudo-relevant documents, the similarity between query and the relevance set, term proximity measures, etc. But it was only recently when some works started to consider the scores of the initial retrieval (Shtok et al., 2009). Shtok et al. argue that query-drift can potentially be estimated by measuring the diversity (e.g., standard deviation) of the retrieval scores of the documents in the ranking.

In this chapter we also exploit the scoring information but in a different way, we use the scores of the initial retrieval for determining the pseudo-relevant set itself, trying to minimise the amount of non-relevant documents in it. For achieving this objective we used a framework for modelling the score distributions of a retrieval model (Manmatha et al., 2001) and adapt the threshold optimization solution for recall-oriented retrieval (Arampatzis et al., 2009) for our particular problem, where we want to stop selecting documents from the top of the initial retrieval when non-relevant documents appear. Score distributions research investigates the idea of using the docu-

ments' scores for separating relevant and non-relevant documents. For doing this, different statistical modelling choices over both groups of documents are taken and the parameters of the statistical distributions are inferred from the observed scores. Although it has been already used for other task such as meta-search and high recall oriented task such as legal retrieval, this is a novel and especially adequate use of the score distributions analysis. We are really pursuing a high precision for our task in such a way that ideally if no relevant documents are present on the top of the initial retrieval we want to return an empty RS producing a way of selective PRF. Furthermore, and not less important, our approach reduces the number of parameters to tune in the training phase of PRF methods by suppressing the necessity of tune r, the number of documents on the RS.

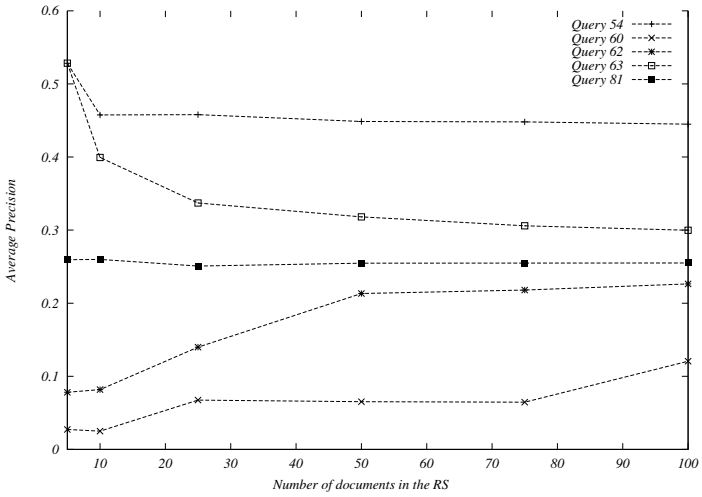

Figure 5.1: RM3 behaviour in terms of Average Precision for different queries from the training query set of the AP-8889 collection with $t = 100$ and $\lambda = 0.8$ and $\mu = 1000$

Although, when averaged over a query set the differences in performance in terms of Average Precision when selecting different top sizes for RS in a particular collection may not differ too much for RM3, it varies a lot at query level (see Figure 5.1). Meanwhile some queries present a stable behaviour (as query 81), most of them have either an increasing behaviour (as queries 60 and 62) or a decreasing behaviour (as queries 54 and 63). Thus, it is clear that it is import to be able to automatically adjust the RS at query level, which

motivates the work in this chapter.

We performed evaluation to assess how our proposal affects to the effectiveness and robustness of RM on standard settings. Results showed that both characteristics are improved with the extra advantage of the reduction of the number of parameters involved in the training phase. The rest of the chapter is as follow: next Section (5.2) starts with some specific background on score distributions, in Section 5.3 we present our proposal for modelling the score distributions and automatically limit the RS size, Section 5.4 shows the evaluation results, related work is briefly reviewed in Section 5.5 and finally we conclude with our main findings in Section 5.6.

5.2 Background

In this section we introduce some theoretical basis for this chapter not reviewed in the general background chapter.

5.2.1 Score Distributions

The Probability Ranking Principle (PRP, Robertson (1997)) states that the ranking of the documents should be according to their probability of relevance. However, retrieval models, in the ideal case where the document ranking strictly honours the PRP, do not provide with a method for delimiting when the non-relevant documents start to appear. In this context, score distributions have been studied and modelled since the early days of IR. Initial works date from the sixties (Swets, 1963), when the idea of using the scores for separating relevant and non-relevant documents was originally formulated. However, it was only recently when the benefit of these approaches was demonstrated for the retrieval task (Manmatha et al., 2001). Score distribution modelling techniques try to infer statistical properties from the seen data (the scores of the ranking documents) and take advantages of such inferred properties, and not directly from the observed data, for classifying documents between relevant and not relevant.

Score distribution models generally assume that the scores of the relevant documents were generated by a different distribution from the distribution of the non-relevant documents. The research efforts have been centred on two aspects: which family of statistical distributions corresponds with each group of documents and how the parameters of the distributions can be learned or

estimated from the observed documents' scores. Different combinations of

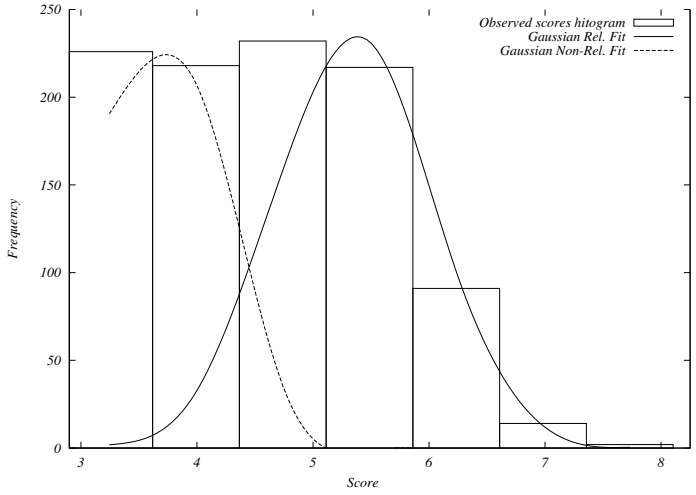

Figure 5.2: Mixture of Gaussians fit to relevant and non-relevant data obtained processing the scores of TREC query 154 over the AP88-89 collection produced with the LM retrieval function with Dirichlet smoothing ($\mu = 1000$)

statistical distributions were proposed for modelling the score distributions. Swets (1963) originally proposed to model the relevant and non-relevant groups as two Gaussian distribution with different parameter values (see Fig. 5.2 as an example), although later on, Sweets considered two negative exponential distributions (Swets, 1969). Bookstein (1977) tested with two Poisson and Baumgarten (1999) with two Gamma. It was only lately when the mixture model of a Gaussian distribution for the relevant and a negative exponential distribution for the non-relevant documents was proposed (Arampatzis et al., 2000). Also recently, Kanoulas et al. (2009) proposed a mixture of Gaussian distributions for relevant documents and a Gamma for non-relevant documents.

In this context, Robertson (2007) presented the convexity hypothesis which stated that for all good systems, the recall-fallout curve (when viewed from the top left (0,1), see Fig. 5.3) is convex. In this case, recall should be interpreted as the proportion of the relevant distribution exceeding a given threshold t and fallout the proportion of the non-relevant distribution exceeding that point. In the graph, the point (0,0) corresponds with a very high threshold that is (nothing retrieved), while the point (1,0) corresponds with

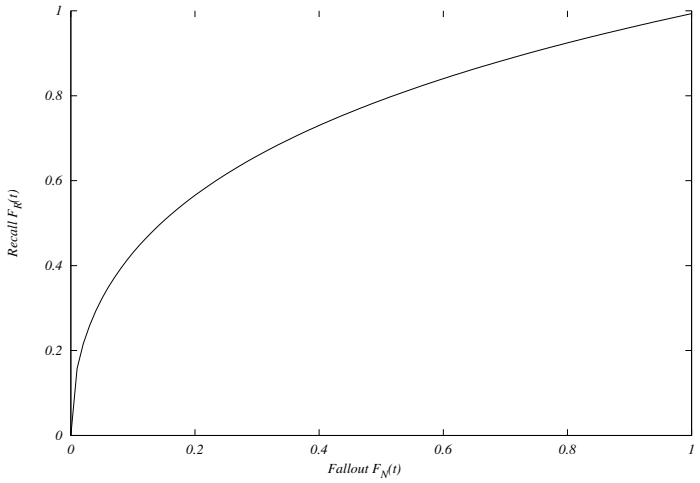

Figure 5.3: Example of idealised Receiver Operating Characteristic (ROC) for a cut-off or threshold t

a very low threshold (everything retrieved). So, if this graph presents concave parts it means that the proportion of the relevant distribution over the non-relevant decreases when the scores increase for some segment of values. This is related, but somewhat stronger than, the inverse recall-precision relationship and it means that the higher the score of a document the higher the probability of relevance. Over the graph, a random ordering of the collection of documents (identical relevant and non-relevant score distributions) would produce a straight line from (0,0) to (1,1). Any other straight segment may also be interpreted of random ordering of sub-sets of the documents. We can easily improve the performance eliminating the concavity segments of the curves by simply randomising the sub-list of scores corresponding with those segments and thus, replacing the concavity parts by straight segments. Indeed, we can just reversing the scores in the sub-list and converting the concavity segments in their convex mirror reflections. In this way, if we depart from a convex curve, we can easily improve the initial performance of our model, so convexity seems to be a desirable property.

In this work, Robertson probes that although the most of the previously presented distributions choice honour the convexity principle, some of them, do not. In particular the Gaussian-negative exponential mixture model (Manmatha et al., 2001; Arampatzis et al., 2009), one of the most popular choices,

does not accomplish this property. In particular, this model presents concavity problems both in the top right end (low threshold values) and the bottom left end (high threshold values) for any parameters' values.

The model presented in Arampatzis et al. (2009), besides not honouring the recall-fallout curve convexity (about the 60% of the queries in the experiments suffer from this anomaly), presents good practical results for a high recall retrieval task such as legal retrieval. One of the most popular effectiveness measures on legal retrieval is the $F_1@K$ where K is the cut-off selected by the system to stop providing with results. The objective pursued with score distributions is to automatically determine the value of K for each query. So for achieving that objective Arampatzis et al. presented a threshold optimisation method over the learned distributions which we adapted for our problem in the next section.

Most of the existing works on score distributions use relevance information and so the learning of the different distributions' parameters is an easy task (the groups of relevant and non-relevant documents are already defined). When there are no relevance judgements, the learning of the distributions' properties from the observed scores also includes the learning of the weights of the mixture. The Expectation Maximisation (EM) algorithm (Dempster et al., 1977) has been the standard approach to finding the mixing and the distribution's properties in this area. Recently, extended versions of this method have been developed for this specific task (Dai et al., 2012). EM is an iterative algorithm which is used for finding maximum likelihood estimates of the parameters in probabilistic models, when dependency exists on unobserved hidden variables

5.3 Modelling Score Distributions for Pseudo-Relevance Feedback

Our objective is the use of score distributions models to automatically determine the size of the pseudo-relevant set, i.e., we want to select for each query the optimal top of documents which will feed the PRF process. Ideally these top documents will be only relevant ones. We formulate this problem as a threshold optimisation task. In order to adapt the score distribution models to work under this paradigm we have to (i) select an appropriate distribution modelling choice, (ii) select a learning strategy for inferring the distributions'

parameters and (iii) formulate the corresponding cut-off conditions.

Referring to the first decision, the straightforward choice should be to use the popular Negative exponential-Gaussian mixture (Arampatzis et al., 2000) or its truncated version (Arampatzis et al., 2009). However, as stated before, this model clearly violates the convexity hypothesis (Robertson, 2007). Moreover, our experiments using these models showed results consistently worse than with our final choice. The model which resulted to perform better than those alternatives was the Gaussian-Gaussian mixture (Swets, 1963) which honours the convexity hypothesis for fixed variances and for almost every situation of different variances (it only presents anomalies in the ends of the intervals). Particularly, we chose to use the later because it presented more robust results across collections, although the former presents greater improvements in some collections.

Regarding to the second point, EM is an efficient and popularly used method to estimate model parameters from a set of observed values by maximising the likelihood. In this case, we decided to use a generalisation of the EM algorithm known a Bregman soft clustering (Banerjee et al., 2005). Bregman soft clustering allows estimating the parameters of a mixture of exponential families (Garcia and Nielsen, 2010), given a set of observations. This Bregman soft clustering algorithm shares with the EM the initialisation, expectation and maximisation steps. The main advantage of using this method instead of the EM algorithm is that it allows to estimate the parameters of *any* mixture of exponential family distributions. The Statistical Exponential Family (Nielsen and Garcia, 2009) is a set of probability distributions admitting the following canonical decomposition:

$$P(x, \Theta) = \exp(\langle t(x), \Theta \rangle - F(\Theta) + k(x)) \tag{5.1}$$

where

- $t(x)$ is the sufficient statistic, a function of the data that fully summarizes the data.
- Θ are the natural parameters,
- $\langle ., . \rangle$ is the inner product,
- $F(.)$ is called the log-normalizer because it is the logarithm of a normalization factor,

- $k(x)$ the carrier measure.

In particular, this family includes the following well-known distributions: Gaussian, Poisson, Bernoulli, binomial, multinomial, Laplacian, Gamma, Beta, negative exponential, Wishart, Dirichlet, Rayleigh, probability simplex, negative binomial, Weibull, von Mises, Pareto distributions, skew logistic, etc. In our case, we use a mixture of Gaussian distributions, in this case the mapping for the canonical decomposition is:

- $t(x) = (x, x^2)$
- $\Theta = (\frac{\mu}{\sigma^2}, -\frac{1}{2\sigma^2})$
- $F(\Theta) = -\frac{\Theta_1^2}{4\Theta_2} + \frac{1}{2}\log(-\frac{\pi}{\Theta_2})$
- $k(x) = 0$.

where μ is, in this case, the mean of the Gaussian distribution and σ its standard deviation. More details of the canonical decomposition can be found in Nielsen and Garcia (2009).

For estimating the parameters of a mixture of exponential families with Bregman soft clustering over the observed scores a general expectation-maximisation procedure is used. As result of this process, the natural parameters of the distributions involved in the mixture are obtained as well as the weights of the distributions in the mixture. In our case, those natural parameters correspond with the means and variances of the Gaussian distribution. Details of the initialisation, expectation and maximisation steps of the process are reported in section 1.5.4 of Nielsen and Garcia (2009). In the initialisation step, the scores are grouped in so many clusters as distributions in the mixture with the K-Means algorithm estimating the weight for each component as the proportion of scores in each cluster. The initial values for the parameters of each distribution are estimated in the corresponding clusters. In the expectation step the probabilities of the observed scores of belonging to each distribution are recomputed. Finally, the maximisation step recomputes the values the parameters of the probability distributions given the new belonging probabilities of the observed scores.

The only remaining aspect to be defined is the cut-off strategy. Arampatzis et al. (2000) states this threshold optimisation problem as follows. The

following definitions are given:

$$\begin{aligned} R &= nG_n \\ R_+(s) &= R(1 - F(s|1)) \\ N_+(s) &= (n - R)(1 - F(s|0)) \\ R_-(s) &= R - R_+(s) \\ N_-(s) &= (n - R) - N_+(s) \end{aligned} \tag{5.2}$$

where R is the number of relevant documents for the query, $R_+(s)$ and $R_-(s)$ the number of relevant documents over and below the given score respectively, $N_+(s)$ and $N_-(s)$ the number of non-relevant documents over and below the given score respectively, G_n is the fraction of relevant documents in the collection, n is the number of documents in the collection and $F(s|1)$ and $F(s|0)$ are values of the cumulative distribution functions at the score s for the relevant and non-relevant distributions respectively.

Then, the optimal score where to perform the cut-off (s_{opt}) is that one such maximise a given effectiveness measure M of the form of a linear combination of the document count of the categories defined in Eq. 5.2:

$$s_{opt} = \arg\max_s \{M(R_+(s), N_+(s), R_-(s), N_-(s))\} \tag{5.3}$$

In our case, we ideally want to obtain a RS for RM where every document is relevant. This is a quite strict condition and for many queries the apparition of a non-relevant document as the highest scored document would produce an empty RS, discarding a lot of useful information. For this reason we decided to relax this constraint and formulate the effectiveness measure for cut-off problem as:

$$M(R_+(s), N_+(s), R_-(s), N_-(s)) = \frac{R_+(s)}{N_+(s)} \tag{5.4}$$

That is, we will cut the top for building the RS in the point of maximum relevance density.

This is our approach to automatically estimate the size of the pseudo-relevance feedback set for RM. Only some final estimation details remain to be explained. As commented before, we chose to model the relevant and non-relevant distributions as a mixture of two Gaussian distributions. From the two Gaussian distributions, learned with the Bregman soft clustering method, the one corresponding with the relevant documents will be assumed that one

with highest mean. The G_n and n parameters will be replaced by their estimated values, corresponding with the fraction of relevant documents in the top and the size of the top, respectively. The fraction of relevant documents in the top will be estimated as the weight in the mixture of the Gaussian distribution corresponding with the relevant documents in the mixture.

5.4 Experiments and Results

The evaluation of our approach was performed over four TREC collections comparing with a baseline retrieval model and the baseline feedback model (training the size of the pseudo-relevant set)

5.4.1 Settings and Methodology

Regarding to the collections and training and test query-sets, we followed the same experimental settings as in Chapter 3 (Section 3.4, Table 3.1). For evaluation datasets we choose a subset of the Associated Press collection corresponding to the 1988 and 1989 years (AP88-89), the Small Web Collection WT2G, the disk 4 and 5 from TREC (TREC-678) and the WT10G collection. In AP88-89, TREC-678 and WT10G we used training and test evaluation (see Table 3.1) meanwhile for the WT2G well-tuned values are reported. Short queries (title only) were used because they are the most suitable to be expanded. All the collections were preprocessed with standard stop-word removal and Porter stemmer.

5.4.2 Compared Methods

We compared four methods:

- **LM**: the baseline Language Modelling retrieval model with Dirichlet smoothing as in Eq. 2.7. This approach was also used by the other methods for producing the initial retrieval.
- **RM3**: the standard formulation of RM3, as explained in Section 2.4 training the size of the RS.
- **SDRM3**: the standard formulation of RM3 but automatically determining for each query the size of the RS as described in Section 5.3

Table 5.1: Values for Mean Average Precision (MAP) on the test topics. Statistical significant improvements (Wilcoxon $p < 0.1$, and Wilcoxon $p < 0.05$ underlined) with respect to LM, RM3, and SDRM3 are superscripted with l, r, and d respectively. Best values are bolded.

	MAP		
Col.	*LM*	*RM3*	*SDRM3*
AP88-89	.2775	.3408$^{\underline{l}}$ (+22%)	**.3794**$^{\underline{lr}}$ (+37%)
WT2G	.3115	**.3376**$^{\underline{l}}$ (+08%)	.3345$^{\underline{l}}$ (+08%)
TREC-678	.1915	.2194$^{\underline{l}}$ (+15%)	**.2245**$^{\underline{l}}$ (+17%)
WT10G	.2182	**.2402**^{l} (+10%)	.2322^{l} (+6%)

5.4.3 Training and Evaluation

As commented we performed a training and test strategy for MAP. There are several parameters to train. Namely, the smoothing parameter μ was tuned for LM, RM3 and SDRM3 ($\mu \in \{10, 100, 1000, 2000, 3000, 4000, 5000, 6000\}$). The parameter e, the number of expansion terms, and λ, the interpolation factor, for the pseudo feedback based query expansion were trained in the RM3, and SDRM3 methods ($e \in \{5, 10, 25, 50, 75, 100\}$ and $\lambda \in \{0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1\}$). Furthermore, for RM3 the parameter $r = |RS|$, the size of the pseudo-relevant set, was also trained ($r \in \{5, 10, 25, 50, 75, 100\}$).

We have to remark that the effectiveness values of the PRF methods can be further improved by producing a separated training of the different μ parameters as we did in Chapter 3, i.e. considering different values for the parameters playing different smoothing roles. Nevertheless, it was not the objective of this chapter to demonstrate the different roles played by the smoothing parameters as it has been already clarified before. So for keeping in reasonable values the training computational costs we decided to follow the traditional approach and we only considered one μ parameter.

5.4.4 Results

The first comment is that, as expected both RM3 and SDRM3 outperform the initial retrieval with statistical significant differences. Analysing the MAP values for the query expansion methods for the test topics (see Table 5.1) the best values are obtained by our proposal in two collections and by traditional RM3 in the other two. However, it has to be notice that the differences in favour of RM3 are never statistically significant and in one case the improvements occur with optimal trained values (WT2G). Meanwhile, our method achieves

Table 5.2: Values for Robustness Index (RI) with respect to the LM baseline model for every collection. Best values are bolded.

	RI	
Col.	*RM3*	*SDRM3*
AP88-89	.23	**.70**
WT2G	.35	**.36**
TREC-678	.22	**.29**
WT10G	**.12**	**.12**

statistical significant improvements in the AP88-89 collection surpassing RM3 in more than 11%.

An interesting fact, is that our proposal seems to be more adequate on the text collection meanwhile it is not able to outperform traditional RM3 (in terms of MAP) in the web collections. This can be partially explained by the fact that the fitting of the chosen score distribution model (mixture of two Gaussian distributions) over the web documents is not as good as it is in textual documents. The retrieval model will produce more separated scores for relevant and non-relevant documents if the documents are more focused and shorter, where the risk of spurious signals of relevance is less

Another important point to analyse is the robustness of the methods, and how this is maintained across collections. Considering the values presented in Table 5.2 we can see that our method obtains the best values in terms of RI in every collection. Again the differences between the RI values of RM3 and SDRM3 are higher in the text collections than in the web collections. In fact, for the AP88-89 text collection the RI for our method is 0.70 which is the highest RI value reported in this thesis, it improves the RM3 method in more than 38% and it is close to the maximum RI, we shall remember that the RI measure spans from -1 to 1.

5.5 Related Work

Related with score distributions *per se*, several works have addressed the finding of best distributions models. In Section 5.2 we already reviewed the most important works about this topic. Recently, some efforts have been presented in the direction of modelling the score distributions in a systematic way (Kanoulas et al., 2010), producing an analytical process based on the form of the scoring formulas of the retrieval models.

Score distributions modelling has been applied to tasks such as information filtering or distributed IR, but, in particular, we shall remark the works of Manmatha et al. (2001) where score distribution modelling was applied in order to combine the outputs of different search engines for the meta search task, and the works of Arampatzis et al. (Arampatzis et al., 2000, 2009) which formulated the threshold optimization problem over the score distributions models for locating a good cut-off point in the legal search task. The objectives in both cases are quite different from ours, for instance, the legal search task is a high recall task, meanwhile in our case we desire the opposite: a high precision cut for determining the RS.

Very few works have been presented in the direction of refining the RS. Winaver et al. (2007) presented a language modelling approach for improving the robustness of the PRF methods. This approach, given a query, computes a set of different language models corresponding with different parameter settings, then, the best computed language model (two different strategies for deciding which one is the best are presented) is selected as initial retrieval. Secondly, different language models are computed using different configurations of r and e over the chosen initial retrieval, selecting that one with the minimum KLD with the query model for processing a second retrieval. Evaluation is not conclusive and no comparison with train and test approach is presented. Moreover, this method requires of a high number of computations of language models for each query, which is quite expensive in terms of computational costs. This last fact is more evident if we compare with our proposal which does not require any extra relevance or language model computation but a very efficient expectation-maximisation process over a limited set of scores.

Huang et al. (2008) remarked the importance of selecting the adequate number of feedback documents for the PRF methods. This work explores two different approaches for query-specific feedback document selection. The first approach determines the size of the RS for a given query using either clarity score or cumulative gain. The second one instead of locating the optimal number of documents in the RS uses a mixture model by combining all the query language models rather than only selecting one with the hope of smoothing the effects of the different models. Neither the clarity score base method, nor the cumulative gain strategy, nor the mixture model are able to achieve significant improvements in any collection over the training-test strategy.

In Zhang et al. (2009) a different view to the problem of the presence of

irrelevant documents in the RS is presented. This paper proposes a distribution separation model than taking as input a seed of non-relevant documents and the mixed distributions of the RS will try to estimate an approximation to the true relevance distribution. Evaluation results are interesting but they depend of the existence of relevance judgements to determine the irrelevant seeds (up to the 30% of the known non-relevant documents in the RS are used by the algorithm).

Another open research line is to produce models less sensitive to the composition of the RS. As commented in previous chapters, Li (2008) presented a new estimation for the relevance models which combines three different aspects: common word discounting, non-uniform document priors and the modification of the traditional pseudo feedback paradigm by considering the original query as a pseudo feedback document rather than combining it with the expanded query. With the introduction of three additional parameters in the model, the method seems to be more robust to the variation in the number of feedback documents than RM3, the effectiveness, once reached the optimal size of the RS, drops slower than for RM3 when increasing the number of pseudo-relevant documents.

5.6 Conclusions

In this chapter we showed how the size of the RS greatly affects to the performance of the RM methods. Motivated by that fact, we presented a method which introduces the use of the threshold optimisation problem over score distribution modelling for automatically selecting the size of the RS. Particularly our method assumes a mixture of two Gaussian distributions and based on this assumption computes the threshold point as the score over which the highest density of relevant documents is obtained.

We have used Bregman soft clustering in order to learn the distributions' parameters from the observed scores. The results of the evaluation showed that in terms of MAP our method is equivalent of better to standard RM3. Important improvements in terms of robustness are obtained with respect to RM3, achieving more than a 38% in the case of the AP88-89 collection. Analyses of the results suggest that our modelling decisions perform better in textual collection than in web collections. Overall, the general objective of improving the robustness of the RM estimations is achieved and moreover, we

present the extra advantage of reducing the number of parameters involved in the estimation of the Relevance Models.

Chapter 6

Relevance-Based Language Modelling of Constrains for Text Clustering

Previous chapters model the retrieval task, in this chapter we will adapt a different task: constrained clustering. Constrained clustering is a recently presented family of semi-supervised learning algorithms. These methods use domain information to impose constraints over the clustering output. The way in which those constraints (typically pair-wise constraints between documents) are introduced is by designing new clustering algorithms that enforce the accomplishment of the constraints. In this chapter we present an alternative approach for constrained clustering where, instead of defining new algorithms or objective functions, the constraints are introduced modifying the document representation by means of their language modelling. More precisely the constraints are modelled using Relevance Models. To the best of our knowledge this is the first attempt to try such approach. The results show that the presented approach is an effective method for constrained clustering even improving the results of existing constrained clustering algorithms.

6.1 Introduction and Motivation

Clustering is an important data mining tool in order to exploit the knowledge present in the document collections. Lately it has been also demonstrated as an useful tool not only by itself but also for other Information Retrieval (IR) tasks such as cluster-based retrieval (Lee et al., 2008) or clustering of search results (Zeng et al., 2004). Recently a new family of constrained clustering algorithms (Basu et al., 2008) has achieved great importance because they enabled the introduction of domain knowledge in the clustering process. In these semi-supervised methods the domain knowledge is introduced as rules in a generalized framework making the algorithm itself still domain-independent. In this way knowledge that was unused in traditional clustering algorithms is exploited to improve the grouping of data.

Till this moment, the way in which this new clustering task was carried out was by designing new specifically tailored algorithms. Due to the popularity of the task, several new algorithms appeared based on traditional clustering algorithms: partitional algorithms (Wagstaff et al., 2001; Ares et al., 2009), hierarchical algorithms (Klein et al., 2002; Bae and Bailey, 2006), probabilistic approaches (Basu et al., 2004; Zhai et al., 2011), matrix decomposition based methods (Ji et al., 2006; Wang et al., 2008a), etc. In fact, we have previously published results in this area (Ares et al., 2009, 2010, 2011, 2012), but all these algorithms force the accomplishment of the constraints in the document to cluster assignment or by modifying the objective functions, in contrast, here we propose an approach based on maintaining the simplicity of the clustering algorithms. The idea explored in this chapter is to avoid the creation of new constrained clustering algorithms and keep using the well-known and tested clustering algorithms for this new semi-supervised clustering task. So the question is how unsupervised clustering algorithms can be used for constrained clustering? To the best of our knowledge this is the first time that this question is answered. Our proposal is by introducing the constraints directly in the document representation by means of their Relevance-Based Language Modelling.

The main contributions of this chapter are on one hand the design of a new approach to constrained clustering which allows the use of unsupervised clustering algorithms instead of the specially tailored new ones and on the other hand to allow so by modifying the document representation by means of the language modelling of the constraints. More precisely our proposal

Document Retrieval	Constrained Text Clustering
query (q)	constrained document (d)
query words ($q_1 \dots q_n$)	words of the document ($d_1 \dots d_n$)
pseudo-relevant set (RS)	constrained set ($C(d)$)
candidate terms for query expansion	candidate terms for document expansion

Figure 6.1: Analogies between the use of Language Based Relevance Models for document retrieval and for constrained text clustering

is to expand the documents that are affected by constraints using Relevance Models (Lavrenko and Croft, 2001).

In this chapter we will use the RM framework to alter the original document representations (Parapar and Barreiro, 2012). In RM the query and the documents in the relevance set are assumed as samples of the same Relevance Model, in our proposal we assume that there exists a Relevance Model which generates a document and the set of documents that share constraints with the given document. Therefore, for every document we can estimate the Relevance Model given the documents that constrain it. Meanwhile in the PRF task a query is expanded with the best terms of the relevance model obtained from the relevance set, in the clustering task, every document which is affected by a set of constraints, will be expanded with the best terms of the relevance model obtained from the set of documents which it shares constraints with (see Figure 6.1).

The rest of the paper is organized as follows. Section 6.2.1 presents the proposed method for the language modelling of the constraints. Section 6.3.1 explains the clustering algorithms with which the approach is tested with some considerations about distance functions. In Section 6.4 the evaluation and its results are reported. Section 6.5 describes the related work and, finally, conclusions are reported in Section 6.6.

6.2 Background

In this section we introduce some general basis for this chapter not reviewed in the general background chapter (2).

6.2.1 Constrained Clustering

As previously exposed constrained clustering algorithms use the background knowledge to drive the clustering process. Constrained clustering is different from a classification task, where it is exactly known which groups exist in the data and examples of those categories are provided to the algorithm. In constrained clustering the domain knowledge gives the clustering algorithm rules over documents. These rules reflect some preferences about whether or not the documents should be in the same cluster, being still the algorithm which finds the groups in the data.

With these constrained clustering approaches, knowledge that was unused in traditional clustering algorithms is used to improve the grouping of data in real domains. This is useful, for example, in collections where data instances contain information that comes from multiple evidence sources, such as medical reports. When wanting to cluster such a collection using the report text as the main source of evidence, it could be also useful to introduce in the process some domain knowledge about dates, geolocalisation, race or gender of the patients. Constrained clustering can also be useful in collections where data points have an obvious grouping and the traditional algorithms tend to be biased to that clustering, being that one not interesting for the users. Using the same domain as in the previous example, a clustering algorithm could point out a well-known relation between a disease and the patients' age, however, the medical doctors might want to get an alternative explanation of the data, relating it to other factors.

Most of existing constrained clustering algorithms rely over the so called instance level constraints (Wagstaff and Cardie, 2000). Instance level constraints can be defined as rules between two documents referring to whether (positive constraints) or not (negative constraints) they must be part of the same clustering. Depending on the algorithm design and the enforcement desired for the constraints they are commonly classified in: absolute constraints, constraints that the algorithm can not violate and must mandatory honour at the end of the clustering process (Must-Link and Cannot-Link for positive and negative constraints respectively); and soft constraints, non absolute constraints that the algorithm could not honour at the end of the clustering process (May-Link and May-Not-Link for positive and negative constraints respectively).

When working in real scenarios, dealing with non categorical information

is the most common situation. Therefore, soft constraints are commonly used taking advantage of the adjustment of the parameters that controls the enforcement of the soft constraints in the algorithms which support that kind of constraints. From now, when talking about constraints we will refer to positive soft constraints, i.e., May-Links.

6.3 Relevance-Based Language Modelling of the Constraints

One important point in every clustering algorithm is the way in which the documents are represented. Over that representation will relay the computation of the similarity/distance functions among documents and/or centroids. When dealing with textual documents, they are usually represented according to the Vector Space Model, assigning one dimension to each term in the lexicon. The way in which each term is weighted for every document varies being the TF·IDF and the pointwise Mutual Information the most used weighting schemas due to their good performance.

In this chapter we want to introduce the constraints in the document representation under the LM framework. In order to do so, we have to consider the document representations as probability distributions. So we decided to weight the terms by means of the maximum likelihood estimator. Once that the original document representation is defined we proceed to the constraint modelling. Let us define $C(d) = \{\hat{d}^1, \cdots, \hat{d}^{|C(d)|}\}$ as the set of documents that share a constraint with the document d. In order to introduce those constraints in the document representation our proposal is:

1. Let us suppose that for every affected document d and its $C(d)$ a supporting Relevance Model exists.

2. That Relevance Model can be estimated under the RM framework.

3. From the estimated Relevance Model the e best terms are selected to alter the original representation of the document d. Then a linear interpolation is done as in RM3, being λ in this case the parameter that weight the importance of the constraints in the interpolated representation.

4. Use the altered document representation in the clustering process with unsupervised algorithms.

$$P(w|R) \propto \sum_{\hat{d} \in C(d)} P(\hat{d}) \cdot P(w|\hat{d}) \cdot \prod_{i=1}^{|d|} P(d_i|\hat{d}) \tag{6.1}$$

In Eq. 6.1 the reformulation of the Eq. 2.10 for our task is presented. Equation 6.1 gives the estimation of probabilities in the Relevance Model underlying d and the set of documents $C(d)$ that constrains it. In practice $P(\hat{d})$ can be considered to be uniform. In our task the role of q in the query likelihood presented in Eq. 2.10 is played by the document affected by constraints d meanwhile the role of the RS is played by $C(d)$. As result of the way of how $C(d)$ is constructed $\prod_{i=1}^{|d|} P(d_i|\hat{d})$ should be considered uniform because the constraints are defined explicitly having everyone the same weight. Talking in terms of relevance each time a constraint is explicitly established between two documents d^x and d^y it is equivalent to assess that the document d^x is relevant for the document d^y and *vice versa*, non existing any grading in the relevance assessment. Therefore the final estimation used in this approach is presented in Eq. 6.2, the final document representation is then computed as in Eq. 6.3.

$$P(w|R) \propto \sum_{\hat{d} \in C(d)} P(w|\hat{d}) \tag{6.2}$$

$$P(w|d') = (1 - \lambda) \cdot P(w|d) + \lambda \cdot P(w|R) \tag{6.3}$$

6.3.1 Clustering Algorithms

Before presenting the clustering algorithms that we will use to assess our proposal (K-Means family and Normalized Cut family) we have to do some consideration about the similarity/distance functions. As previously stated when working in the LM framework we will work with probability distributions, so in order to be scrupulous with that fact we have to work with similarity/distance functions according to that. In IR usually Kullback Leibler Divergence (KLD) as in Eq. 6.4 is used in such cases. Unfortunately KLD is only defined when $Q(i) > 0$ for any i such that $P(i) > 0$ and also is a non-symmetric measure. One of the algorithms that we will use, the Normalized Cut algorithm, requires a symmetric function so we decided to use the I-Divergence

to the mean (IDM). This is a symmetric version of the I-Divergence (both previously successfully used in the clustering task (Basu et al., 2004)) that is a Bregman divergence, a family of divergence functions including the KLD and squared Euclidean distance that guarantees the decrease of the K-Means objective function (Banerjee et al., 2005). So the distance function between two documents, d^x and d^y, used in every algorithm is IDM defined as in Eq. 6.5.

$$KLD(P \parallel Q) = \sum_i P(i) \cdot \log \frac{P(i)}{Q(i)} \tag{6.4}$$

$$IDM(d^x, d^y) = \sum_{i=1}^{n} d_i^x \log \frac{2d_i^x}{d_i^x + d_i^y} + d_i^y \log \frac{2d_i^y}{d_i^x + d_i^y} \tag{6.5}$$

In Section 6.4 a preliminary experiment is presented comparing the presented set-up (MLE as document representation with IDM as distance function) with the traditional set-up for text clustering (TF·IDF and cosine distance) in the unsupervised algorithms, showing that our proposal is not only competitive but also significantly improves the traditional set-up.

In this chapter we will asses our proposal with two clustering families: partitional and spectral algorithms. Next we will briefly revise the algorithms:

6.3.1.1 Partitional algorithms

The batch K-Means (KM, MacQueen and McQueen (1967)) algorithm is a well-known efficient iterative clustering algorithm. It is one of the most popular ones due to its simplicity and good performance, which enables its use in large datasets.

A constrained counterpart of KM is the Soft Constrained K-Means (SCKM, Ares et al. (2009)). SCKM is an extension to KM which allows the introduction of soft constraints in the clustering by altering the similarity values between documents and centroids: the similarity score is initialised with the similarity between the document and the centroid of the cluster, and it will be modified depending on the soft constraints affecting the data instance. Namely, the score of a cluster is increased a certain amount w for each document which was last assigned to that cluster and has a constraint with the document being assigned.

6.3.1.2 Spectral Algorithms

Spectral Clustering algorithms use graph spectral techniques to tackle the clustering problem transforming it into a graph cut problem. Thus, finding a *good* clustering of the data in k clusters can be reformulated in terms of finding a *good* cut of a weighted graph where each vertex corresponds to a data point and the weight of an edge is proportional to the similarity between data points. One of the most popular is Normalised Cut (NC, Shi and Malik (2000)), defined in a way such a cut of the graph with a low NC value corresponds to a good (as defined above) clustering of the data. Hence, the Normalised Cut (NC) algorithm proceeds building the graph from the data and finding a cut of it with a small NC value.

It can be shown that the minimisation of NC can be presented as a matrix trace minimisation problem (Shi and Malik, 2000), which, if subject to some constraints, would yield the exact solution. Unfortunately this is NP-hard problem, and so the constraints have to be relaxed in order to make the algorithm computationally affordable. With this relaxation, the documents are projected in a reduced space ($\mathbb{R}^k$, where k is the desired number of clusters) using the smallest k eigenvectors of a Laplacian matrix of the graph. Given these projections, K-Means is used to find a discrete segmentation of this space. Once this segmentation has been performed, we can backtrace each projected document to the original one, obtaining the final outcome of the NC clustering algorithm.

Ji et al. (2006) proposed a Constrained Normalised Cut (CNC) algorithm which introduces soft constraints in NC. In order to do so, they altered the function minimised in the NC algorithm to obtain a new one, such that the cut of the graph which minimises this function would convey a grouping which is still a good one but also tries to respect the constraints supplied by the user. To achieve this, they built a new matrix which encodes positive constraints and introduced it in the core of the minimisation problem, controlling the degree of enforcement of the constraints with a parameter β, with higher values of this parameter meaning a tighter enforcement. The result of the minimisation is a projection of the points in $\mathbb{R}^k$, and so a segmentation of the projected documents has to be performed in order to produce the final clustering of the data.

Two considerations have to be done about the spectral methods. It is very common to pre-process the similarity matrix between documents with a

Gaussian Filter. When using IDM as distance function its form is:

$$e^{\left(\frac{-IDM(d^x,d^y)}{2\sigma^2}\right)} \tag{6.6}$$

Also in practice the dimension of the reduced space is taken greater than k (the number of desired clusters) because it performs considerably better (Jin et al., 2005; Ares et al., 2012), let us call this dimension δ, the number of eigenvectors kept in the projection phase.

6.4 Experiments and Results

In this section we report the results of the evaluation of the different clustering approaches. Some methodological remarks particular to the clustering field have to be done due to the difference with retrieval evaluation.

6.4.1 Constraints and Seed Initialisations

All the presented algorithms are affected by the seed initialization problem of the KM algorithm. In order to reduce that problem, for every algorithm we did ten runs with different seeds, the same seeds in each collection for the six different algorithms. The results reported in the table 6.2 are the average for the ten different initialisations.

KM and NC are not affected by constraints (their values are reported as baselines), for SCKM, CNC, KM_{RM} and NC_{RM} we have to consider also the constraint generation. So for every seed initialisation, we did five different randomly chosen constraints sets. These constraints represent the 1% of all the possible constraints and the same constraints are used in each collection for the four different algorithms. The result for every seed initialisation in the algorithms affected by constraints is the average of the five different constraints sets. The constraints were created from the reference grouping used as clustering ground truth by randomly selecting pairs of documents which belonged to the same cluster, as it is traditionally done in constrained clustering evaluation.

6.4.2 Collections

We run experiments with publicly available datasets that have been widely used in the evaluation of clustering algorithms:

1. ModApte: a split of Reuters-21578 with documents belonging to one of the biggest ten categories considering only the documents categorised in only one group (7282 documents, 10 groups)

2. WebKBUniversities: the WebKB dataset with the golden truth corresponding to universities, and taking only the documents from Cornell, Texas, Washington and Wisconsin universities and removing those corresponding to "misc", "other" and "department" (1087 documents, 4 groups).

3. WebKBTopics: the same dataset as (2), but this time distributed in five groups, corresponding to the topics "course", "faculty", "project", "staff", and "student" (1087 documents, 5 groups).

4. News3Related: a sample of three categories of the 20 Newsgroups collection. Following the same approach in Basu et al. (2004), we have chosen 300 documents randomly from each of the categories `talk.politics.misc`, `talk.politics.guns`, and `talk.politics.mideast` (900 documents, 3 groups).

We decided to choose the WebKB collection and both of its categorization because in this collection the bias problem occurs, tending the clustering algorithms to follow one of the categorizations. Dealing with this problem is a very common task for the constrained clustering algorithms (avoiding bias task), therefore it is an interesting collection for the evaluation of constrained clustering algorithms. The use of small datasets comprised by sparse high-dimensional data is interesting because the clustering task is notably difficult, as the clustering algorithms are more prone to fall in local minima (Basu et al., 2004).

6.4.3 Compared Methods

The primary objective of this chapter is to assess the use of Relevance Modelling to modify the document representations enabling the use of unsupervised clustering algorithms for the constrained clustering task. So in the ex-

periments we will compare the performance of two different family of clustering algorithms, partitional and spectral ones, by their traditional formulation (**KM** and **NC**), the constrained counterparts (**SCKM** and **CNC**), and the traditional formulation with the constraints modelled in the document representation ($\mathbf{KM_{RM}}$ and $\mathbf{NC_{RM}}$).

6.4.4 Metrics

In order to assess the effectiveness of the different clustering algorithms we have compared the outcomes of the algorithms with the reference groupings using three metrics: Adjusted Rand Index (ARI), Purity and Entropy. However, as the results for the three metrics show the same trends, only the results for Adjusted Rand Index (Hubert and Arabie, 1985) are reported in this thesis.

This metric measures the ratio of good decisions made by the algorithm on a pairwise basis (Eq. (6.7)). It is based on Rand Index (Rand, 1971), correcting certain deficiencies of that metric, namely, it is corrected for chance. To do so, the expected value of the index is subtracted from the unadjusted index and the result is divided by the maximum value of index (from which the expected value has been subtracted as well). Higher values of Adjusted Rand Index indicate a greater similarity between the results and the reference.

$$ARI(\Omega;\mathbb{C}) = \frac{\sum_{ij}\binom{|\omega_i \cap c_j|}{2} - \frac{\sum_i \binom{|\omega_i|}{2} \sum_j \binom{|c_j|}{2}}{\binom{N}{2}}}{\frac{1}{2}[\sum_i \binom{|\omega_i|}{2} + \sum_j \binom{|c_j|}{2}] - \frac{\sum_i \binom{|\omega_i|}{2} \sum_j \binom{|c_j|}{2}}{\binom{N}{2}}} \tag{6.7}$$

where $\{\omega_1, \omega_2, \ldots, \omega_k\}$ are the set of clusters and $\{c_1, c_2, \ldots, c_k\}$ is the set of classes defined in the golden truth.

6.4.5 Training

To deal with the values of the parameters involved in the different approaches we decided to use traditional training and test methodology. We tuned the parameters for ARI in the ModApte collection, and the trained values were used in the other collections.

However the parameters σ (the Gaussian filter parameter) $\sigma \in \{0, 0.05, 0.10, 0.15 ... 0.90, 0.95, 1\}$ and δ (the number of eigenvectors keep in the projection phase) $\delta \in \{1, 5, 10, ... |C|\}$ involved in the spectral algorithms had to be tuned for every collection because they are very sensitive, they were tuned

in the NC algorithm and those trained values were used by the constrained versions.

So the parameters tuned were: the parameters w and β for the enforcement of the constraints in the SCKM and CNC algorithms take values in $\{0.00250,\ 0.00500, 0.0125, 0.0250, 0.0500\}$ and $\{5, 10, 20, 30\}$ respectively. The parameters involved in the RM estimation namely, the Dirichlet smoothing parameter μ which takes values in $\{5, 10, 15, 25, 50, 100, 500, 1000\}$, was trained in the KM algorithm and the same values used in the NC algorithm, the parameter e (the number of terms selected from the Relevance Model) was set to 500 without tuning it. Furthermore, the interpolation parameter λ which takes values in $\{0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1\}$ was tuned using the same strategy as with μ.

In the experiments, as it is common practice on clustering evaluation, we have considered that the number of clusters (k) in the grouping used as reference was known, and so the number of desired clusters was set to that amount in each of the tested clustering algorithms

6.4.6 Statistical Significance

Finally, we have assessed the statistical significance of the results of the experiments using the Sign Test (Conover, 1971), a choice which was motivated by its reduced number of assumptions about the data in comparison with other tests such as Wilcoxon's or Student's t. The results of each approach were compared with the rest of the methods for every collection.

For each test ten observations $(ARI_{xi}, ARI_{yi}), i \in [1..10]$ were considered, one for each initialisation of the seeds, where ARI_{xi} is the ARI of the method X and ARI_{yi} is the ARI for Y. Over these observations we performed a Lower-Tailed test, where the null hypothesis was $H_0 : P(+) \geq P(-)$, i.e. , that the values ARI_{xi} were greater or equal to ARI_{yi} (meaning that the quality of the results of the method X was greater or comparable to that of the Y method), and the alternative hypothesis was $H_1 : P(+) < P(-)$.

6.4.7 Results

In order to clarify the competitiveness of the baselines given the experimental conditions in terms of document representations and distance measures, a preliminary experiment was carried out in the ModApte collection compar-

ing for both KM and NC the averaged ARI values when using classical TF·IDF document representation and cosine distance function and when using the experimental conditions designed in this chapter. Results are reported in Table 6.1 showing not only that the probabilistic representation in combination with the IDM measure performs well but it also significantly outperforms the classical clustering set-up.

Table 6.1: Adjusted Rand Index values, statistical significant improvements w.r.t to the alternative set-up for each algorithm according with the Sign Test are starred (the null hypothesis is rejected for a p-value ≤ 0.0547).

	ARI	
Set-up	*KM*	*NC*
TF·IDF and Cosine	0.319	0.311
MLE and IDM	0.446*	0.648*

Table 6.2: Adjusted Rand Index values, statistical significant improvements w.r.t KM, SCKM, KM_{RM}, NC, CNC and NC_{RM} according with the Sign Test are marked as k, s, κ, n, c, η respectively (the null hypothesis is rejected for a p-value ≤ 0.0547). Best values bolded.

	ARI		
Collection	*KM*	*SCKM*	KM_{RM}
ModApte (Training)	0.446	$\mathbf{0.983}^{k\kappa ncn}$	0.820^{kncn}
WebKBUniversities	0.073	0.311^{kn}	$\mathbf{0.581}^{ksncn}$
WebKBTopics	0.230	0.574^{kn}	0.505^{kn}
News3Related	0.183	0.712^{knn}	$\mathbf{0.833}^{ksncn}$
	ARI		
Collection	*NC*	*CNC*	NC_{RM}
ModApte (Training)	0.648^{k}	0.771^{kn}	0.781^{kn}
WebKBUniversities	0.009	0.342^{kn}	0.377^{ksnc}
WebKBTopics	0.331^{k}	$\mathbf{0.734}^{ks\kappa nn}$	$0.668^{ks\kappa n}$
News3Related	0.258^{k}	0.783^{ksn}	0.617^{kn}

In Table 6.2 an effectiveness comparison between the different approaches is presented in terms of ARI. When analysing the results the first consideration is that, as expected, the presented approach performs significantly much better than the unconstrained algorithms, showing that it is a valid approach for the constrained clustering task.

When comparing with the ad-hoc constrained clustering algorithms we

have to remark that in WebKBUniversities and News3Related collections the best method is one based on the language modelling of the constraints and performs significantly better than both constrained algorithms, meanwhile only the CNC can achieve significant improvements over both RM based approaches in one testing collection (WebKBTopics). The evaluation, as commented before, showed similar trends for the other metrics (Purity and Entropy). These numbers show that the proposed approach is valid for the constrained clustering task, achieving results comparable or even better than specially tailored clustering algorithms.

Furthermore, training values helped us to observe that the parameters of the constrained clustering algorithms are much less stable (to the point that, for instance in the SCKM method the performance of some parameters settings fell behind the unconstrained KM method) than the λ parameter of the presented approach. Also it is an advantage that the λ parameter behaviour has been widely studied in other retrieval tasks. On the other hand the interpretability of the role of the parameter λ is very easy and it will only depend on the importance that we want to assign to the constraints in the interpolated model.

6.5 Related Work

So far with the approach presented in this thesis three alternatives exists for the introduction of constraints in the clustering process. (a) The approach presented in this chapter to introduce the constraints directly in the document representation. (b) The design of new specially tailored algorithms as the ones commented in Section 6.1 based on forcing the accomplishment of the constraints in the document to cluster assignment or by modifying the objective functions. (c) An alternative approach introduces the constraints in the clustering process through the use of distance learning methods. Xing et al. (2002) presented an algorithm that given some constraints learns a distance metric over $\mathbb{R}^n$ respecting those constraints, however the efficiency of this approach is compromised by the need of solving a convex optimization problem.

We could frame the method presented in this chapter as a document expansion technique. One of the first successful applications of document expansion in IR was carried out by Singhal and Pereira (1999). In this paper, the

authors presented the application of document expansion techniques to alleviate the effect of transcription mistakes for speech retrieval under the VSM. More recently, Tao et al. (2006) presented the application of document expansion techniques in the LM framework with the objective of improving the retrieval effectiveness lightening the effect of insufficient sampling of documents. In this paper, the authors extend representation of a document by smoothing the document model with the language model of its neighbours.

6.6 Conclusions

In this chapter we have proposed the use of unsupervised clustering algorithms for the constrained clustering task. The main contributions are two: the use of the document representation to code the constraints and the use of Relevance Models under the LM framework to model those constraints. The evaluation showed that the use of our proposal with the traditional clustering algorithms achieves comparable and even better results than specially tailored constrained clustering algorithms, allowing in this way the use of the unsupervised algorithms for the constrained clustering task.

Our proposal has been built upon a strong and well-studied theoretical base as is the Language Modelling framework which allows the interpretability of the elements involved in the approach pretty straightforward.

Chapter 7

Relevance-Based Language Modelling of Recommender Systems

The Recommender Systems field is a fertile research area where users are provided with personalised recommendations in several applications. In this chapter, we propose an adaptation of the Relevance Modelling framework to effectively estimate a user's recommendations. We also propose a probabilistic clustering technique to perform the neighbour selection process as a way to achieve a better approximation to the set of relevant documents in the pseudo-relevance feedback process. Empirical evaluation results show that both proposals outperform individually the baseline methods. Furthermore, by combining both approaches even larger effectiveness improvements are achieved.

7.1 Introduction and Motivation

Recommender Systems have traditionally been a fertile research area due to the existence of a wide range of scenarios where users may benefit from automatic personalised recommendations. This research area has its roots in the eighties, and started to attract wider attention in the mid-nineties when the first works on collaborative filtering were published (Resnick et al., 1994; Hill

et al., 1995). Collaborative Filtering (CF) is one of the three classical approaches to recommendation (Adomavicius and Tuzhilin, 2005): *content-based recommendation*, based on the user's history; *collaborative filtering*, based on the history of similar users; and *hybrid approaches*, based on combining content-based recommendation and collaborative filtering.

In CF (Herlocker et al., 2002), the input evidence about user preferences consists of data records collected from user interaction with items. In the simplest form, this evidence consists of explicit user ratings, which are graded relevance values assigned by end-users to items of interest. CF algorithms exploit the target user's ratings to make preference predictions, and have the interesting property that no item descriptions are needed to provide recommendations, since the algorithms merely exploit information about past interaction between users and items. Moreover, CF has the salient advantage that a user benefits from others' experience, being exposed to novel recommendations produced from the personal preferences of affine users.

Two different types of CF approaches exist: model-based approaches, which learn user/item rating patterns to build statistical models that provide rating estimations, and memory-based approaches, which compute user/item similarities based on distance and correlation metrics (Desrosiers and Karypis, 2011). Memory-based approaches find either like-minded people for the target user (user-based approach), or pairs of items that are liked by common users. In the user-based approach, the set of similar-minded users are called neighbours, and their preferences are combined to predict ratings for the active user. In the item-based approach, items similar to the ones the user has liked in the past are recommended.

The recommendation task has been traditionally formulated and evaluated as a rating prediction problem (Adomavicius and Tuzhilin, 2005). However, in practical terms, the effectiveness of recommendations depends on what items are presented to the user and in what order. Thus the ranking of recommender items, rather than the numeric system scores that determine this ranking, is the essential problem in common recommendation scenarios, whereby recommendation can be seen as an IR task (one where there is no explicit query). Considering this, several attempts have been recently made to formalise the recommendation task as a relevance ranking problem (Wang et al., 2006b, 2008c,b; Bellogín et al., 2011). The objective is to take advantage of well-studied and highly-performing Information Retrieval (IR) techniques to model the notion of relevance. Such attempts have explored

Document Retrieval	Recommendation
query (q)	target user (u)
query words ($q_1 \dots q_n$)	items rated by user ($I(u)$)
pseudo-relevant set (RS)	user neighbourhood (V)
candidate terms for query expansion	candidate items for recommendation

Figure 7.1: Analogies between the use of Language Based Relevance Models for document retrieval and for item recommendation

the adaptation of the vector-space IR model (Bellogín et al., 2011), the extended Boolean model (Bellogín et al., 2011b), the binary independent retrieval model (upon the PRP) (Wang et al., 2008c,b), and statistical Language Models Wang et al. (2006b). However, to the best of our knowledge, no attempt has been made yet at a similar adaptation of so-called Relevance-Based Language Models (Lavrenko and Croft, 2001).

The adaptation of RM to recommendation is non-trivial as, to begin with, there are neither queries nor words in the generic recommendation task. In our proposed approach (Parapar et al., 2013), the role of the query is played by the user to whom we want to provide with item recommendations. In our adaptation of RM to recommendation, query expansion shall thus become a form of user profile expansion. The role of the pseudo-relevant documents shall be played in our model by the set of similar users (based on profile similarity to the target user). Hence, the objective is to select good items to recommend from the profile of those similar users. The set of analogies on which our approach is based is shown in Figure 7.1).The aim of our approach is thus to leverage the effectiveness of the Relevance Models to estimate the probabilities of relevance, even when the probability distributions are not expressed in terms of words as originally proposed for text retrieval.

A good approximation of the set of relevant documents is critical to the effectiveness of pseudo-relevance feedback methods. Analogously, a good selection of user neighbours (which we are taking as the equivalent of pseudo-relevant documents) can be expected to heavily influence the effectiveness of our approach. In the context of a probabilistically formalised framework as we intend to build, Posterior Probabilistic Clustering (PPC) (Ding et al., 2008) provides a rigorous probabilistic basis for neighbourhood formation by user clustering, based on Non-negative Matrix Factorisation (NMF). Besides the probabilistic interpretability of this method, the NMF family of algorithms

has proved to have a very good performance in terms of clustering effectiveness (Xu et al., 2003). This method is particularly convenient as an effective neighbouring technique, providing an indicator of the strength (degree of membership) between a user and her neighbourhood. In this chapter we shall explore the use of this particular probabilistic clustering both in isolation (as an enhancement of neighbour selection in CF recommendation), and in combination with the relevance modelling of the recommendation process.

In summary, we present a new recommendation approach based on the Relevance Modelling of the problem under the Statistical Language Modelling framework, the use of probabilistic clustering methods for the neighbour selection problem, in particular, the use of Posterior Probabilistic Clustering and the combination of both contributions, leading to even better performance than their separate application.

The remaining of the chapter is structured as follows: Section 7.2 presents the adaptation of Relevance Modelling framework to the recommendation problem. In Section 7.3 we introduce our proposal for neighbour selection based on Posterior Probabilistic Clustering. Section 7.4.1 reports the empirical evaluation of the proposed approaches and analyses the results of different experiments. In Section 7.5 we present a study of the works related to our proposal. Finally, conclusions are presented in Section 7.6.

7.2 Relevance-Based Language Modelling for Recommendation

Prior to present our estimations of RM for the recommending task we shall present the RM2 formulation that we use in this chapter. In Chapter 2 we present the final estimations under the i.i.d. sampling assumption for RM1 (Eq.2.10). In RM2 (*conditional sampling*) the main assumption is that the query words are independent from each other but dependent on the words of the relevant documents. As a result of that, $P(w|R)$ is computed as in Eq. 7.1 (Lavrenko and Croft, 2001):

$$P(w|R) \propto P(w) \prod_{i=1}^{n} \sum_{d \in \mathcal{C}} P(q_i|d) \frac{P(w|d)P(d)}{P(w)} \tag{7.1}$$

As stated before, the final objective in pseudo-relevance feedback is to select from the pseudo-relevant set good terms which are related to the original

query terms. In the case of retrieval, the goodness of those selected terms is evaluated by how their adding to the original query produces a more effective second document ranking. In recommendation and, particularly, in collaborative filtering, the user is modelled as a set of already scored items. For our proposed approach, those items shall play the role of the query words in IR relevance models, and the objective is to provide the user with more good items corresponding with her already assessed interests. We thus propose a formulation of the recommendation process as a profile expansion problem, where the items to recommend play the role of the candidate expansion terms in the pseudo-relevance feedback task. In this way, the recommendation problem can be accommodated as a profile expansion process where the models for pseudo-relevance feedback can be tested.

Specifically, we propose new Relevance Models estimations for the recommending task. In order to accommodate the recommendation process in such a way, we have to suppose that for every target user $u \in \mathcal{C}$ and set of *relevant users* or *neighbours* (V) a supporting Relevance Model R_u exists. This underlying relevance model can be estimated under the RM framework and, from this estimation, the ranking of best items to recommend to the user u are selected. It is important to note that this model is agnostic with respect to how the relevant users are determined, that is, different neighbour selection methods can be incorporated in a straightforward way. Indeed, we will go back to this point later on and show how different selection approaches can be integrated into our model.

7.2.0.1 Method 1: i.i.d. sampling

Analogously to the RM1 estimation, we produce an RM1 based recommendation. In this context, we assume that the items in the user's profile and the items rated by the user's neighbours are sampled identically and independently from a unigram distribution. Eq. 7.2 defines the estimation of probabilities in the Relevance Model underlying u and V. For every item i in the set of items scored by the similar users V (where V acts as the relevance set) the probability of the item i given the relevance model R_u for user u is computed as:

$$P(i|R_u) \propto \sum_{v \in V} P(v) \times P(i|v) \times \prod_{j \in I(u)} P(j|v) \tag{7.2}$$

where $I(u)$ is the set of items already rated by the user u.

Therefore, assuming the prior for a user's neighbour as uniform and being $\prod_{j \in I(u)} P(j|v)$ the user's profile likelihood for the neighbour v, we can estimate the probability of an item under the Relevance Model for a given user, as the weighted average of the language model probabilities for the item in the neighbourhood of the user, where the weights are the user profile likelihood scores for her neighbours.

Given this scoring formula, the top items can be selected for recommendation by ranking the items according to the probability $P(i|R_u)$. Additionally, an explicit rating estimate can be computed, in case it is required by the recommendation algorithm.

7.2.0.2 Method 2: conditional sampling

Alternatively, we can make use of the conditional sampling assumption as in the RM2 method. In this case, we assume that items in the user's profile are independent from each other but dependent on the items present in the profiles of the user's neighbours. In this situation, the item preference is computed as follows:

$$P(i|R_u) \propto P(i) \prod_{j \in I(u)} \sum_{v \in V} P(v|i) P(j|v) \tag{7.3}$$

where $P(v|i)$ is estimated with Bayes as $P(i|v)P(v)/P(i)$, that is, the preference score is:

$$P(i|R_u) \propto P(i) \prod_{j \in I(u)} \sum_{v \in V} \frac{P(i|v)P(v)}{P(i)} P(j|v) \tag{7.4}$$

Therefore, as Eq. 7.3 shows, in this case the association between each item and the user's profile is computed using the neighbours that contain both the profile's items and the item as "bridges".

7.2.0.3 Final Estimation Details

For both methods we can initially consider that the prior $P(v)$ is uniform, i.e. every neighbour ($v \in V$) has the same probability of being sampled. The estimation of the probability of an item given a user will be computed by smoothing the maximum likelihood estimate with the probability in the

collection (background collection model), in this case using Jelinek-Mercer smoothing (Zhai and Lafferty, 2004):

$$P_\lambda(i|u) = (1-\lambda)P_{ml}(i|u) + \lambda \cdot P(i|\mathcal{C}) \tag{7.5}$$

where $I(\mathcal{C})$ is the set of items in the collection and $p_{ml}(i|u)$ is estimated as:

$$P_{ml}(i|u) = \frac{\mathrm{rat}(u,i)}{\sum_{j \in I(u)} \mathrm{rat}(u,j)} \tag{7.6}$$

$\mathrm{rat}(u,i)$ is the rating assigned by user u to item i, and $P(i|\mathcal{C})$ is estimated as a maximum likelihood in the whole collection:

$$P(i|\mathcal{C}) = \frac{\sum_{v \in \mathcal{C}} \mathrm{rat}(v,i)}{\sum_{j \in I(\mathcal{C}), v \in \mathcal{C}} \mathrm{rat}(v,j)} \tag{7.7}$$

In the Language Modelling framework and retrieval tasks, Dirichlet smoothing outperforms Jelinek-Mercer (Zhai and Lafferty, 2004). However, when modelling the recommendation problem, Dirichlet can suffer from the undesired effect of demoting those items that have been recently introduced in the system and so have very few recommendations. In fact, in (Wang, 2009) the smoothing for the LM based recommendation with Dirichlet smoothing presents significantly worse performance than using Jelinek-Mercer in one of the experiments reported there. For estimating $P(i)$ we decided to keep it simple and a uniform distribution was chosen.

Finally, depending on the proposed methods, different strategies were used in this chapter to compute the neighbourhood of a given user (V), as we present in the next section.

7.3 A Probabilistic Neighbour Selection Technique

A crucial step in order to rank the items according to the RM framework is to properly select the relevance set, we have addressed that problem in Chapter 5 for the retrieval task. In our adaptation of the RM framework to user-based collaborative filtering, this relevance set is composed by the target user's neighbourhood, that is, the set of her most akin users. Next, we will define an alternative probabilistic approach to the computation of such neighbourhoods. This is not enforced by the RM approach itself, and

other alternatives could thus be considered as well, which we leave as future work. A probabilistic neighbour selection approach provides nonetheless for a smoother global user-based CF framework. In particular, the approach proposed here builds on the Posterior Probabilistic Clustering algorithm, as we present next.

7.3.1 Posterior Probabilistic Clustering

The lack of probabilistic interpretation of Non-negative Matrix Factorisation (NMF) (Lee and Seung, 2001) clustering methods and their ad-hoc document-to-cluster assignments motivated the development of the Posterior Probabilistic Clustering (PPC) method (Ding et al., 2008). PPC provides with a posterior probability interpretation, removes uncertainty in the clustering assignment and has a very close relation to probabilistic latent semantic indexing when performing co-clustering of documents and words.

Given a collection of n documents and m words, let $X = (X_{ij})$ be the words-to-documents matrix where $X_{ij} = X(w_i, d_j)$ is the term frequency of the term w_i in the document d_j. The traditional formulation of the NMF method consists in solving the following optimisation problem, given a number of clusters κ:

$$\underset{F\geq 0, G\geq 0}{min} \parallel X_{m\times n} - F_{m\times \kappa} G^T_{\kappa \times n} \parallel^2 \tag{7.8}$$

Once that the solution (G^*, F^*) to the optimisation problem is obtained, every document d_j is assigned to the cluster C_k such that:

$$k = \arg\max_z (G^*_{jz}) \tag{7.9}$$

where z ranges from 1 to κ.

PPC is a posterior probability interpretation of the NMF algorithm. PPC considers the rows of G^* as the posterior probabilities that a given document belongs to the different clusters, i.e. $P(d_j|C_l) = G^*_{jl}$. In order to enforce a proper probability distribution, a PPC optimisation function is formulated as follows:

$$\underset{F\geq 0, G\geq 0}{min} \parallel X_{m\times n} - F_{m\times \kappa} G^T_{\kappa \times n} \parallel^2, \quad s.t. \sum_{k=1}^{\kappa} Gj_k = 1 \tag{7.10}$$

which results, after using Lagrangian multipliers, in the next updating rules:

$$G_{ik} \longleftarrow G_{ik}\frac{(X^TF)_{ik} + (GF^TFG^T)_{ii}}{(GF^TF)_{ik} + (X^TFG^T)_{ii}} \tag{7.11}$$

$$F_{ik} \longleftarrow F_{ik}\frac{(XG)_{ik}}{(FG^TG)_{ik}} \tag{7.12}$$

This alternative interpretation of the Non-negative Matrix Factorisation algorithm also allows the classical hard clustering task based on the same cluster selection procedure as in NMF (Eq. 7.9). Furthermore, it also represents a probabilistic interpretation of the clustering problem supplying degrees of membership of documents to clusters. This information can also be exploited in the recommendation problem as we shall explain in the next section.

7.3.2 Neighbour Selection based on PPC

As described before, we want PPC to find better neighbourhoods (clusters) for the users. Therefore, we have to adopt certain decisions in order to model the neighbour selection problem in recommender systems with the PPC algorithm. Which representation fits better this particular problem determines our first decision. In the recommendation problem, the role of documents will be played by users and the role of terms will be played by items which, in collaborative filtering, are the constituent elements of the user representation. So we apply the PPC algorithm under the following settings. Having a collection of n users and m items, let $X = (X_{ab})$ be the items-to-users matrix. The weight of $X_{ab} = X(i_a, u_b)$ will be the rating assigned by the user u_b to the item i_a, i.e., $\mathrm{rat}(u_b, i_a)$. In this initial approach to the problem, we assign zero weight when no rating was produced by the user to the item.

Given this formulation of the clustering scenario, once the minimisation problem formulated in Eq. 7.10 is solved, the elements of G^* contain the posterior probabilities of the users given the clusters, i.e., $P(u_b|C_l) = G^*_{bl}$. Having this information, traditional neighbour selection can be done as in hard-clustering by assigning each user only to the cluster C_k such that $k = \arg\max_z(G^*_{bz})$ where z ranges from 1 to κ.

Therefore, for each user u we obtain a neighbourhood V as the cluster to which the user belongs. Given this situation we build a recommender which predicts the rating for user u and item i in the following way:

$$\widetilde{\mathrm{rat}}(u,i) \quad = \quad \frac{\sum_{v \in V} \mathrm{sim}(u,v)\,\mathrm{rat}(v,i)}{\sum_{v \in V} |\mathrm{sim}(u,v)|} \tag{7.13}$$

where $\widetilde{\mathrm{rat}}(u,i)$ represents a predicted rating (as opposed to an actual rating, denoted as $\mathrm{rat}(u,i)$); besides in this case we estimate $\mathrm{sim}(u,v)$ as

$$\mathrm{sim}(u,v) \propto P(v|V) = G^*_{cl} \tag{7.14}$$

provided that the index of user v is c (that is, $v = u_c$) and that $V = C_l$ is the cluster assigned to the target user u.

The only remaining decision is to choose the desired number of neighbours (in our case, the number of clusters that we want to obtain with PPC, i.e., κ). We discuss this point in the following section.

7.4 Experiments and Results

In this section we present three different experiments and discuss the results by comparing the performance of our proposals against standard recommendation techniques.

7.4.1 Collections

In the evaluation of the recommendation methods, we have used two publicly available datasets commonly named as *Movielens 100K* and *Movielens 1M*[1]. Some characteristics about these datasets are shown in Table 7.1. Note that these datasets are different, in particular, the smaller dataset is not a subset of the larger one (although the films are similar, there is no relation between the user information of each dataset), to further emphasise this issue, we have incorporated information about the time span each dataset was collected. Furthermore, as we shall see later we have used the smaller dataset to analyse the sensitivity of our approach to different parameters and the larger one to validate the results.

[1] Both are available at http://www.grouplens.org/node/73

Table 7.1: Statistics about the datasets used in the experiments.

Dataset	#users	#items	#ratings	Sparsity	Recollection Period
Movielens 100K	943	1,682	100,000	6.30%	1997/1998
Movielens 1M	6,040	3,900	1,000,000	4.24%	2000

7.4.2 Compared Methods

In this work, we are proposing different user-based recommendation approaches. Therefore, most of the selected baselines are also user-based methods. We compared our proposals with:

- **UB**: a standard User-Based collaborative filtering method (Resnick et al., 1994) where the neighbourhood is selected among the set of 100 most similar users (according with Pearson's correlation)

- **MF**: a state of the art method which does not use any neighbour selection but is based on Matrix Factorisation through Singular Value Decomposition (SVD) using 50 dimensions (Koren, 2008) and which is generally among the best performing recommendation methods to date.

Moreover, we also tested against other existing proposals based on modelling of the recommendation problem as an Information Retrieval task. We will discuss in Section 7.5 the differences to our proposals. In particular, we test our methods against:

- **UIR-User**: the user-based formulation of the probabilistic interpretation of the relevance models for log-based CF proposed in Wang et al. (2006b), formulated in the Eq. 16 of that paper, that is:

$$P(i|R_u) \propto \sum_{\substack{v \in L_i \\ c(u,v)>0}} \log\left(1 + \frac{(1-\lambda)p_{ml}(v|u,r)}{\lambda P(v|r)}\right) + |L_i| \log \lambda \quad (7.15)$$

  where the sum is over the set of users who have expressed interest for item i ($v \in L_i$) and, at the same time, the number of items rated in common with the target user u ($c(u,v)$) is greater than zero. The maximum likelihood estimator for the user v given the target user u assuming relevance (r) is estimated as follows:

$$p_{ml}(v|u,r) \propto \frac{c(v,u)}{c(u)}$$

And the probability of a user v assuming relevance is estimated by the count of items rated by the user:

$$P(v|r) \propto c(v)$$

- **User-basedRM** the user-based model presented in Wang et al. (2008b), which allows the introduction of ratings in the probability estimations. More specifically, we use the Eq. 40a from (Wang et al., 2008b) which goes as follows:

$$P(i|R_u) = \widetilde{\text{rat}}(u,i) = \frac{\sum_{v \in L_i} \text{rat}(v,i) e^{-\frac{1-\cos(u,v)}{h_u^2}}}{\sum_{v \in L_i} e^{-\frac{1-\cos(u,v)}{h_u^2}}} \tag{7.16}$$

where $\cos(u,v)$ is a cosine kernel based similarity measure (Liu et al., 2004) between the user u and v represented as vectors in the item space, where the missing ratings can be replaced by a constant value of 0 or by the average rating value. As we shall discuss in the related work, this approach requires a prior learning of the value h_u (the kernel bandwidth window parameter) based on an expectation-maximisation process (Wang et al., 2008b). In order to provide a fair comparison, we shall use here the best value reported in Wang et al. (2008b), which was tuned on the very same collection ($h_u^2 = 0.79$).

7.4.3 Training and Evaluation

We performed a standard 5-fold cross-validation evaluation using the splits provided with the collections. This is a typical experimental approach in the recommender systems field, where in each split the 80% of the data is retained in order to produce item recommendations which are evaluated with the 20% of the held out data. Note that this cross-validation has solely evaluation purposes and it is independent from the parameter training. The methodology used in the evaluation corresponds to the *TestItems* approach described in Bellogín et al. (2011a), where, for each user, a ranking is generated by predicting a score for every item in the test set, only ignoring those items already rated

by the user (i.e., in training). We also tested alternative methodologies, such as the one proposed by Koren (2008) where a ranking is generated for each item in the test set based on N additional not-relevant items. We observed similar trends to those reported herein with that methodology in preliminary experiments.

Once a ranking has been generated for each user, e.g., with the TestItems methodology, its performance can be measured using, for instance, the *trec_eval* program[2]. In this way, standard IR metrics such as precision, normalised Discounted Cumulative Gain (nDCG) or Mean Reciprocal Rank could be used. In the following we report effectiveness values for precision at 5 (P@5), precision at 50 (P@50) and normalised discounted cumulative gain with cut-offs at 5 and 10 (nDCG@5 and nDCG@10, respectively). Note that, as already acknowledged in McLaughlin and Herlocker (2004) and Wang et al. (2008b), the rated items in the test users represent only a fraction of the items that the user truly liked, and therefore, the measured metrics may underestimate the true metric values.

Precision at the different cut-offs was used as defined in Section 3.4, in the case of the item recommendation task:

$$P@k = \frac{1}{|\mathcal{C}|_u} \sum_{u \in \mathcal{C}} \frac{Rel_u@k}{k} \tag{7.17}$$

where $|\mathcal{C}|_u$ is the number of users in the collection $Rel_u@k$ is the number of relevant recommended items at the cut-off k.

The normalised Discounted Cumulative Gain (Järvelin and Kekäläinen, 2002) uses graded relevance that is accumulated starting at the top of the ranking and may be reduced, or discounted, at lower ranks:

$$nDCG@k = \frac{1}{|\mathcal{C}|_u} \sum_{u \in \mathcal{C}} \frac{1}{IDCG_u^k} \sum_{p=1}^{k} fdis(rel(u, i_p), p) \tag{7.18}$$

where $rel(u, i_p)$ is the graded relevance of i_p (the item at position p in the ranking), for the user u, $fdis(rel(u, i_p), p)$ is the discount function which is defined by the *trec_eval* software as $fdis(x, p) = \frac{x}{\log(p+1)}$ and $IDCG_u^k$ denotes the score obtained by an ideal perfect ranking for user u at the position k, which acts as a normalisation factor for convenient comparison across different users and datasets.

[2]Available at http://trec.nist.gov/trec_eval/

Regarding the experimental results, we tuned the values of the parameters involved in the different compared methods by optimising P@5 on the small *Movielens 100K* collection, that is, we perform a 5-fold cross validation evaluation as described above in this dataset and report the best values for each parameter. In the case of one of the baselines, we did not have to perform this tuning process, since the optimal parameter values for the same collection were previously reported in Wang et al. (2008b), as we shall point out again in the next section. We also report coverage values following the definition given in Shani and Gunawardana (2011) of *user space coverage*, that is, the number of users for which the system is able to recommend at least one item. After tuning the parameters on *Movielens 100K*, we evaluate the methods in the larger *Movielens 1M* collection, where again, we performed a 5-fold cross-validation setting retaining the 80% of the data for training, and using the same evaluation methodology with the optimal parameters obtained for the first dataset. For this reason, sometimes we will refer to the first dataset as the *training collection*, whereas the second would be the *test collection*.

Finally, to analyse the statistical significance of the results, we performed Wilcoxon Signed-Rank Test (Wilcoxon, 1945), where the performance at user level of two methods are compared. In this case the two paired samples are the concatenation of the user-level effectiveness values of the five different folds.

7.4.4 Results

We present now the experimentation carried out, and the obtained results, in order to validate our contributions and answer the following research questions: (i) are Relevance Based Language Models an effective framework for modelling the recommendation problem? (ii) is it possible to achieve a better neighbourhood selection by applying probabilistic clustering techniques? and (iii) is it possible to achieve further improvements by the combination of both approaches?

7.4.4.1 Experiment 1: Relevance-Based Language Models

In this experiment, we assess the validity of our relevance modelling of the recommendation problem. In order to do so, item recommendations are generated using Eq. 7.2 and 7.4 and the neighbourhoods are constructed with

Table 7.2: Summary of the results for each approach, best values for each collection and metric bolded. Statistical significant improvements according to Wilcoxon Test ($p < 0.01$) w.r.t. MF, UB, User-basedRM, UIR-User, RM1, RM2, PPC, PPC+RM1 and PPC+RM2 are superscripted with a, b, c, d, e, f, g, h and i respectively.

Method	*Movielens 100K (training collection)*				
	P@5	nDCG@5	P@50	nDCG@10	Cvg.
MF	0.081^{bcd}	0.076^{bcd}	0.060^{bcd}	0.074^{bcd}	100%
UB	0.026^{cd}	0.020^{cd}	0.057^{cd}	0.029^{cd}	100%
User-basedRM	0.005	0.003	0.054^{d}	0.018^{d}	100%
UIR-User	0.004	0.002	0.002	0.002	100%
RM1	0.240^{abcdfg}	0.221^{abcdfg}	0.141^{abcdfg}	0.214^{abcdfg}	100%
RM2	0.181^{abcdg}	0.161^{abcdg}	0.089^{abcd}	0.153^{abcdg}	100%
PPC	0.135^{abcd}	0.114^{abcd}	0.108^{abcdf}	0.123^{abcd}	95%
PPC+RM1	$0.320^{abcdefg}$	$0.294^{abcdefg}$	$0.162^{abcdefg}$	$0.282^{abcdefg}$	100%
PPC+RM2	$\mathbf{0.327}^{abcdefg}$	$\mathbf{0.297}^{abcdefg}$	$\mathbf{0.168}^{abcdefgh}$	$\mathbf{0.290}^{abcdefg}$	100%

Method	*Movielens 1M (test collection)*				
	P@5	nDCG@5	P@50	nDCG@10	Cvg.
MF	0.062^{bcdg}	0.061^{bcdg}	0.045^{bcd}	0.060^{bcdg}	100%
UB	0.052^{d}	0.049^{d}	0.038^{d}	0.048^{d}	100%
User-basedRM	0.001	0.001	0.034^{d}	0.006^{d}	100%
UIR-User	0.001	0.001	0.001	0.001	100%
RM1	0.205^{abcdfg}	0.192^{abcdfg}	0.112^{abcdfg}	0.182^{abcdfg}	100%
RM2	0.115^{abcdg}	0.109^{abcdg}	0.064^{abcdg}	0.104^{abcdg}	100%
PPC	0.050^{d}	0.044^{d}	0.059^{ad}	0.050^{d}	98%
PPC+RM1	$0.258^{abcdefg}$	$0.243^{abcdefg}$	$0.133^{abcdefg}$	$0.225^{abcdefg}$	100%
PPC+RM2	$\mathbf{0.294}^{abcdefgh}$	$\mathbf{0.275}^{abcdefgh}$	$\mathbf{0.152}^{abcdefgh}$	$\mathbf{0.258}^{abcdefgh}$	100%

traditional nearest neighbours approach as in the UB method, then we compare the results obtained with these methods against the baselines. The results of the experiments are presented on Table 7.2, denoting **RM1** the results of the RM1 estimation based on the i.i.d. sampling assumption (Eq. 7.2) and **RM2** the results of the RM2 estimation based on the conditional sampling assumption (Eq. 7.4). Furthermore, we present in Figure 7.2 and Figure 7.3 an analysis on the parameter stability of λ (the amount of smoothing in Jelinek-Mercer) in the *Movielens 100K* collection. In all cases, we use the parameter estimation approach described in Section 7.2.0.3.

The results reported in Table 7.2, validate our proposal for the relevance modelling of the recommendation process, showing considerable improvement. Both methods achieve a statistically significant advantage against every baseline. The performance enhancement is considerable over every baseline method (between 120% and 200% of improvement in terms of P@5, depending on the dataset). This clearly indicates that the estimates obtained through our relevance modelling of the recommendation problem are more suitable to obtain good effectiveness values. Profile-expansion style recommendation proves to be a better strategy than pure item ranking based recommendation. The poor behaviour of the UIR-User method was expected because this method does not exploit rating information but only co-rating. Meanwhile, the User-basedRM, which achieved good results in terms of Mean Average Error (MAE) in the original paper, does not perform well in precision oriented tasks. It only achieves comparable results with the other baselines for the $P@50$ metric. The large difference with respect to this method can thus be partly explained by the fact that in the original paper the method is optimised for a different metric from the ones we use here, which are ranking-oriented rather than error-based, as corresponds to a retrieval task.

Overall, this experiment confirms that our proposal of combining neighbourhood information and relevance estimations under the same method is very beneficial to the recommendation task. Furthermore, when analysing the behaviour in terms of the parameter stability in the training collection, it can be observed that both methods are very robust over the parameter values. Meanwhile the optimal λ for the RM1 method is achieved when the amount of smoothing applied is the maximum (hence, the background model is used, which turns into a pure popularity-based recommender).

In the case of the RM2 method, the optimal value is achieved for $\lambda = 0.1$ which indicates that the estimation benefits both from the background model

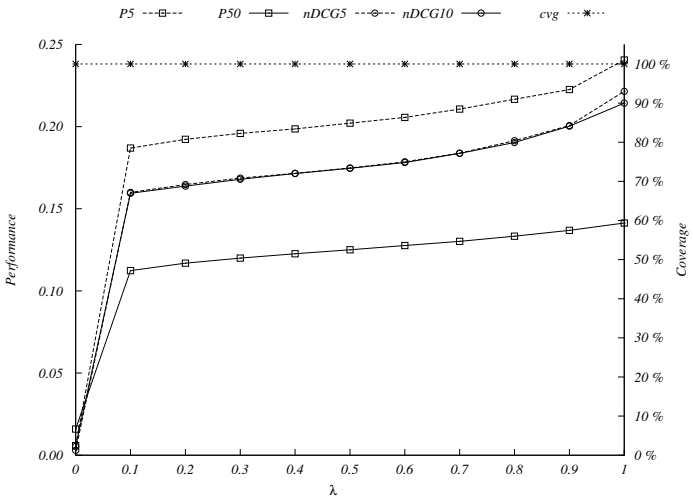

Figure 7.2: Performance and coverage (cvg) in the Movielens 100K collection when varying the amount of smoothing applied for the RM1 method

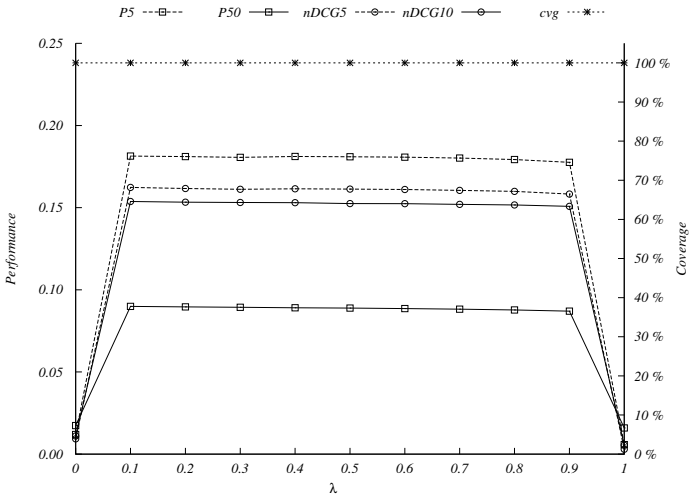

Figure 7.3: Performance and coverage (cvg) in the Movielens 100K collection when varying the amount of smoothing applied for the RM2 method

and the users' models. In this case, the performance of the method based on the conditional sampling assumption is less sensitive than that of the RM1 method. In this experiment, we used traditional neighbourhood selection techniques for user-based collaborative filtering, that is, based on Pearson's correlation and nearest neighbours. In the next experiment, we assess if more elaborated approaches for such task based on probabilistic clustering can improve the performance of the recommendation process.

7.4.4.2 Experiment 2: Probabilistic Clustering for Neighbourhood Selection

The objective of this experiment is to evaluate the suitability of the PPC algorithm for the neighbourhood selection task. We followed the experimental set-up described in Section 7.3.2 and the rating prediction was performed using Eq. 7.13. The results of applying this method for the neighbourhood selection task instead of using a standard nearest neighbour selection (e.g., computing Pearson's correlation) are presented in Table 7.2 denoted as PPC. The most important finding is that the neighbourhood selection based on applying probabilistic forms of clustering greatly enhances the performance of the recommendation. Particularly, this method beats every baseline in the training collection, achieving statistically significant improvements.

It is important to highlight, regarding the test collection, that our PPC method outperforms the UB approach for nDCG@10 and every baseline for P@50. We believe the different performance improvements observed for the two collections may be due to the optimal parameter (κ) found in the training collection, which seems to be insufficient for the test collection. This makes sense since the properties of each collection are very different (943 vs. 6040 users, see Table 7.1). The results, nonetheless, are very promising, and underline the fact that improvements of up to 30% for P@50 are possible by tuning on a separate – but not very different – collection and not using the optimal parameters.

As explained before, only one parameter value has to be determined in this experiment, namely the number of clusters κ. In order to study the behaviour of this method when varying the number of clusters, we report the results over the training collection in Figure 7.4. Interestingly, when increasing the number of clusters the recommendation effectiveness tends to improve but at the expense of coverage. This is explained by the fact that when increasing

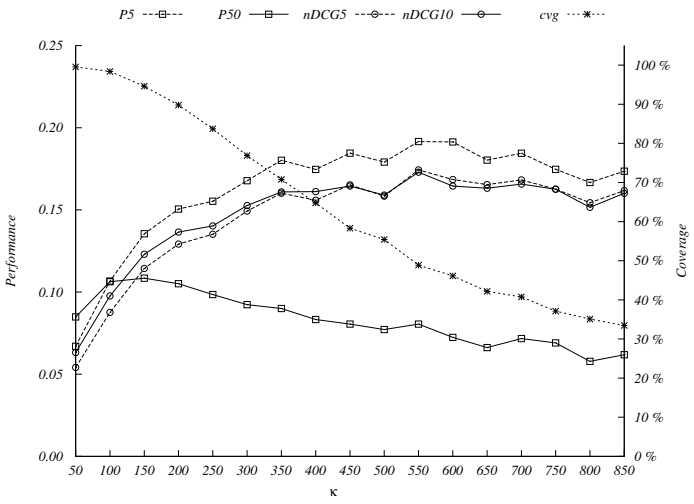

Figure 7.4: Performance and coverage (cvg) in the Movielens 100K collection when varying the number of clusters PPC method

the number of clusters and, at the same time, working with hard-clustering methods, clusters with very few users tend to appear. For very small clusters, it is not possible to produce a good recommendation for the users belonging to them. It can be observed that a value of $\kappa = 150$ (which corresponds with the values reported for the training collection in Table 7.2) provides a good trade-off between coverage and effectiveness in this experiment.

7.4.4.3 Experiment 3: Probabilistic Clustering and Relevance-Based Language Models

Once determined that both approaches, separately, are able to greatly improve the effectiveness of the baselines, we take into consideration the combination of both. In this combination, the neighbourhood selection phase is addressed by applying the PPC method, while the recommendation output is obtained by applying Eq. 7.2 (PPC+RM1) or Eq. 7.4 (PPC+RM2). In this case, we have to train two parameters, namely, the number of clusters for PPC (κ) and the value of the Jelinek-Mercer smoothing parameter (λ). As in previous experiments, we trained and validated those values in the *Movielens 100K* collection and tested them in the *Movielens 1M* dataset. The results for both collections are summarised in Table 7.2. The effectiveness of both

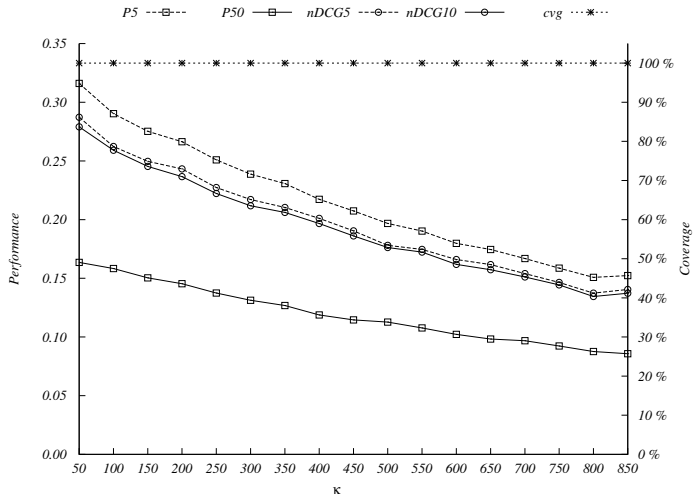

Figure 7.5: Performance and coverage (cvg) in the Movielens 100K collection when varying the number of clusters PPC+RM1 method

methods clearly outperforms the four baselines, where these improvements are statistically significant.

Moreover, these combinations also outperform the isolated application of the two approaches – relevance modelling (RM1 and RM2) and probabilistic neighbour selection (PPC) – outperforming the results obtained for Experiments 1 and 2 in every analysed metric. Note that, in this situation, the variation in performance when using the different RM estimations – together with a neighbourhood based on PPC clustering – are negligible in the training collection but significant in the test one. Furthermore, the best value is obtained by the method PPC+RM2, which is slightly better than PPC+RM1, just the opposite to what was found in Experiment 1, for both datasets. Finally, the optimal neighbourhood size found in training was the same for both methods ($\kappa = 50$), and the performance decreases when more clusters are considered (see Figures 7.5 and 7.6 for a sensitivity analysis in the training collection). Interestingly, in this experiment the best result is obtained without affecting the coverage, an interesting effect since although a user would be isolated in a singleton neighbourhood by means of the PPC, with the RM modelling it will still benefit from the background knowledge of the collection in the recommendation process.

As an additional checking, we show in Table 7.3 how these methods are

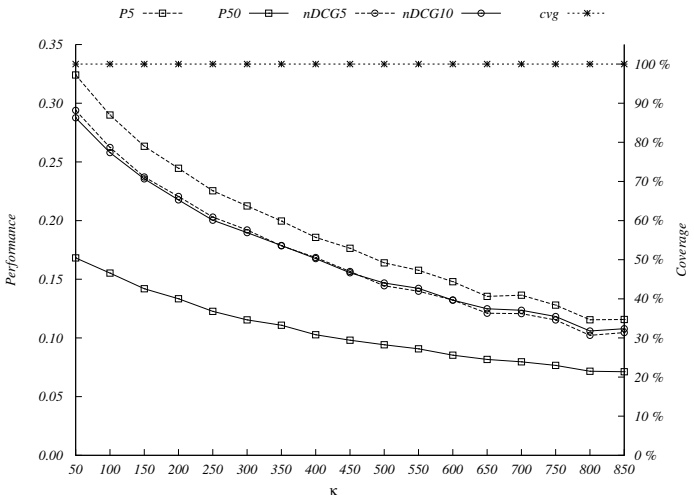

Figure 7.6: Performance and coverage (cvg) in the Movielens 100K collection when varying the number of clusters PPC+RM2 method

Table 7.3: Performance results for the combination of PPC and RM models, for P@5 and 50 clusters.

Method	λ value					
	0.0	0.1	0.2	0.3	0.4	0.5
PPC+RM1	0.006	0.316	0.318	**0.320**	0.319	0.317
PPC+RM2	0.024	0.324	0.324	0.324	0.325	0.326

Method	λ value				
	0.6	0.7	0.8	0.9	1.0
PPC+RM1	0.309	0.299	0.288	0.272	0.240
PPC+RM2	0.326	**0.327**	**0.327**	0.326	0.006

sensitive to the value of the λ parameter in the training collection. It can be observed that effectiveness values close to the ones of the best performing λ can be obtained in a wide range of the parameter space for both methods, stressing the robustness of these approaches.

7.4.4.4 Discussion

When globally analysing the results of the three experiments, we can conclude that (i) the proposed language-based relevance modelling of the recommendation process performs better than other similar approaches which also capture the relevance notion for this problem, (ii) the probabilistic clustering for neighbourhood selection clearly outperforms traditional neighbourhood selection techniques based on Pearson's correlation on nearest neighbours or matrix factorisation techniques, and (iii) the combination of both approaches enhances even further the recommendation results.

A very desirable characteristic, which is worth analysing, was presented in Experiment 1. Our relevance model estimations for the recommendation problem have the capability of, depending on the amount of smoothing applied, producing a range of different recommendation strategies, from a pure popularity-based recommendation to a (standard) neighbour-based recommender algorithm. As a result of this, in the first experiment, we obtained that the best performance of the RM1 method was produced by a popularity-based recommendation, once its optimal configuration of the parameters is analysed. We believe this fact indicates that the quality of the standard neighbourhood techniques is not good enough, and more emphasis on the collection statistics (popularity) should be taken into account. For that reason, we decided to test alternative neighbour selection techniques such as PPC.

From the results of the Experiment 2, we can clearly conclude that PPC obtains better neighbourhoods than standard techniques, in terms of the resulting recommendation performance. In fact, if we compare the results obtained by the classic UB against our method, the precision may be multiplied up to a factor of 5x in the best situation (training collection), whereas a decent improvement of 30% for P@50 has been obtained for the test collection. It is true that this improvement is achieved at the expense of lower coverage, but, as shown in Figure 7.4, even when few clusters are exploited (and, hence, coverage is still high) precision is doubled. Moreover, for the values reported the coverage is quite high, as we can observe in Table 7.2.

Finally, if we analyse the results of Experiment 3, we can observe that now the method based on RM1 is not producing solely popularity-driven recommendations for its optimal parameters, but instead a combination of both the background model and the neighbourhood information. In our opinion, this is another evidence to support the high quality of the neighbourhoods obtained by the PPC method. As another consequence, we observe in this experiment an important improvement in terms of effectiveness with respect to the results obtained in Experiments 1 and 2.

In summary, the combination of Relevance Modelling and PPC approaches leads to more robust techniques (since the sensitivity of λ has decreased and coverage is now independent of the number κ of clusters), more computationally efficient algorithms (because lower values for κ are required), and better performing techniques in terms of precision and nDCG. Moreover, since the optimal parameters found in the training collection have proved to be effective in the test collection, we may conclude that our methods are also flexible and general enough to be trained and tested in two collections with different properties while showing good performance in both situations.

7.5 Related Work

This is not the first attempt to use probabilistic modelling from Information Retrieval in Collaborative Filtering. In Wang et al. (2008c), the authors found interesting analogies between implicit CF (based on frequencies, instead of ratings) and IR, introducing the concept of binary relevance into CF and applying the Probability Ranking Principle of IR to it. Similarly, in Wang et al. (2006b) a generative relevance model is proposed for implicit CF, and in (Wang, 2009), the author made use of a language modelling formulation to propose a risk-aware ranking for (implicit) CF. However, these attempts have been partial and focused purely on the model level with regards to CF; moreover, these approaches do not tackle the problem of having ratings instead of frequencies, like we did in the work presented herein.

Regarding rating-based CF, one of the first works which explicitly dealt with a generative probabilistic framework in a rating-based collaborative filtering scenario is Wang et al. (2006a). In that work, together with (Wang et al., 2008b), the authors presented a probabilistic relevance framework, where three models are derived: one based on users, other based on items,

and a unified relevance model. This modelling was based on the probabilistic interpretation of the Relevance Models for language modelling (Lafferty and Zhai, 2002). This is a good attempt to model the recommendation problem as a relevance ranking problem. However it has a different philosophy to our approach. Wang et al.'s idea can be interpreted as an attempt to produce an initial ranking meanwhile our proposal is based on mimicking the typical query expansion process and produce the recommendation as a user profile expansion.

When comparing both user-based proposals, the evaluation shows that our methods achieve better results with respect to precision-based evaluation metrics, whereas Wang's methods perform well with error-based metrics such as Mean Average Error. One problem of Wang's method is the need of an intensive training phase (based on Expectation Maximisation) to learn the optimal parameter values (bandwidth h_u in Wang et al. (2008b)) over the whole data, while the methods proposed herein can be tuned by means of a classical cross-validation experiment. Furthermore, in our method, even parameters' values from other collections achieve a decent performance, as observed in the experiments with the second dataset.

Regarding the use of clustering for recommendation, these algorithms have not been widely exploited yet. Some authors split the set of users or items in order to improve the scalability of the recommender systems and their accuracy (O'Connor and Herlocker, 1999; Xue et al., 2005). Most of these approaches use old-fashioned clustering methods such as k-Means or hierarchical clustering, which, in general, produce good results but at the expense of lower coverage (Xue et al., 2005). Furthermore, some authors make use of external information for the data partition, such as the content of the item (e.g. genres or tags, in the film domain). Recently, we have studied in Bellogín and Parapar (2012) the use of the Normalised-Cut algorithm showing important improvements over classical clustering approaches but not reaching the performance of the presented PPC-based approach.

Although there are no Posterior Probabilistic Clustering applications in the field of recommendation, Non-negative Matrix Factorisation (NMF) methods have been previously used mainly for the rating prediction task, as a model-based recommender similar to, for instance, SVD. In Gu et al. (2010), Gu et al. proposed a unified model for collaborative filtering based on a type of non-negative matrix factorisation. While, in our case, Posterior Probabilistic Clustering is only a better performing tool for locating good neighbourhoods,

in that work the NMF algorithm also produces recommendations itself, by combining both model-based and memory-based information to improve the recommendation effectiveness. The evaluation against existing methods exhibited modest improvements in terms of Mean Average Error. Among other baselines, they compared the results with a previous work by Zhang et al. (2006) that also uses different types of NMF algorithms. The latter was a pioneering work on tackling the problem of incomplete ratings when applying recommendations based on weighted NMF, obtaining small improvements against user-based and matrix factorisation techniques, again in terms of Mean Average Error.

7.6 Conclusions

In this chapter, we have proposed a relevance modelling approach for the recommendation problem. Our proposal addresses the item recommendation task as a profile expansion problem, using the mechanisms for query expansion provided by the Relevance-Based Language Models. The empirical evaluation of our proposal shows considerable improvements in terms of effectiveness (measured by ranking quality metrics) against different related baselines. Furthermore, in order to obtain better neighbours for the memory-based recommendation, we proposed the application of the Posterior Probabilistic Clustering algorithm. This proposal by itself also achieves important effectiveness improvements over traditional neighbour-based approaches, while at the same time it outperforms standard matrix factorisation algorithms.

Furthermore, we show that the combination of both proposals improves the results of their individual application, demonstrating in this way that the better the neighbourhood (which acts as the pseudo-relevant set in the explicit search scenario), the better the estimations of the underlying relevance model, and therefore, better item recommendations are produced as expansions of the user profile. This fact is consistent with the previous results obtained in the application of RM on text retrieval.

Chapter 8

Conclusions and Future Research

In this chapter we summarise the conclusions of this thesis and suggest some future research directions.

8.1 Conclusions

This thesis presents new estimations and new application fields for the Relevance-Based Language Models. On one hand we have presented new estimations for the traditional retrieval task with pseudo-relevance feedback by affecting the selection of expansion terms (Chapter 3) with the objective of select better expansion terms, and by modifying the pseudo-relevant set over which the Relevance Models are estimated (Chapters 4 and 5) with the objective of selecting better pseudo-relevant documents. On the other hand we have adapted the Relevance Based Language Modelling Framework for other tasks such as text clustering (Chapters 6) and recommendation (Chapters 7), producing the corresponding estimations. The main objective of this work was not only improving the performance in term of effectiveness for the given tasks, but also we have been very worrying on studying the robustness of our proposals. This objective was accomplished not only in the retrieval task but also to an important extent in the clustering and recommendation tasks.

- Chapter 3 presents a new RM estimation where the idea of promot-

ing the divergent terms is introduced in the core of the RM computation. The objective is to promote those terms in the pseudo-relevant set which are divergent from the background collection model. Results show that of our estimation outperforms RM3 estimation (Abdul-jaleel et al., 2004), the best performing RM estimation to the date, both in retrieval effectiveness and query robustness. In order to provide a baseline which also incorporates the divergence idea we have also produced a formulation of the traditional KLD QE method from the Rocchio's framework in the LM framework. Interestingly, this method seems competitive with RM3 and having one parameter less.

- Chapter 4 formulates a new cluster based RM method. Taking advantage of overlapping clustering techniques we processed the initial retrieval and produced a reranking of the top documents considering the document query likelihood and the cluster query likelihoods of the best and worst ranked clusters to which the documents belongs to. With this reranking the RS is modified and RM3 is used to estimate an expanded query and to produce a second retrieval. Our proposal outperforms in both effectiveness and query robustness the traditional RM3 method. Furthermore, the proposed method outperforms the Resampling (Lee et al., 2008) method, another cluster based pseudo-relevance feedback baseline.

- Chapter 5 introduces the application of the Relevance-Based Language Modelling framework to the score distribution modelling and the threshold optimisation problems. The objective was to produce a RS with less non-relevant documents given the fact of the sensibility of the PRF method to the presence of such noise in the pseudo-relevant set. We proposed the modelling of the score distributions for our task as a mixture of two Gaussian distributions. We used Bergman soft-clustering technique to infer the distributions and mixture's parameters. And finally the threshold optimisation problem considers the location of the cut-off point were the highest relevance density is estimated. With this framework the RS size is automatically determined at query level. Results comparing our proposal with the standard approach of training the size of the RS show that our method is equivalent or better to the traditional method. Particularly our method achieves important improvements in the text collections (up to a 11% in terms of MAP and more than 38%

in terms of query robustness).

- Chapter 6 presents the adaptation of the Relevance Modelling framework to a non-retrieval task such as clustering. Particularly we addressed a kind of semi-supervised clustering task known as Constrained Clustering. Here, we introduced the application of relevance models for modelling the clustering constraints directly in the document representation. Some assumptions and definitions are introduced for applying the framework for this task. The modelling of the constraints in such forms allows the use of traditional un-constrained clustering algorithm for performing the constrained clustering task. Evaluation shows, not only that our proposal is valid for the given task but also it outperforms existing specially tailored constrained clustering methods in effectiveness and robustness to the parameter settings.

- Chapter 7 formulates two new estimations for relevance models. In this case these formulations are used in the recommendation task. The estimations are derived as equivalent to RM1 and RM2 (Lavrenko and Croft, 2001) for document retrieval. The proposed estimations outperform four different state-of-the-art recommending methods. Furthermore, we also introduced a new way of constructing the neighbourhoods of users in the recommender system. The neighbourhood of a user in our modelling play the role of the relevance set, so with this proposal we reiterate the importance of achieving a good RS for the estimation of the relevance models. The combination of both proposals achieves impressive results in terms of effectiveness (around 38% over the baselines). Furthermore, the study of parameter sensitivity and behaviour across collections shows that our methods are quite robust.

8.2 Future Research

Several research opportunities have been opened by the works presented in this thesis not only for further improve the effectiveness and robustness of the methods but also for new application fields for them.

- In the promotion of the divergence in the estimation of relevance models several research lines remain open. In particular it seems worth to try alternative ways for estimating of the background model in the

collection. Furthermore, research on more elaborated methods for introducing the demotion of those common collection terms in the RM estimates is also interesting. We also want to study how the presented ideas may be applied to improve existing techniques for selective query expansion and adaptive relevance feedback.

- Regarding to our work on cluster based relevance models, in the future, we want to test our approach in additional clustering frameworks. Particularly it seems interesting to explore probabilistic clustering methods such as the PPC algorithm (Ding et al., 2008) which provide with belonging probability of documents to clusters, and study how to integrate that information directly in the computation of the cluster query likelihood. Also we want to explore further methods for cluster representation as input to the computation of the cluster query likelihood. In line with our work in score distributions it seems appropriate considering the cluster scores in order to discern between relevant and non-relevant information.

- Score distribution modelling is not an easy field because it depends in weak assumptions on distributions choice. More work on selecting good mixtures of appropriate statistical distributions has to be carried out. In particular we have observed that the distribution fitting depends not only on the queries but also in the nature of the collection of documents. We envisage future work on automatically selecting for each situation the distribution combination that best fit with the observed data in an attempt of improving the performance of these methods.

- Referring to our work on modelling the constraints in the document representation for constrained clustering we consider two different problems to approach. We want to test our method in other clustering framework and also test other RM estimations and PRF methods. We also will study how to accommodate other kind of constraints, in particular absolute and negative constraints and check how negative PRF methods perform for this task. The combination of our proposal with other methods as a way for reinforcing the effect of the constraints is also interesting to try.

- In the field of recommender systems we plan to further study other options for the construction of the pseudo-relevant set of users, not only

techniques based on neighbours but also other approaches to produce an initial user ranking, as is standard in the retrieval task. We will also consider alternative estimations and smoothing approaches to be applied in our formulation of the problem. We envision additional refinements of our methods, such as only considering positively rated items in the user profile when computing the user likelihood, or tackling differently the absence of rating for an item by the user in the PPC algorithm. We also plan to assess the use of our proposal in order to address the important problem of recommendation diversification. We also envision the recommendation diversification in our modelling, i.e., the diversification of the recommended items in the expanded profile, as the same problem of promoting divergent terms in the estimation of the relevance models.

In accordance with the Regulations of the Ph.D. studies passed by the Governing Council of the University of A Coruña at its meeting of July 17, 2012, it is reproduced below a summary of this thesis in Spanish.

Appendix A

Resumen

Desde las formalizaciones de cómo debe llevarse a cabo el proceso de búsqueda en los archivos de las bibliotecas durante los años cincuenta, las técnicas de Recuperación de Información se han convertido en esenciales para la actividad diaria de la mayoría de los seres humanos. Hoy en día la página principal de casi todos los navegadores web instalados en los ordenadores personales apuntan a un motor de búsqueda web como Google, Yahoo! o Bing, esto no tiene solamente propósitos comerciales, sino también, y más importante aún, se debe a que hoy en día los motores de búsqueda son vitales para acceder a la información. Y estos motores de búsqueda no habrían sido posibles sin los esfuerzos de investigación realizados en el campo de la Recuperación de Información. Recuperación de Información (RI) es en realidad el textit la ciencia de la búsqueda, o quizás una mejor descripción podría ser la *la ciencia de encontrar*. Varias definiciones han sido propuestas para caracterizar esta, todavía joven, área de investigación, en nuestra opinión, una de las más precisas se produjo en Manning et al. (2008):

> *Recuperación de Información es encontrar material (normalmente documentos) de una naturaleza no estructurada (generalmente texto) que satisface una necesidad de información sobre grandes colecciones (normalmente almacenadas en los ordenadores)*

Los motores de búsqueda mencionados anteriormente son complejos sistemas de información con muchos componentes, tales como rastreadores, analizadores, *tokenizadores*, indexadores, buscadores, clasificadores, o interfaces de interacción. Pero por encima de todos ellos, los modelos de recu-

peración son los están en el centro de cualquier motor de búsqueda. Los modelos de recuperación permiten a los motores de búsqueda proporcionar, en respuesta a una necesidad de información usuario, *rankings* de documentos de manera eficaz y eficiente . A lo largo de la historia de la Recuperación de Información varios modelos de recuperación han sido propuestos. En particular, esta tesis se enmarca dentro de los bien conocidos y altamente efectivos Modelos Estadísticos del Lenguaje. Ponte and Croft (1998) introdujeron el uso de los Modelos de Lenguaje en la Recuperación de la Información, modelos cuyas raíces más cercanas se encuentran en el campo del reconocimiento automático del habla. En concreto, en este trabajo se tratará con una expansión posterior de los mismos llamada Modelos de Lenguaje basados en Relevancia (Lavrenko and Croft, 2001). Los Modelos de Relevancia introdujeron en el marco de los Modelos de Lenguaje el concepto de relevancia, concepto que es explícito en otros modelos de recuperación, como los modelos probabilísticos. Los Modelos de Relevancia se han utilizado principalmente para una tarea específica dentro de la Recuperación de Información llamada Retroalimentación por Pseudo Relevancia. Retroalimentación por Pseudo Relevancia es un tipo de técnica local de expansión de consultas donde se asume relevancia sobre un conjunto de documentos del *ranking* inicial, estos documentos se utilizan para seleccionar los términos de expansión para la consulta original y producir una segundo *ranking* más efectivo

En esta tesis investigamos algunas nuevas estimaciones de los Modelos de Relevancia tanto para la tarea de retroalimentación por pseudo relevancia como para otras tareas más allá de la búsqueda de documentos, en particular, el agrupamiento documental y recomendación. Se estudian los beneficios de nuestras propuestas para estas tareas comparándolas con las estimaciones existentes. También prestamos atención sobre algunos aspectos prácticos de los Modelos de Relevancia como el ajuste de parámetros, y especialmente una nueva propuesta para determinar automáticamente el tamaño del conjunto de los documentos seleccionados para la retroalimentación.

A.1 Motivación

La producción de un *ranking* de documentos de calidad con respecto a las necesidades de los usuarios no es todavía un tema cerrado. Aunque para algunos tipos de consultas comunes producidas por los usuarios en buscadores

web, los resultados sean satisfactorios en términos de eficacia, todavía hay mucho margen de mejora en otros escenarios en tareas más allá de la recuperación de documentos. Varias técnicas han sido exploradas para conseguir este objetivo. Entre esas técnicas, las técnicas de retroalimentación de relevancia parecen ser unas de las más prometedoras en cuanto a mejoras en la eficacia. Es comúnmente reconocido que los Modelos de Relevancia es una de las técnicas de retroalimentación con mejores resultados, a pesar de esto, no fue hasta hace poco cuando se llevaron a cabo estudios detallados sobre sus diferentes estimaciones (Lv and Zhai, 2009a). Esta es, de hecho, la principal motivación de esta tesis, producir nuevas estimaciones de los Modelos de Relevancia para la tarea de recuperación de documentos y la comparación de las mismas con los métodos estado del arte. No sólo queremos producir modelos más eficaces, sino también obtener estimaciones más robustas que hagan frente al problema endémico de la *deriva en el tópico* de las técnicas de retroalimentación de relevancia. En línea con esto queremos seguir explorando principalmente dos caminos diferentes: la promoción de los métodos basados en la divergencia y en el agrupamiento documental. En relación a lo primero, la promoción de términos divergentes ha sido previamente explorada con éxito en otros modelos de recuperación (Carpineto et al., 2001), mientras que para los Modelos de Relevancia, los intentos de hacerlo no fueron concluyentes (Zhai and Lafferty, 2001) y fueron aplicados sólo a nivel de modelo, en vez de a nivel de término. En cuanto al uso de técnicas de agrupación de documentos, los esfuerzos de investigación se han centrado en aplicar métodos de recuperación basados en agrupamiento documental para mejorar la selección del conjunto de documentos pseudo relevantes (Lee et al., 2008) y esta es la idea que se quiere explorar más a fondo en esta tesis, no sólo con la objetivo de mejorar la eficacia, sino también la robustez.

Como resultado del buen comportamiento de los Modelos de Relevancia en términos de eficacia para la recuperación de documentos, estos han sido probados para la recuperación más allá de los documentos como la recuperación de frase (Balasubramanian et al., 2007) , recuperación de pasajes (Li and Zhu, 2008) o recuperación de sentimiento y opinión (Eguchi and Lavrenko, 2006). Además, el mismo concepto de Modelos Relevancia fue aplicado con éxito para otras tareas como la anotación automática de imágenes (Jeon et al., 2003) o el etiquetado social (Lavrenko et al., 2002). Esto nos inspiró para ir un paso más allá y, de la misma manera que los modelos de lenguaje fueron reinterpretados para la tarea de recuperación trayéndolos del

campo del reconocimiento automático del habla, queremos producir estimaciones adicionales de los Modelos de Relevancia para otras tareas donde la relevancia no está asociada entre consultas y documentos. En particular, vamos a modelar el problema de recomendación y las tareas de agrupamiento documental con restricciones, dos problemas muy relacionados con la recuperación de documentos en los que hemos trabajado anteriormente en otras líneas de investigación. Los Modelos de Relevancia son en principio una forma de hacer expansión de consultas, nosotros vamos a usar este marco para otras tareas donde no es la consulta el elemento a expandir. En el problema de la recomendación los Modelos de Relevancia se estimarán sobre las preferencias del usuario y las preferencias de los vecinos del usuario, para producir la expansión de perfil del usuario, mientras que en la tarea agrupamiento documental con restricciones el Modelo de Relevancia será calculado sobre el documento de texto y los documentos con los que este documento comparte restricciones, lo que resultará en la expansión documento constreñido.

Al abordar esta investigación es imprescindible para hacer frente a algunos problemas prácticos. Específicamente, una buena elección de los valores de los parámetros es un punto que se conoce crucial en el rendimiento de las diferentes formulaciones de RM. Algunos estudios se han publicado recientemente sobre este tema (Winaver et al., 2007; Huang et al., 2008), pero la eficacia de los enfoques existentes compromete su uso en tiempo de consulta en aplicaciones reales. Debido a esto, se decidió estudiar este punto en detalle, tratando de producir métodos automáticos para determinar automáticamente los valores de algunos de estos parámetros en una forma menos costosa computacionalmente.

A.2 Objetivos

El principal alegato de este trabajo es que es posible formular nuevas estimaciones de los Modelos de Relevancia para mejorar aún más la eficacia de la tarea de búsqueda de documentos y tareas más allá de la búsqueda. Estos nuevos modelados serán capaces de no sólo mejorar la eficacia de las estimaciones y los métodos existentes, sino también de superar su robustez, un factor crítico cuando se trata de métodos de retroalimentación de relevancia. Estos objetivos se perseguirán mediante diferentes caminos: la promoción de términos divergentes en la propia estimación, presentando nuevas técnicas

basadas en el agrupamiento documental, la introducción de nuevos métodos para determinar automáticamente el tamaño del conjunto de documentos pseudo relevantes a nivel de consulta, y produciendo nuevas y originales estimaciones para los Modelos de Relevancia en tareas como la recomendación y el agrupamiento documental con restricciones.

Alineadas con los medios anteriormente indicados, las contribuciones de esta tesis son las siguientes. Tres métodos diferentes propuestos con el fin de mejorar la eficacia de la tarea de búsqueda cuando se utilizan Modelos de Relevancia. En primer lugar, un nuevo estudio sobre la forma de introducir la promoción de los términos divergentes en las estimaciones de los Modelos de Relevancia, incorporando en este marco teórico la idea intrínseca a otros modelos de retroalimentación como los métodos de expansión de consultados usando la divergencia Kullback-Leibler bajo el marco de Rocchio (Carpineto et al., 2001) o el modelo de recuperación basado en la Divergencia de la Aleatoriedad (Amati and Van Rijsbergen, 2002). En segundo lugar, presentamos un nuevo enfoque basado en el agrupamiento documental y diseñado con el objetivo de seleccionar los mejores documentos candidatos para formar el conjunto de pseudo relevantes, comparando nuestra propuesta con formas alternativas de construir el conjunto de pseudo relevantes. Y en tercer lugar, diseñamos un sistema para la selección automática del número de los documentos en el conjunto de pseudo relevantes usado para la expansión, estudiando su efectividad y las consecuencias en términos de eficiencia con respecto a los métodos existentes.

Además, contribuimos con dos nuevos modelos más allá de la tarea de búsqueda de documentos. Por un lado, un nuevo modelado de la tarea agrupamiento documental con restricciones, donde derivamos las estimaciones correspondientes para los Modelos de Relevancia con el fin de introducir la información de las restricciones directamente en la representación de los documentos. Por otra parte, formulamos el problema recomendación de elementos en el contexto de un sistema de recomendación como una estimación de los Modelos de Relevancia. Todas estas técnicas son validadas en la evidencia empírica, a través de varias series de experimentos cuidadosos, demostrado ser métodos robustos a través de diferentes colecciones.

A.3 Estructura

Las principales aportaciones de esta tesis se presentan en los capítulos 3, 4, 5, 6 y 7. El capítulo 2 contiene una introducción general a los Modelos de Relevancia que un especialista en Recuperación de Información podría evitar, pero en la que cualquier lector interesado puede encontrar una breve introducción al estado del arte.

La organización de capítulos es el siguiente [1]:

- El capítulo 2 es un breve resumen de los principales conceptos de Recuperación de Información, Retroalimentación de Pseudo Relevancia y, sobre todo, Modelos de Relevancia. Las diferentes etapas de cómo se ha abordado esta tarea tradicionalmente en un proceso de búsqueda son aclarados y particularmente como estas se encuadran en el marco teórico específico asociado. Se presenta una breve revisión de la literatura en términos de técnicas de retroalimentación y diferentes estimaciones de los Modelos de Relevancia el objetivo de proporcionar una visión clara de las alternativas existentes a las técnicas que se proponen en este trabajo.

- El capítulo 3 presenta el primero de nuestros esfuerzos para producir mejores estimaciones de los Modelos de Relevancia, en este caso basado en la promoción de los términos divergentes cuando se calcula el Modelo de Relevancia para una consulta determinada. Una evaluación detallada contra la formulación estándar y otras técnicas de expansión no basadas en Modelos de Relevancia avala la eficacia y los valores de robustez de nuestra propuesta.

- El capítulo 4 presenta la propuesta para modificar la selección de los documentos que pertenecen al conjunto de pseudo relevantes mediante la aplicación de estrategias basadas en búsqueda sobre agrupamientos de documentos. Nuestro objetivo no es sólo explotar en este proceso la información grupos *buenos* de documentos, sino también de grupos *malos*. La evaluación se realiza contra los de los esfuerzos anteriores en esta línea y se informa de los resultados comparando la eficacia y las cifras de robustez.

[1] Siguiendo las recomendaciones de Evans et al. (2012) decidimos mantener independientes y auto-contenidos los capítulos, exponiendo en cada uno la literatura y el estado del arte específico del tema de cada capítulo. También seguimos las recomendaciones en términos de formato y estilo

- El capítulo 5 describe nuestros intentos de proporcionar de un método automático para determinar, a nivel de consulta, el número de documentos seleccionados del *top* de la primera recuperación para formar parte del conjunto de pseudo relevantes. Nuestra propuesta, que se basa en el estudio de las distribuciones de las puntuaciones de los documentos, trata de mejorar la eficacia pero, sobre todo, la robustez de los Modelos de Relevancia. Estudiamos cómo nuestro método se compara con otras técnicas anteriores comentando sus ventajas.

- El capítulo 6 comienza con nuestras propuestas para el modelado de nuevas tareas en el marco de los Modelos de Relevancia. En este caso nos acercamos a la tarea de agrupamiento de texto con restricciones. En este capítulo se presenta cómo acomodar las restricciones directamente en la representación de documentos, evitando el uso de algoritmos de agrupamiento especialmente creados para esta tarea y consiguiendo valores comparables de eficacia e incluso mejores que dichos métodos.

- El capítulo 7 examina otro problema más allá de la recuperación de documentos: recomendación de elementos. Se propone una formulación alternativa de los métodos de filtrado colaborativo encuadrada en el marco de los Modelos de Relevancia, se modela el problema de recomendación de elementos como un problema de expansión de perfil y se formulan las estimaciones correspondientes para los Modelos de Relevancia asociados. Se lleva a cabo una comparación detallada con los métodos de recomendación estado del arte mostrando impresionantes mejoras para las diferentes métricas de eficacia.

- Por último, en el capítulo 8 se presentan las conclusiones de la tesis y un resumen de las líneas de investigación futuras.

A.4 Publicaciones

A.4.1 Publicaciones Recientes en Congresos de Referencia

- Alejandro Bellogín, Javier Parapar. Using Graph Partitioning Techniques for Neighbour Selection in User-Based Collaborative Filtering. In Proceedings of the 6th ACM Conference on Recommender Systems, RECSYS 2012, Dublin, Ireland, September 9-13, 2012. (s.p., ratio aceptación:

31 %, cualificación: CORE ERA 2008: B) - **Best s.p. Award**

- Javier Parapar, Alvaro Barreiro. Language Modelling of Constraints for Text Clustering, Proceedings of the 34th European Conference on Information Retrieval Research ECIR 2012, Barcelona, Spain, 1-5 April 2012, Lecture Notes in Computer Science vol. 7224, pp. 352-363, 2012 (ISBN:978-3-642-28996-5) (f.p., ratio aceptación 21 % (35/163), cualificación: CORE ERA 2008: B)

- Javier Parapar, María M. Vidal, José Santos. Finding the Best Parameter Setting: Particle Swarm Optimisation. Proceedings of the 2nd Spanish Conference on Information Retrieval, CERI'12, Valencia, Spain, June 18-19, pp. 49-60, 2012, (ISBN 978-84-8021-860-3)

- Javier Parapar, Alvaro Barreiro. A Cluster Based Pseudo Feedback Technique which Exploits Good and Bad Clusters. Proceedings of the 14th Conference of the Spanish Association for Artificial Intelligence, CAEPIA 2011, 7-11 November 2011, Tenerife, Spain, Lecture Notes in Artificial Intelligence 7023, pp 403-412, 2011 (ISSN 0302-9743) (ISBN 978-3-642-25273-0)(f.p., ratio aceptación 33 % (50/149))

- Javier Parapar, Alvaro Barreiro. Promoting Divergent Terms in the Estimation of Relevance Models, Proceedings of the Third International Conference on the Theory of Information Retrieval, ICTIR 2011, 12-14 September 2011, Bertinoro, Italy, Lecture Notes in Computer Science vol. 6931, pp. 77-88, 2011. (ISSN 0302-9743) (ISBN 978-3-642-23318-0) (f.p., ratio aceptación 38 % (25/65))

- M. Eduardo Ares, Javier Parapar, Alvaro Barreiro. Improving Text Clustering with Social Tagging, Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media (ICWSM 2011), pp. 430-433, 17-21 July 2011, Barcelona, Spain.(s.p. ratio de aceptación: 23 % f.p. + 18 % s.p)

- M. Eduardo Ares, Javier Parapar, Alvaro Barreiro. Improving Alternative Text Clustering Quality in the Avoiding Bias Task with Spectral and Flat Partition Algorithms, Proceedings of the 21st International Conference on Database and Expert Systems Applications DEXA'10, Bilbao, Spain, August 30 - September 3 2010, Lecture Notes in Computer Science vol. 6262, Part II, pp. 407-421, 2010. (ISBN 978-3-642-15250-4) (f.p., ratio aceptación 22,8 %, cualificación: CORE ERA 2008: B)

- Renato de Freitas Bulcão-Neto, José Antonio Camacho-Guerrero, Alvaro Barreiro, Javier Parapar, Alessandra Alaniz Macedo. An automatic linking service of document images reducing the effects of OCR errors with latent semantics. Proceedings of the ACM Symposium on Applied Computing SAC 2010, pp. 13-17. (ISBN 978-160558-638-0) (f.p., ratio de aceptación 26,5 %, cualificación: CORE ERA 2008: B)

- Javier Parapar, Jorge-López-Castro, Alvaro Barreiro. Blog Snippets: A Comments-Biased Approach, Proceedings of the 33rd ACM International Conference on Research and Development in Information Retrieval SIGIR'10, Geneva, Switzerland, July 19-23, pp. 711-712, 2010. (p.p. ratio aceptación: 32 %, cualificación CORE ERA 2008: A*)

- Ronald T. Fernández , Javier Parapar, David E. Losada, Alvaro Barreiro. Where to Start Filtering Redundancy? A Cluster-Based Approach, Proceedings of the 33rd ACM International Conference on Research and Development in Information Retrieval SIGIR'10, Geneva, Switzerland, July 19-23, pp. 735-736, 2010.(p.p. ratio aceptación: 32 %, cualificación CORE ERA 2008: A*)

- Javier Parapar, Jorge-López-Castro, Álvaro Barreiro. Blog Posts and Comments Extraction and Impact on Retrieval Effectiveness, Proceedings of the 1st Spanish Conference on Information Retrieval CERI'10, Madrid, Spain, June 15-17, pp. 5-16, 2010.

- E. Ares, J. Parapar, A. Barreiro. Avoiding Bias in Text Clustering Using Constrained K-means and May-Not-Links, Proceedings of the 2nd International Conference on the Theory of Information Retrieval ICTIR

2009, Cambridge, UK, September 10-12, 2009, Lecture Notes in Computer Science vol. 5766, pp. 322-329, 2009. ISBN: 978-3-642-04416-8. Short paper (s.p. ratio de aceptación: 22 % f.p. + 17 % s.p)

- J. Parapar, A. Barreiro. Evaluation of text clustering algorithms with n-gram-based document fingerprints, Proceedings of the 31st European Conference on Information Retrieval Research ECIR 2009, Toulouse, France, April 2009, Lecture Notes in Computer Science vol. 5478, pp. 645-653, 2009. ISBN: 978-3-642-00957-0. (s.p. ratio de aceptación: 22 % , cualificación: CORE ERA 2008: B)

- J. Parapar, A. Freire, A. Barreiro. Revisiting n-gram based models for retrieval in degraded large collections, Proceedings of the 31st European Conference on Information Retrieval Research ECIR 2009, Toulouse, France, April 2009, Lecture Notes in Computer Science vol. 5478, pp. 680-684, 2009. ISBN: 978-3-642-00957-0. (p.p. ratio de aceptación: 41 %, cualificación: CORE ERA 2008: B)

- J. Parapar, D. Losada, A. Barreiro. Compression-based Document Length Prior for Language Models.Poster session of the 32nd ACM International Conference on Research and Development in Information Retrieval, SIGIR 2009, Boston (USA), Jul 2009. (p.p. ratio aceptación: 33 %, cualificación CORE ERA 2008: A*)

- J. Parapar, A. Barreiro. Winnowing-based text clustering, Proceedings of 17th ACM Conference on Information and Knowledge Management CIKM 2008, pp. 1353-1354, Napa Valley, California, October 2008, ISBN: 978-1-59593-991-3. (p.p. ratio aceptación: 17 % f.p + 12 % p.p., cualificación CORE ERA 2008: A)

A.4.2 Publicaciones Recientes en Revistas JCR

- M. Eduardo Ares, Javier Parapar, Alvaro Barreiro. An Experimental Study of Constrained Clustering Effectiveness in Presence of Erroneous Con-

straints. Information Processing and Management 48(3) pp. 537-551, 2012.

- Ismael Hasan, Javier Parapar, Alvaro Barreiro. Improving the Extraction of Text in PDFs by Simulating the Human Reading Order, Journal of Universal Computer Science, vol. 18, no. 5, pp. 623-649, 2012.

- Renato de Freitas Bulcão-Neto, José Antonio Camacho-Guerrero, Márcio Dutra, Álvaro Barreiro, Javier Parapar, Alessandra Alaniz Macedo. The use of latent semantic indexing to mitigate OCR effects of related document images. Journal of Universal Computer Science, vol. 17, no. 1, pp. 64-80, 2011.

Bibliography

Abdul-jaleel, N., Allan, J., Croft, W. B., Diaz, O., Larkey, L., Li, X., Smucker, M. D., and Wade, C. (2004). UMass at TREC 2004: Novelty and HARD. In *Proceedings of TREC-13*, NIST Special Publication. National Institute for Science and Technology.

Adomavicius, G. and Tuzhilin, A. (2005). Toward the Next Generation of Recommender Systems: A Survey of the State-of-the-Art and Possible Extensions. *IEEE Trans. on Knowl. and Data Eng.*, 17(6):734–749.

Amati, G., Carpineto, C., and Romano, G. (2004). Query Difficulty, Robustness and Selective Application of Query Expansion. In McDonald, S. and Tait, J., editors, *Proceedings of the 26th European conference on Advances in Information Retrieval*, volume 2997 of *ECIR'04*, pages 127–137, Berlin, Heidelberg. Springer Berlin Heidelberg.

Amati, G. and Van Rijsbergen, C. J. (2002). Probabilistic models of information retrieval based on measuring the divergence from randomness. *ACM Trans. Inf. Syst.*, 20(4):357–389.

Arampatzis, A., Beney, J., Koster, C. H. A., and van der Weide, T. P. (2000). Incrementality, Half-life, and Threshold Optimization for Adaptive Document Filtering. In *Proceedings of TREC-9*, NIST Special Publication. National Institute for Science and Technology.

Arampatzis, A. and Kamps, J. (2008). A study of query length. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '08, pages 811–812, New York, NY, USA. ACM.

Arampatzis, A., Kamps, J., and Robertson, S. E. (2009). Where to stop reading a ranked list? In *Proceedings of the 32nd international ACM SIGIR conference*

on Research and development in information retrieval, SIGIR'09, pages 524–531, New York, New York, USA. ACM Press.

Ares, M. E., Parapar, J., and Barreiro, A. (2009). Avoiding Bias in Text Clustering Using Constrained K-means and May-Not-Links. In *Proceedings of the 2nd International Conference on Theory of Information Retrieval: Advances in Information Retrieval Theory*, ICTIR '09, pages 322–329, Berlin, Heidelberg. Springer-Verlag.

Ares, M. E., Parapar, J., and Barreiro, A. (2010). Improving alternative text clustering quality in the avoiding bias task with spectral and flat partition algorithms. In *Proceedings of the 21st international conference on Database and expert systems applications: Part II*, DEXA'10, Berlin, Heidelberg. Springer-Verlag.

Ares, M. E., Parapar, J., and Barreiro, A. (2011). Improving text clustering with social tagging. In *Proceedings of the 5th International AAAI Conference on Weblogs and Social Media*, ICWSM '11, pages 430–433, New York, NY, USA. AAAI Press.

Ares, M. E., Parapar, J., and Barreiro, A. (2012). An experimental study of constrained clustering effectiveness in presence of erroneous constraints. *Inf. Process. Manage.*, 48(3):537–551.

Bae, E. and Bailey, J. (2006). Coala: A novel approach for the extraction of an alternate clustering of high quality and high dissimilarity. In *Proceedings of the Sixth International Conference on Data Mining*, ICDM '06, pages 53–62, Washington, DC, USA. IEEE Computer Society.

Balasubramanian, N., Allan, J., and Croft, W. B. (2007). A comparison of sentence retrieval techniques. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '07, pages 813–814, New York, NY, USA. ACM.

Banerjee, A., Merugu, S., Dhillon, I. S., and Ghosh, J. (2005). Clustering with Bregman Divergences. *J. Mach. Learn. Res.*, 6:1705–1749.

Basu, S., Bilenko, M., and Mooney, R. J. (2004). A probabilistic framework for semi-supervised clustering. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '04, pages 59–68, New York, NY, USA. ACM.

Basu, S., Davidson, I., and Wagstaff, K. (2008). *Constrained Clustering: Advances in Algorithms, Theory, and Applications*. Chapman & Hall/CRC.

Baumgarten, C. (1999). A probabilistic solution to the selection and fusion problem in distributed information retrieval. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '99, pages 246–253, New York, NY, USA. ACM.

Belkin, N., Oddy, R., and Brooks, H. (1982). ASK for Information Retrieval: Part I. Background and Theory. *Journal of Documentation*, 38(2):61–71.

Bellogín, A., Castells, P., and Cantador, I. (2011a). Precision-oriented evaluation of recommender systems: an algorithmic comparison. In *Proceedings of the fifth ACM conference on Recommender systems*, RecSys '11, pages 333–336, New York, NY, USA. ACM.

Bellogín, A. and Parapar, J. (2012). Using graph partitioning techniques for neighbour selection in user-based collaborative filtering. In *Proceedings of the sixth ACM conference on Recommender systems*, RecSys '12, pages 213–216, New York, NY, USA. ACM.

Bellogín, A., Wang, J., and Castells, P. (2011b). Structured collaborative filtering. In *Proceedings of the 20th ACM international conference on Information and knowledge management*, CIKM '11, pages 2257–2260, New York, NY, USA. ACM.

Bellogín, A., Wang, J., and Castells, P. (2011). Text retrieval methods for item ranking in collaborative filtering. In *Proceedings of the 33rd European conference on Advances in Information Retrieval*, ECIR'11, pages 301–306, Berlin, Heidelberg. Springer-Verlag.

Blanco, R. (2008). *Index compression for information retrieval systems*. PhD thesis, University of A Coruña, A Coruña, Spain.

Blanco, R. and Barreiro, A. (2008). Probabilistic document length priors for language models. In *Proceedings of the 30th European conference on Advances in Information Retrieval*, ECIR'08, pages 394–405, Berlin, Heidelberg. Springer-Verlag.

Bookstein, A. (1977). When the most "pertinent" document should not be retrieved - An analysis of the Swets model. *Information Processing & Management*, 13(6):377–383.

Carmel, D., Yom-Tov, E., Darlow, A., and Pelleg, D. (2006). What makes a query difficult? In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '06, page 390, New York, New York, USA. ACM Press.

Carpineto, C., de Mori, R., Romano, G., and Bigi, B. (2001). An information-theoretic approach to automatic query expansion. *ACM Trans. Inf. Syst.*, 19(1):1–27.

Carpineto, C. and Romano, G. (2012). A Survey of Automatic Query Expansion in Information Retrieval. *ACM Comput. Surv.*, 44(1):1:1—-1:50.

Collins-Thompson, K. and Callan, J. (2007). 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. ACM.

Conover, W. J. (1971). *Practical nonparametric statistics*. John Wiley & Sons, New York, third edition.

Croft, W. B., Metzler, D., and Strohman, T. (2009). *Search Engines: Information Retrieval in Practice*. Addison-Wesley Publishing Company, USA, 1st edition.

Cronen-Townsend, S., Zhou, Y., and Croft, W. B. (2002). Predicting query performance. In *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '02, page 299, New York, New York, USA. ACM Press.

Cutting, D. R., Karger, D. R., Pedersen, J. O., and Tukey, J. W. (1992). Scatter/Gather: a cluster-based approach to browsing large document collections. In *Proceedings of the 15th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR'92, pages 318–329, New York, NY, USA. ACM Press.

Dai, K., Pavlu, V., Kanoulas, E., and Aslam, J. A. (2012). Extended expectation maximization for inferring score distributions. In *Proceedings of the 34th European conference on Advances in Information Retrieval*, ECIR'12, pages 293–304, Berlin, Heidelberg. Springer-Verlag.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society*, 39(1):1–38.

Desrosiers, C. and Karypis, G. (2011). A comprehensive survey of neighborhood-based recommendation methods. In Ricci et al. (2011), pages 107–144.

Ding, C., Li, T., Luo, D., and Peng, W. (2008). Posterior probabilistic clustering using NMF. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '08, pages 831–832, New York, NY, USA. ACM.

Doszkocs, T. E. (1978). ID, an associative interactive dictionary for online searching. *Online Review*, 2(2):163–173.

Eguchi, K. and Lavrenko, V. (2006). Sentiment retrieval using generative models. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, EMNLP '06, pages 345–354, Stroudsburg, PA, USA. Association for Computational Linguistics.

Evans, D., Gruba, P., and Zobel, J. (2012). *How to Write a Better Thesis*. Melbourne University Publishing.

Fernandez, R. T., Parapar, J., Losada, D. E., and Barreiro, A. (2010). Where to start filtering redundancy?: a cluster-based approach. In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '10, pages 735–736, New York, NY, USA. ACM.

Fuhr, N. (1992). Probabilistic models in information retrieval. *Comput. J.*, 35(3):243–255.

Garcia, V. and Nielsen, F. (2010). Simplification and hierarchical representations of mixtures of exponential families. *Signal Processing*, 90(12):3197–3212.

Gu, Q., Zhou, J., and Ding, C. H. Q. (2010). Collaborative Filtering: Weighted Nonnegative Matrix Factorization Incorporating User and Item Graphs. In *Proceedings of the SIAM Conference on Data Mining*, SDM 2010, pages 199–210.

He, B. and Ounis, I. (2007). Combining fields for query expansion and adaptive query expansion. *Inf. Process. Manage.*, 43(5):1294–1307.

Hearst, M. A. and Pedersen, J. O. (1996). Reexamining the cluster hypothesis: scatter/gather on retrieval results. In *Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR'96, pages 76–84, New York, NY, USA. ACM Press.

Herlocker, J. L., Konstan, J. A., and Riedl, J. (2002). An Empirical Analysis of Design Choices in Neighborhood-Based Collaborative Filtering Algorithms. *Information Retrieval*, 5(4):287–310.

Hill, W., Stead, L., Rosenstein, M., and Furnas, G. (1995). Recommending and evaluating choices in a virtual community of use. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, CHI '95, pages 194–201, New York, NY, USA. ACM Press/Addison-Wesley Publishing Co.

Hitwise, E. C. (2011). News release: Experian Hitwise reports Bing-powered share of searches at 29 percent in October 2011. Technical report, Hitwise, a Experiean Company.

Huang, Q., Song, D., and Rüger, S. (2008). Robust query-specific pseudo feedback document selection for query expansion. In *Proceedings of the 30th European conference on Advances in Information Retrieval*, ECIR'08, pages 547–554, Berlin, Heidelberg. Springer-Verlag.

Hubert, L. and Arabie, P. (1985). Comparing partitions. *Journal of Classification*, 2(1):193–218.

Ide, E. (1971). New experiments in relevance feedback. In Salton, G., editor, *The SMART Retrieval System: Experiments in Automatic Document Processing*, pages 337–354. Prentice Hall, Inc.

Järvelin, K. and Kekäläinen, J. (2002). Cumulated gain-based evaluation of ir techniques. *ACM Trans. Inf. Syst.*, 20(4):422–446.

Jeon, J., Lavrenko, V., and Manmatha, R. (2003). Automatic image annotation and retrieval using cross-media relevance models. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, SIGIR '03, pages 119–126, New York, NY, USA. ACM.

Ji, X., Xu, W., and Zhu, S. (2006). Document clustering with prior knowledge. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR'06, pages 405–412. ACM Press.

Jin, R., Ding, C., and Kang, F. (2005). A Probabilistic Approach for Optimizing Spectral Clustering. In *Advances in Neural Information Processing Systems 18*, NIPS 2005.

Kanoulas, E., Dai, K., Pavlu, V., and Aslam, J. A. (2010). Score distribution models: assumptions, intuition, and robustness to score manipulation. In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '10, pages 242–249, New York, NY, USA. ACM.

Kanoulas, E., Pavlu, V., Dai, K., and Aslam, J. A. (2009). Modeling the score distributions of relevant and non-relevant documents. In *Proceedings of the 2nd International Conference on Theory of Information Retrieval: Advances in Information Retrieval Theory*, ICTIR '09, pages 152–163, Berlin, Heidelberg. Springer-Verlag.

Klein, D., Kamvar, S. D., and Manning, C. D. (2002). From instance-level constraints to space-level constraints: Making the most of prior knowledge in data clustering. In *Proceedings of the Nineteenth International Conference on Machine Learning*, ICML '02, pages 307–314, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Koren, Y. (2008). Factorization meets the neighborhood: a multifaceted collaborative filtering model. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '08, pages 426–434, New York, NY, USA. ACM.

Kurland, O. (2008). The opposite of smoothing: a language model approach to ranking query-specific document clusters. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR'08, pages 171–178, New York, NY, USA. ACM.

Kurland, O. (2009). Re-ranking search results using language models of query-specific clusters. *Inf. Retr.*, 12(4):437–460.

Kurland, O. and Domshlak, C. (2008). A rank-aggregation approach to searching for optimal query-specific clusters. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR'08, pages 547–554, New York, NY, USA. ACM.

Kurland, O. and Lee, L. (2004). Corpus structure, language models, and ad hoc information retrieval. In *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR'04, pages 194–201.

Lafferty, J. and Zhai, C. (2002). Probabilistic relevance models based on document and query generation. In *Language Modeling and Information Retrieval*. Kluwer Academic Publishers.

Lancaster, F. W. and Fayen, E. G. (1973). *Information Retrieval On-Line*. Melville Publishing Company, Los Angeles, California.

Lavrenko, V., Allan, J., DeGuzman, E., LaFlamme, D., Pollard, V., and Thomas, S. (2002). Relevance models for topic detection and tracking. In *Proceedings of the second international conference on Human Language Technology Research*, HLT '02, pages 115–121, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Lavrenko, V. and Croft, W. B. (2001). Relevance based language models. In *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR'01, pages 120–127, New York, NY, USA. ACM.

Lee, D. D. and Seung, H. S. (2001). Algorithms for Non-negative Matrix Factorization. In Leen, T. K., Dietterich, T. G., and Tresp, V., editors, *Advances in Neural Information Processing Systems 13*, pages 556–562. MIT Press.

Lee, K. S., Croft, W. B., and Allan, J. (2008). 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. ACM.

Li, X. (2008). A new robust relevance model in the language model framework. *Information Processing & Management*, 44(3):991–1007.

Li, X. and Zhu, Z. (2008). Enhancing relevance models with adaptive passage retrieval. In *Proceedings of the 30th European conference on Advances in Information Retrieval*, ECIR'08, pages 463–471, Berlin, Heidelberg. Springer-Verlag.

Liu, Q., Lu, H., and Ma, S. (2004). Improving kernel Fisher discriminant analysis for face recognition. *Circuits and Systems for Video Technology, IEEE Transactions on*, 14(1):42–49.

Liu, X. and Croft, W. B. (2004). Cluster-based retrieval using language models. In *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR'04, pages 186–193, New York, NY, USA. ACM Press.

Liu, X. and Croft, W. B. (2008). Evaluating Text Representations for Retrieval of the Best Group of Documents. In *Proceedings of the 30th European conference on Advances in Information Retrieval*, ECIR '08, pages 454–462.

Losada, D. E. and Azzopardi, L. (2008). An analysis on document length retrieval trends in language modeling smoothing. *Inf. Retr.*, 11(2):109–138.

Lu, X. A., Ayoub, M., and Dong, J. (1996). Ad Hoc Experiments Using EUREKA. In *Proceedings of TREC-5*, NIST Special Publication, pages 229–240. National Institute for Science and Technology.

Lv, Y. and Zhai, C. (2009a). A comparative study of methods for estimating query language models with pseudo feedback. In *Proceeding of the 18th ACM conference on Information and knowledge management*, CIKM '09, pages 1895–1898, New York, NY, USA. ACM.

Lv, Y. and Zhai, C. (2009b). 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. ACM.

MacKay, D. J. C. and Peto, L. C. B. (1994). A Hierarchical Dirichlet Language Model. *Natural Language Engineering*, 1:1–19.

MacQueen, J. B. and McQueen, J. (1967). Some Methods for Classification and Analysis of Multivariate Observations. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, 1:281–297.

Manmatha, R., Rath, T., and Feng, F. (2001). Modeling score distributions for combining the outputs of search engines. In *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR'01, pages 267–275, New York, New York, USA. ACM Press.

Manning, C. D., Raghavan, P., and Schtze, H. (2008). *Introduction to Information Retrieval*. Cambridge University Press, New York, NY, USA.

Maron, M. E. and Kuhns, J. L. (1960). On Relevance, Probabilistic Indexing and Information Retrieval. *Journal of the ACM*, 7(3):216–244.

McLaughlin, M. R. and Herlocker, J. L. (2004). A collaborative filtering algorithm and evaluation metric that accurately model the user experience. In *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '04, pages 329–336, New York, NY, USA. ACM.

Mei, Q., Fang, H., and Zhai, C. (2007). A study of Poisson query generation model for information retrieval. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR'07, page 319, New York, New York, USA. ACM Press.

Mitra, M., Singhal, A., and Buckley, C. (1998). Improving automatic query expansion. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '98, pages 206–214, New York, New York, USA. ACM Press.

Nielsen, F. and Garcia, V. (2009). Statistical exponential families: A digest with flash cards. *CoRR*, abs/0911.4863.

O'Connor, M. and Herlocker, J. (1999). Clustering Items for Collaborative Filtering. In *ACM SIGIR Workshop on Recommender Systems*.

Parapar, J. and Barreiro, A. (2011a). A cluster based pseudo feedback technique which exploits good and bad clusters. In *Proceedings of the 14th international conference on Advances in Artificial Intelligence: Spanish Association for Artificial Intelligence*, CAEPIA'11, pages 403–412, Berlin, Heidelberg. Springer-Verlag.

Parapar, J. and Barreiro, A. (2011b). Promoting Divergent Terms in the Estimation of Relevance Models. In *Proceedings of 3rd International Conference on the Theory of Information Retrieval*, ICTIR'11, pages 77–88, Berlin, Heidelberg. Springer-Verlag.

Parapar, J. and Barreiro, A. (2012). Language Modelling of Constraints for Text Clustering. In *Proceedings of the 34th European conference on Advances in Information Retrieval*, ECIR'12, pages 352–363, Berlin, Heidelberg. Springer-Verlag.

Parapar, J., Bellogín, A., Castells, P., and Barreiro, A. (2013). Relevance-Based Language Modelling for Recommender Systems. *Inf. Process. Manage.*, (Under Review).

Parapar, J., Losada, D. E., and Barreiro, A. (2009). Compression-based document length prior for language models. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, SIGIR'09, page 652, New York, New York, USA. ACM Press.

Ponte, J. M. and Croft, W. B. (1998). A language modeling approach to information retrieval. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '98, pages 275–281, New York, NY, USA. ACM.

Rand, W. (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, 66(336):846–850.

Resnick, P., Iacovou, N., Suchak, M., Bergstrom, P., and Riedl, J. (1994). Grouplens: An open architecture for collaborative filtering of netnews. In *CSCW*, pages 175–186.

Ricci, F., Rokach, L., Shapira, B., and Kantor, P. B., editors (2011). *Recommender Systems Handbook*. Springer.

Robertson, S. (2007). On score distributions and relevance. In *Proceedings of the 29th European conference on Advances in Information Retrieval*, ECIR'07, pages 40–51, Berlin, Heidelberg. Springer-Verlag.

Robertson, S. E. (1991). On term selection for query expansion. *J. Doc.*, 46(4):359–364.

Robertson, S. E. (1997). The probability ranking principle in IR. In Sparck Jones, K. and Willett, P., editors, *Readings in information retrieval*, pages 281–286. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

Robertson, S. E. and Sparck Jones, K. (1976). Relevance weighting of search terms. *Journal of the American Society of Information Science*, 27(3):129–146.

Rocchio, J. (1971). Relevance feedback in information retrieval. In Salton, G., editor, *The SMART Retrieval System: Experiments in Automatic Document Processing*, pages 313–323. Prentice Hall, Inc.

Sakai, T., Manabe, T., and Koyama, M. (2005). Flexible pseudo-relevance feedback via selective sampling. *ACM Transactions on Asian Language Information Processing (TALIP)*, 4(2):111–135.

Sakai, T. and Robertson, S. E. (2001). Flexible pseudo-relevance feedback using optimization tables. In *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '01, pages 396–397, New York, New York, USA. ACM Press.

Salton, G. (1971). *The SMART Retrieval System: Experiments in Automatic Document Processing*. Prentice-Hall, Inc., Upper Saddle River, NJ, USA.

Salton, G. (1989). *Automatic text processing: the transformation, analysis, and retrieval of information by computer*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA.

Salton, G. and Buckley, C. (1990). Improving retrieval performance by relevance feedback. *Journal of the American Society for Information Science*, 41:288–297.

Salton, G., Wong, A., and Yang, C. S. (1975). A vector space model for automatic indexing. *Commun. ACM*, 18(11):613–620.

Shani, G. and Gunawardana, A. (2011). Evaluating recommendation systems. In Ricci et al. (2011), pages 257–297.

Shi, J. and Malik, J. (2000). Normalized Cuts and Image Segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 22(8):888–905.

Shtok, A., Kurland, O., and Carmel, D. (2009). Predicting Query Performance by Query-Drift Estiamation. In *Proceedings of the 2nd International Conference on Theory of Information Retrieval: Advances in Information Retrieval Theory*, volume 5766 of *ICTIR '09*, pages 305–312, Berlin, Heidelberg. Springer.

Singhal, A. and Pereira, F. (1999). Document expansion for speech retrieval. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '99, pages 34–41, New York, NY, USA. ACM.

Song, F. and Croft, W. B. (1999). A general language model for information retrieval. In *Proceedings of the eighth international conference on Information and knowledge management*, CIKM '99, pages 316–321, New York, NY, USA. ACM.

Swets, J. A. (1963). Information retrieval systems. *Science*, 141(3577):245–250.

Swets, J. A. (1969). Effectiveness of information retrieval methods. *American Documentation*, 20(1):72–89.

Tao, T., Wang, X., Mei, Q., and Zhai, C. (2006). Language model information retrieval with document expansion. In *Proceedings of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*, HLT-NAACL '06, pages 407–414, Stroudsburg, PA, USA. Association for Computational Linguistics.

Tao, T. and Zhai, C. (2006). 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. ACM.

Thijs Westerveld, Wessel Kraaij, and Djoerd Hiemstra (2002). Retrieving Web Pages using Content, Links, URLs and Anchors. In E.M. Voorhees and D.K. Harman, editors, *Proceedings of TREC-10*, volume SP 500-25 of *NIST Special Publication*, pages 663–672, Gaithersburg. National Institute for Science and Technology.

Tombros, A., Villa, R., and Van Rijsbergen, C. J. (2002). The effectiveness of query-specific hierarchic clustering in information retrieval. *Inf. Process. Manage.*, 38(4):559–582.

Van Rijsbergen, C. J. (1979). *Information Retrieval*. Butterworths, London.

Voorhees, E. M. (1985). The cluster hypothesis revisited. In *Proceedings of the 8th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR'85, pages 188–196, New York, NY, USA. ACM Press.

Wackerly, D., Mendenhall, W., and Scheaffer, R. (2008). *Mathematical Statistics With Applications*. Thomson, Brooks/Cole.

Wagstaff, K. and Cardie, C. (2000). Clustering with Instance-level Constraints. In *Proceedings of the Seventeenth International Conference on Machine Learning*, ICML '00, pages 1103–1110. Morgan Kaufmann Publishers Inc.

Wagstaff, K., Cardie, C., Rogers, S., and Schrödl, S. (2001). Constrained K-means Clustering with Background Knowledge. In *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01, pages 577–584. Morgan Kaufmann Publishers Inc.

Wang, F., Li, T., and Zhang, C. (2008a). Semi-Supervised Clustering via Matrix Factorization. In *SDM 2008*, pages 1–12. Proceedings of the SIAM Conference on Data Mining.

Wang, J. (2009). Language Models of Collaborative Filtering. In *Proceedings of the 5th Asia Information Retrieval Symposium on Information Retrieval Technology*, AIRS '09, pages 218–229, Berlin, Heidelberg. Springer-Verlag.

Wang, J., de Vries, A. P., and Reinders, M. J. T. (2006a). Unifying user-based and item-based collaborative filtering approaches by similarity fusion. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '06, pages 501–508, New York, NY, USA. ACM.

Wang, J., de Vries, A. P., and Reinders, M. J. T. (2006b). A user-item relevance model for log-based collaborative filtering. In *Proceedings of the 28th European conference on Advances in Information Retrieval*, ECIR'06, pages 37–48, Berlin, Heidelberg. Springer-Verlag.

Wang, J., de Vries, A. P., and Reinders, M. J. T. (2008b). Unified relevance models for rating prediction in collaborative filtering. *ACM Trans. Inf. Syst.*, 26(3):16:1—-16:42.

Wang, J., Robertson, S. E., Vries, A. P., and Reinders, M. J. (2008c). Probabilistic relevance ranking for collaborative filtering. *Information Retrieval*, 11(6):477–497.

Wang, X., Fang, H., and Zhai, C. (2007). Improve retrieval accuracy for difficult queries using negative feedback. In *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, CIKM '07, pages 991–994, New York, NY, USA. ACM.

Wang, X., Fang, H., and Zhai, C. (2008d). A study of methods for negative relevance feedback. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR'08, pages 219–226, New York, NY, USA. ACM.

Wilcoxon, F. (1945). Individual Comparisons by Ranking Methods. *Biometrics Bulletin*, 1(6):80–83.

Winaver, M., Kurland, O., and Domshlak, C. (2007). Towards robust query expansion: model selection in the language modeling framework. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '07, pages 729–730, New York, NY, USA. ACM.

Xing, E. P., Ng, A. Y., Jordan, M. I., and Russell, S. (2002). Distance Metric Learning, with Application to Clustering with Side-information. In *Advances in Neural Information Processing Systems 15*, NIPS 2002, pages 505–512.

Xu, J. and Croft, W. B. (1999). Cluster-based language models for distributed retrieval. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR'99, pages 254–261, New York, NY, USA. ACM Press.

Xu, W., Liu, X., and Gong, Y. (2003). Document clustering based on non-negative matrix factorization. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, SIGIR '03, pages 267–273, New York, NY, USA. ACM.

Xue, G. R., Lin, C., Yang, Q., Xi, W., Zeng, H. J., Yu, Y., and Chen, Z. (2005). Scalable collaborative filtering using cluster-based smoothing. In *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR'05, pages 114–121. ACM.

Ye, Z., He, B., Huang, X., and Lin, H. (2010). Revisiting rocchio's relevance feedback algorithm for probabilistic models. In *Proceedings of the 6th Asia Information Retrieval Symposium on Information Retrieval Technology*, AIRS '10, pages 151–161, Berlin, Heidelberg. Springer-Verlag.

Zamir, O. and Etzioni, O. (1998). Web document clustering: a feasibility demonstration. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '98, pages 46–54, New York, NY, USA. ACM.

Zeng, H.-J., He, Q.-C., Chen, Z., Ma, W.-Y., and Ma, J. (2004). Learning to cluster web search results. In *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '04, pages 210–217, New York, NY, USA. ACM Press.

Zhai, C. (2008). Statistical Language Models for Information Retrieval A Critical Review. *Found. Trends Inf. Retr.*, 2(3):137–213.

Zhai, C. and Lafferty, J. (2001). Model-based feedback in the language modeling approach to information retrieval. In *Proceedings of the tenth international conference on Information and knowledge management*, CIKM '01, pages 403–410, New York, NY, USA. ACM.

Zhai, C. and Lafferty, J. (2004). A study of smoothing methods for language models applied to information retrieval. *ACM Trans. Inf. Syst.*, 22(2):179–214.

Zhai, Z., Liu, B., Xu, H., and Jia, P. (2011). Constrained LDA for Grouping Product Features in Opinion Mining. In Huang, J., Cao, L., and Srivastava, J., editors, *Advances in Knowledge Discovery and Data Mining*, volume 6634 of *Lecture Notes in Computer Science*, pages 448–459. Springer Berlin / Heidelberg.

Zhang, P., Hou, Y., and Song, D. (2009). Approximating true relevance distribution from a mixture model based on irrelevance data. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in*

information retrieval, SIGIR'09, page 107, New York, New York, USA. ACM Press.

Zhang, S., Wang, W., Ford, J., and Makedon, F. (2006). Learning from Incomplete Ratings Using Non-negative Matrix Factorization. In *Proceedings of the SIAM Conference on Data Mining*, SDM 2006.