Topical Relevance Models

Debasis Ganguly

B.Tech., M.Tech.

A dissertation submitted in fulfilment of the requirements for the award of

Doctor of Philosophy (Ph.D.)

to the

Dublin City University
School of Computing

Supervisor: Gareth Jones

Jul 2012

I hereby certify that this material, which I now submit for assessment on the programme of study leading to the award of Ph.D. is entirely my own work, that I have exercised reasonable care to ensure that the work is original, and does not to the best of my knowledge breach any law of copyright, and has not been taken from the work of others save and to the extent that such work has been cited and acknowledged within the text of my work.

Signed:

(Candidate) ID No.:

Date:

Contents

List of Figures

List of Tables

Abstract

An inherent characteristic of information retrieval (IR) is that the queries expressing the information need are often multi-faceted, that is, they encapsulate more than one specific sub-information need. This multi-facetedness of the queries manifests itself as a topic distribution in the retrieved set of documents, where each document can be considered as a mixture of topics, some of which correspond to the sub-information needs expressed in the query. In some specific domains of IR, such as patent prior art search, where the queries are full patent articles and the objective is to (in)validate the claims contained therein, the queries themselves are multi-topical in addition to the retrieved set of documents. The overall objective of the research described in this thesis involves investigating techniques to recognize and exploit these multi-topical characteristic of the retrieved documents and the queries in the relevance feedback step of IR.

First, we hypothesize that segments of documents in close proximity to the query terms is indicative of the fact that these segments are topically related to the query terms. An intuitive choice for the unit of such segments, in close proximity to query terms within documents, is the sentences, which characteristically represent a collection of semantically related terms. This way of utilizing term proximity through the use of sentences is empirically shown to select the potentially relevant topics from among those present in a retrieved documents set and thus improve the relevance feedback in IR.

Secondly, to handle the very long queries of patent prior art search which are essentially multi-topical in nature, we hypothesize that segmenting these queries into more topically focused segments and then using these topically focused segments as separate queries for retrieval can retrieve potentially relevant documents for each of these segments. The results for each of these segments then need to be merged to obtain a final retrieval resultset for the whole query.

These two approaches of utilizing the topical relatedness of terms in both the retrieved documents and the queries are then integrated more formally under a single statistical generative model, called the topical relevance model (TRLM). This model is able to utilize the underlying multi-topical nature of both retrieved documents and the query. Moreover, the model is used as the basis for construction of a novel search interface, called *TopicVis*, which lets the user visualize the topic distributions in the retrieved set of documents and the query. This visualization of the topics is beneficial to the user in the following ways. Firstly, through visualization of the ranked retrieval list TopicVis facilitates the user to choose one or more facets of interest from the query in a feedback step, after which it retrieves documents primarily composed on the selected facets at top ranks. Secondly, the system provides an access link to the first segment within a document focusing on the selected topic and also supports navigation links to subsequent segments on the same topic in other documents.

The methods proposed in this thesis are evaluated on datasets from the TREC IR benchmarking workshop series, and the CLEF-IP 2010 data, a patent prior art search data set. Experimental results show that relevance feedback using sentences and segmented retrieval for patent prior art search queries significantly improve IR effectiveness for the standard ad-hoc IR and patent prior art search tasks. Moreover, the topical relevance model (TRLM), designed to encapsulate these two complementary approaches under a single framework, significantly improves IR effectiveness for both standard ad-hoc IR and patent prior art search. Furthermore, a task based user study experiment shows that novel features of topic visualization, topic-based feedback and topic-based navigation, implemented in *TopicVis*, lead to effective and efficient task completion, achieving good user satisfaction.

Acknowledgments

I would first like to express my profound sense of gratitude to my supervisor Gareth Jones for introducing me to this research topic and providing his valuable guidance and unfailing encouragement throughout the course of the work. His sharp insight and enormous support not only helped to shape the work reported in this thesis, but also has constructed outstanding guiding examples to follow for my future career.

I am grateful to Prof. Josef van Genabith and Dr. Cathal Gurrin, who provided insightful suggestions on my transfer report, much of which is incorporated into this thesis. Much of the work in this thesis would not have materialized without the support from Dr. Johannes Leveling, the post-doctoral research fellow, who always provided insightful ideas, constant support and inspiration throughout the course of this thesis.

I would like to express a big thanks to my the past and present members of our research group just to name a few - Jin Ming, Walid Magdy, Maria Eskevich, Wei Li, Sandipan Dandapat, Pratyush Banerjee, Sudip Naskar, Aswarth Dara among many others for their support and interest in my work and for some of the oddest conversations during the tea breaks. Special thanks to Joachim for maintaining our computing cluster in excellent shape and consistently providing technical tips ranging from Linux scripting to what not! Thanks to Eithne, Riona, and Fiona for their kind help since the first day I arrived in Ireland.

I have had a consortium of supporters outside of DCU to whom I am most grateful for reminding me of the outside world. Words are not enough to thank my family members, specially my wife and my parents, for supporting me throughout the ups and downs of the entire duration of my PhD study. A special thanks to my wife Manisha Ganguly for taking up the opportunity to come to Dublin all the way from Calcutta wiith our son thus ending my solitude of two and a half years. She also deserves a heartful of thanks for developing a major part of the search engine interface relevant to my PhD work. She used all her industry experience to develop

a working version of the search interface within a week, which would have taken me at least a month to finish.

Last but certainly not the least, I am thankful to Science Foundation of Ireland for the research grant (grant 07/CE/I1142, Centre for Next Generation Localisation) to support this research.

Chapter 1

Introduction

Information retrieval (commonly known by its abbreviated form IR) is, broadly speaking, the science of retrieving *relevant* information in order to satisfy a user's *information need*. In contemporary IR, the user information need is typically represented as an unstructured *query* statement comprising a number of words which the user hopes is a sufficient representation of their information need to be able to identify relevant documents. The information retrieved in response to a user query, is usually in the form of a set of documents. In practice, it is impossible to compute the exact set of relevant documents for a given query because the mere presence of a query term in a document does not necessarily imply its relevance to the query, or more strictly the information need underlying the query terms. The main challenge in IR is thus in modelling the relevance of a document in the collection to a given query, as accurately as possible. Another challenge is to determine a similarity measure between documents in the collection and the query to define the order in which the retrieved documents are to be presented to the user. The IR system then returns a ranked list of documents sorted by the decreasing values of their similarities with the query.

It is of utmost importance to determine the retrieval effectiveness of an IR system to determine which IR systems prevail over others. In order to compare the retrieval effectiveness between IR systems, it is important to qualitaively measure the amount of user satisfaction achieved by the outputs of the respective systems, that is the ranked list of retrieved documents returned by them. A major hindrance in this approach is that the very notion of satisfiability is highly subjective, and hence difficult to approximately quantify through a subjective judgement of the individual ranked lists. As a practical approximation to measuring user satisfaction with the information retrieved, retrieval effectiveness of an IR system is measured by using quantities such as the number of relevant documents retrieved out of the total number of relevant documents in the collection, known as *recall*, and the number of documents which are relevant out of the total number of documents retrieved, known as *precision*. The former approximates how much of the total existing relevant information the system has been able to retrieve for the user, whereas the latter in turn approximates how much of the total retrieved information is actually relevant.

In general, it is often difficult to achieve a satisfactory retrieval effectiveness during the first phase of retrieval due to the incomplete specification of the user's information need in their initial queries leading to the so-called "vocabulary mismatch" problem between the documents in the collection and the query. For example, a query "atomic power" may not retrieve relevant documents using the vocabulary "nuclear energy" although these phrases refer to the same concept. Problems of vocabulary mismatch and incomplete specification can be addressed in IR systems by exploiting user feedback on the results of the initial retrieval process, so as to improve the quality of retrieval results in a subsequent retrieval step. This process of incorporating feedback is referred to as *relevance feedback*, and involves modifying the initial result list on the basis of relevance information collected from the user. Moreover, since real users are often unwilling to provide manually assessed relevance information for every document that he reads, to make use of relevance feedback techniques, it is a common practice in IR to attempt improvement on the initial search results by assuming that all documents retrieved to a certain rank are relevant and then extracting terms from these documents, adding these to the query and re-retrieving with the expanded query. This process is known as the *pseudo relevance feedback* (PRF) or blind relevance feedback (BRF), the name originating from the implicit assumption that the top ranked documents are *pseudo relevant*.

1.1 Focus of this thesis

An obvious limitation of PRF is in the assumption that the top ranked documents as a whole are relevant to the information need, which is often not true in practice. This is particularly likely to be the case for long documents containing a segment of a relevant piece of information within otherwise non-relevant information. In fact, such long documents are comprised of multiple topics, where it is seldom likely to be true that each such topic is relevant to the query. However, the standard relevance feedback methods in IR do not take this multi-topical nature of a document into consideration, while extracting terms from these documents. As a matter of fact, the multi-topical nature of the retrieved documents is a an indirect manifestation of the multifacetedness in a query. Although standard IR methodologies assume there is only a single aspect of information need expressed in a query, the information need however, can often be categorized into more fine grained aspects. For example, a query may encapsulate the different information needs about the polio disease, its outbreaks, medical protection against the disease and post-polio problems, by the keywords "Poliomyelitis and Post-Polio". This example illustrates that the query entered into a retrieval system, in spite of being short, can be multi-faceted, where the information need is to retrieve relevant content against each facet or aspect. Sometimes, the information need itself will probably only relate to one facet, instead of relating to all of them. However, an IR system has to aim for retrieving against all of the facets because short queries do not in general express the information need sufficiently enough. For this particular example, a retrieval methodology should aim at retrieving several classes of documents, one catering to the disease information, one associated with the prevention of the disease, one pertaining to the post-polio problems and so on. However, current retrieval models do not attempt to exploit this multi-facetedness of the information need, often manifested as clusters of topics in the retrieved set of documents (Xu and Croft, 1996).

A complementary problem arises in the case of long queries, particluarly those which express potentially diverse information needs. A real-life example is the patent prior art search, where prior articles are required to be retrieved and checked in order to (in)validate the novelty of a new patent claim. Standard IR methods do not perform satisfactorily well when the queries are very long, in this case nearly as long as the documents in the target collection. Whole long documents with an obvious lack of focus on a single particular information need if used as queries, may create problems in identifying relevant documents. The reason, again, can be attributed to the fact that the multi-topical nature of the queries is not taken into account during computation of the similarities between the query and the documents in the collection, as a result of which documents related to some topics in the query do not get retrieved at top ranks.

The next problem, which we explore in this thesis, relates to the presentation of search results to the user. The standard paradigm of IR is to present a ranked list of documents to the searcher, sorted in decreasing order by their similarities to the query. Search engines such as Google[1] and Bing[2] display a short text snippet along with the title of each retrieved document with highlighted query keywords. The snippet is intended to indicate the likely relevance of the contents of the full document to the user.

While this standard snippet-based approach of indicating the likely relevance of a document is in general suitable for locating relevant information, the snippets are not likely to be beneficial in cases when the documents and/or the queries are multi-topical. The fact that the patent prior art search, where both documents and queries are multi-topical, is conducted by professional searchers, namely the patent

[1] `http://www.google.com`

[2] `http://www.bing.com`

examiners, illustrates the complexity of such a search task. The patent examiners often have to meticulously read through hundreds of documents for finding the prior art of a submitted patent application (Azzopardi et al., 2010), which in turn shows that the standard paradigm of relevant information access through ranked lists of documents with associated snippets is not fully effective for complex search tasks.

The work in this thesis seeks to address each of these problems of the traditional IR paradigm. Our work is centred around the hypothesis that these problems can be alleviated by utilizing information from segments[3] of documents or queries, as the case may be, the primary reason being that segments, unlike full documents, are more focused on an individual topic. Generally speaking, for the standard search problem with short keyword type queries, we focus on exploiting document segments for a more careful selection of the feedback terms, whereas for the case of very long queries, we devise a technique of retrieving against each aspect of the query. Finally, we design a graphical user interface to facilitate navigation through selective topics, which we refer to as topic-based navigation. Topic-based feedback, i.e. feedback where documents predominantly expressing a particular topic of interest can be returned at top ranks, can also potentially be supported by such an interface.

We now introduce the research questions explored in this thesis, and conclude the chapter by providing a roadmap for the rest of this thesis.

1.2 Research Questions

The previous section has introduced important limitations of standard IR systems, and has discussed how we plan to extend the standard IR paradigm to attempt address these. The work in this thesis in general is motivated by the hypothesis that standard IR can be extended to mine topically related information from documents and queries so as to improve retrieval effectiveness. This section formulates our

[3] We henceforth refer to document or query sub-parts as document or query segments. Most of the times, we will simply use *segments*, and whether it pertains to a document or a query will be clear from the context.

specific research questions, associated with this objective.

In our first research question, we explore whether term proximity can play a part in identifying terms from retrieved documents that are topically related and hence perhaps likely to be relevant, to the given query terms (Luhn, 1958). If the query terms are topically related, they themselves are likely to occur in close proximity to each other for instance within a single sentence, whereas if they are multi-topical, which can often be the case, the terms are likely to occur scattered in multiple sentences. The hypothesis is that additional terms from each such sentence, being topically related to the query terms occuring in it, can potentially play a pivotal role in expanding the initial query and enriching the initial information need, leading likely to a improved retrieval effectiveness. The first research question is thus stated as follows.

RQ-1: *Can additional terms in close proximity to query terms from retrieved documents enrich the information need of the query and improve the retrieval effectiveness of ad-hoc IR?*

We have already pointed out the potential benefit of segmenting very long queries, expressing several diverse information needs, into more focused segments concentrating on a single and more precise information need. Our second research question is thus directed towards exploring whether segmentation of very long queries into smaller units can better represent the more fine grained information needs expressed in topically coherent segments and thus help improving retrieval effectiveness.

RQ-2: *Can segmentation of very long queries into topically coherent segments be utilized to improve IR effectiveness?*

In our first two research questions, we hypothesize that term proximity, at a granularity level of sentences in the case of retrieved documents or paragraphs in the case of very long queries, implies that all terms in such segments (either a sentence or a paragraph) are topically related or in other words belong to the same topical class. In some documents, this assumption may be too restrictive, where even proximal terms may be associated with different topics. In such cases, a more flexible option is to consider each document as a mixture model of topics. Such a representation of documents is in fact realized by statistical approaches to *topic modelling*, which generally speaking involve inferring a probability distribution from terms to a set of latent topics (Hofmann, 1999; Blei et al., 2003). The advantages are that: a) co-occurrence patterns rather than positions determine the likelihood of terms belonging to a topic; b) a term can belong to multiple topics with different probabilities; and c) the segments thus need not neccessarily be comprised of contiguous blocks of text. Another motivation for the third research question is that research questions RQ-1 and RQ-2 are based on two complementary approaches of mining topical relations within retrieved documents on one hand and the query on the other. An interesting question then is whether these two approaches can be encapsulated within the framework of a single model.

Not only does the third research question RQ-3 therefore attempts to generalize the proximity hypothesis of term relatedness addressed in research questions RQ-1 and RQ-2 by explicitly modelling topics, but it also aims to unify within a single framework the two complementary approaches pursued in them.

RQ-3: *Can topic modelling prove beneficial in improving the retrieval effectiveness for both short and long queries thus unifying the solutions of RQ-1 and RQ-2?*

The last research question, explored in this thesis, is about exploring the potential benefits of segmentation for providing more convenient access to relevant information.

RQ-4: *Can topical segmentation of the documents and the queries be helpful in providing a topic-based access to relevant information?*

Towards answering this question, we develop and evaluate a user interface facilitating automatically guided topic-based navigation through search results, and topic-based feedback to rerank search results on the basis of a topic selected by the user.

1.3 Structure of the Thesis

This thesis is structured as follows.

- **Chapter 2** provides a comprehensive literature survey of related work, highlighting the differences of our methodologies with the existing ones. In particular, we revisit the fundamental methods of IR, starting with standard techniques of how documents and queries are represented, followed by a brief description of the standard retrieval models ranging from the basic tf-idf model to more effective ones such as the probabilistic model and the language model. We then review relevance feedback in IR introducing both term re-weighting and query expansion with new terms. We also review standard metrics used for retrieval effectiveness evaluation. This is followed by an overview of existing topic modelling approaches; we first survey topic modelling approaches in general, before reviewing their applications in IR.

- **Chapter 3** presents an overview of the resources, tools and the characteristics of datasets used for the experiments in the subsequent chapters of this thesis. We describe the TREC dataset used for the experiments involving research questions RQ-1 and RQ-3, where the queries are short comprising a few keywords. We then describe the dataset characteristics of CLEF-IP 2010, which is the dataset used for experimenting with much larger queries with an aim to explore research question RQ-4. This chapter also introduces the tools and resources used for the experiments performed in this thesis.

- **Chapter 4** introduces our work on relevance feedback in IR pertaining to research question **RQ-1**. According to the hypothesis that whole documents are seldom relevant to the query, and that long documents often contain a piece of relevant information within otherwise non-relevant information, in this chapter we propose a relevance feedback methodology where the information to be used for relevance feedback is extracted from the sentences of relevant documents which are most similar to the query. The motivation is that terms in close proximity to the query terms are likely to be topically related to the query terms and hence are likely to enrich the information need expressed in the initial query. Experimental investigations show that our proposed method of relevance feedback, which we call sentence based query expansion (SBQE), outperforms the standard approaches of relevance feedback which use information from whole documents.

- **Chapter 5** discusses the complementary method of segmenting queries instead of documents, thus addressing research question **RQ-2**. The main hypothesis underlying the research question RQ-2 is that a very long query document often encompasses several information needs. Using the whole query for retrieval in such a case may not result in effective retrieval against each such fine-grained information need. In this chapter, we propose a method of segmenting the whole query in separate segments, and then using these segments separately for retrieval. We demonstrate that our proposed method of merging resultlists obtained by retrieval with separate query segments outperforms the standard approach of retrieving with the full query.

- **Chapter 6** proposes a formal probabilistic generative model of topic or aspect based relevance combining the ideas of Chapters 4 and 5, thus exploring research question **RQ-3**. The work in this chapter explicitly models the topical representation of documents and queries in contrast to the work in Chapters 4 and 5 pertaining to the previous two research questions respectively, where the assumption was that proximity alone plays a part in identifying topically related terms. Topical segmentation infers a posterior distribution of how likely it is for a word to belong to each of the topical classes. These word-topic membership probabilities are then used in the generative model to estimate relevance models for each topic. The proposed model is evaluated on both short keyword queries and very long queries. The results are also compared with the approaches of Chapter 4 and Chapter 5 respectively.

- **Chapter 7** explores the final research question **RQ-4**. The objective of this research question is to explore ways of providing a topic based access to relevant information to the users of a search system. To this end, we describe how the model developed in Chapter 6 can be applied to design a graphical user interface to support an topic-based access to relevant content. Our developed search interface provides visualization of the topics in each retrieved document and the query in order to enable a user to match the related topics between the retrieved documents and the query with the help of visual cues. Morevover, the interface also provides quick naviagation links between related parts of documents. These new features facilitate the search interface in serving a two-fold advantage. Firstly, the interface helps in saving the reading effort of a user to locate relevant pieces of information within long expository articles. Secondly, the interface through visualization of the topics, some of which may in fact relate to more fine grained aspects of the overall information need, help in the discovery of these latent aspects and hence in the reformulation of the user query towards any of these aspects.

- Finally, **Chapter 8** concludes the thesis by summarizing the research achievements and providing directions for future research.

Chapter 2

Information Retrieval and Topic Modelling

This chapter primarily builds up the background necessary to fully understand the subsequent chapters of this thesis. It starts with a comprehensive survey of existing IR approaches, including a summary of standard retrieval models and relevance feedback methods. Next in Section 2.2), we present an overview of the topic modelling literature, which is a necessary background to read Chapters 6 and 7, which focus on applying topic modelling for PRF and topic visualization.

2.1 Overview of Information Retrieval

The architecture of a generic IR system is schematically shown in Figure 2.1 (reproduced from (Croft, 1993)). Each document in the collection needs to be processed before it can be used for retrieval. This processing of the documents, which enables them to be retrieved effectively, as well as efficiently on entering of a query, is shown as the *representation* box below the *documents* entity in Figure 2.1. Analogous to the document processing phase, the information need of the user also has to be processed to form a query, which can be used in the retrieval step. This is shown as the *representation* box below the entity *information problem* in Figure 2.1. The

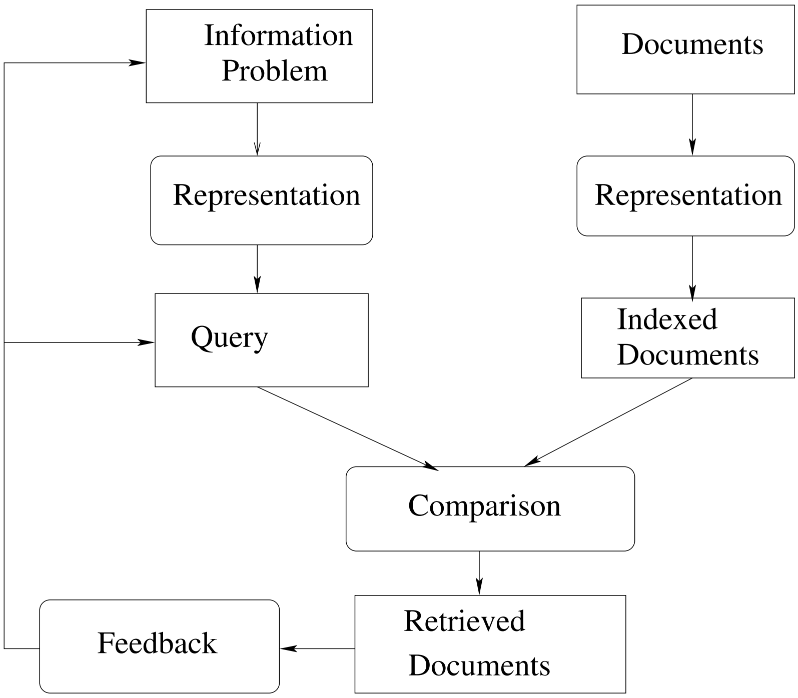


Figure 2.1: Information Retrieval Process (Croft, 1993)

retrieval process then involves the *comparison* step in which the *query representation* is *compared* or *matched* against the *documents representation* to return a set of documents most *similar* to the query. In the following sections, we examine each of these processes in more details.

2.1.1 Document Representation

An important component of an IR system is the way in which documents are represented. This representation, commonly known as the *indexing* process, has to be independent of the query because the set of queries is not known to an IR system a priori. Although the query is not known, an IR system needs to organize the documents in such a way so that they can be retrieved at search time very efficiently when a new query is entered into the system. The system constructs a list of documents available for retrieval in response to a query term. This list is typically is the set of documents in which a particular term occurs. In practice, given a term, the system must be able to constitue this list of documents in which this term occurs very quickly. Given a collection of documents, an efficient way to achieve this is to compute the list of documents in which a particular term occurs, and store each such list (called *postings list* or simply *postings*) computed over the set of all terms of the collection in a file. This is what is done in the *indexing*[1] phase which produces an *inverted file* as an output. It is called an *inverted file* bacause contrary to the direct approach of obtaining the list of terms from a document, with the help of this file it is possible to access the list of documents, given a term. While retrieving with a query, the individual lists obtained for each query term need to be merged as required by the retrieval model adopted. For example, if the query is a boolean AND of two terms, then the set of retrieved documents is the intersection of the two lists, whereas if the query is a boolean OR of the terms then a set union needs to be performed over the postings lists need in order to constitute the list of retrieved documents.

In addition to storing the term presence information, it is often required in practice to store the *importance* or *weight* of a term in a document for effectively predicting relevance of a document in response to a query. Moreover, in addition to storing the per-document weights of a term in the inverted list, an index also stores collection level information, such as the collection frequency, for each term. Furthermore, an index may also contain additional information such as term positions for proximity or phrase search.

In summary, the process of indexing involves organizing a given document collection into an inverted file which for each term in the collection contains the collection statistics of the term along with a list of documents in which the term occurs. The inverted list supports efficient access with the term identifier of a query term used as the *key*. Each term in the collection contributes to a head node in an inverted list, accessed by a hash map or a trie. Each head node in turn points to a sorted list (commonly called *postings*) of document identifiers and the *importance* of that term

[1] This excludes *dynamic indexing*, in which an existing index can be updated with additional documents without the need of creating a new index from scratch. Although dynamic indexing is highly useful in applications such as commercial web search engines, it is however a standard practice to use a static collection of documents for research purposes. The term *indexing*, henceforth in this thesis, will imply *static indexing*.

in each document. The sorted list data structure helps to efficiently accumulate the total similarity for a query in linear time over the individual postings for each query term.

2.1.2 Query Representation

The *information problem* (or *information need*), as shown in Figure 2.1, is an abstract entity which is transformed into a physically existing query string by the users of a search system through the *query representation* process.

The query representation process can encode complex information needs such as those involving:

- *Boolean search*, where the query is a Boolean predicate with operators such as AND, OR etc.

- *Field search*, where document field names are specified in the query and the objective is to seek matching terms in each of these specified fields.

- *Phrase search*, where the objective is to retrieve relevant documents containing a particular phrase. Note that since the meaning of a phrase can be entirely different to the meaning of its constituent words, a match of the whole phrase may not be equivalent to matching any of the individual words, e.g. a phrase query such as "German shepherd" should not retrieve documents having isolated existences of the constituent terms "German" and "shepherd".

- *Proximity search*, a generalization of phrase search, where documents with matching query terms in any order within a specified span is sought for.

Despite the complex query representation processes as outlined above, the most simple and user friendly way to represent a query is to accumulate the key terms describing the information need into a structure-free text string.

2.1.3 Retrieval Models

The aim of a retrieval model is to retrieve relevant documents satisfying a user's information need. To achieve this, generally speaking, a retrieval model needs to *compare* the given query with the documents in the collection and use the results of these comparisons to decide which documents to retrieve and if ranked, the order in which they should be shown to the users.

The oldest and the simplest retrieval model used in IR is the *Boolean model.* The query in a Boolean retrieval model is represented as a sequence of terms separated by the Boolean operators, such as the AND, OR and NOT. The relevant documents, in this case are those which satisfy the Boolean predicate function expressed by the query. For example, if the query is *relativity AND theory*, the Boolean retrieval model retrieves documents containing both the terms *relativity* and *theory.* Recall from Section 2.1.1, that this can be achieved by an intersection of the postings lists for the terms *relativity* and *theory.*

A major limitation of the Boolean model is that it is not possible to obtain a ranking of the retrieval results, For example, if two documents satisfy the Boolean predicate of a query, the model does not specify which document to report first. This is turn does not conform to the user expectation of finding the most relevant document at the first rank, followed by the ones which are progressively less and less relevant. A second major disadvantage is that the information need itself can be more complex than a simple Boolean predicate function [2] For example, a document containing the term *relativity* may still be relevant to the query *relativity theory* even if it does not contain the term *theory.*

To address these limitations, a retrieval model needs to compute a relevance score of some sort between the query and each document to predict how much a document is likely to be relevant to a given query so that it can be used to rank them. Standard retrieval models involve computation of a similarity score between

[2] A strength of the Boolean retrieval model is that it is possible to specify the exact relevance criterion as a Boolean function in some search domains, e.g. a Boolean predicate for relevance is likely to satisfy a user searching for a book in a library.

documents and the query (cf. Figure 2.1). This similarity score considers the relative importance of a query term match in a document and then accumulate these values for all the query terms to yield the total score of a document. The purpose of a *retrieval model* is to define the method for term importance prediction of a query term match in a document and how to combine these predicted values eventually in constituting the final score of a document. Here, we review some of the well known retrieval models in IR.

Vector Space Model

The oldest of the established ranked retrieval models is the vector space model (VSM) (Salton et al., 1975). In the VSM, the query q and each document d are represented as vectors over the term space of the entire vocabulary (say of size n) of the document collection. The basic assumption for the operation of the VSM is that the potential relevance of a document to a query is related to the similarity of their vector representation. The advantage of such a representation is that the concept of distance is well defined in a vector space. A query and a document are similar if their vector representation is close, i.e. if the angle between their vectors is small. The Euclidean distance is not particularly suitable for IR, because it heavily depends on the length of vectors, and as a matter of fact, documents are typically much longer than the queries. Thus, the angle between two vectors is used as a measure of distance, which is in fact proportional to the Euclidean distance between length normalized unit vectors. The cosine of the angle (say ϕ) between two vectors, which is simply the dot product of two normalized vectors, as shown in Equation 2.1, is easier to calculate than the angle itself. The cosine of the angle between a document and the given query vector is thus used directly as a measure of inverse distance or similarity (i.e. closeness).

$$sim_{VSM}(d,q) = \sum_{i=1}^{n} d_i q_i = |d||q| \cos\phi \tag{2.1}$$

An important issue with regard to the vector space similarity function is how to define the components of the document and query vectors. Clearly, the retrieval effectiveness in the VSM depends primarily on how the components of the document and the query vectors are defined. The process of defining the vector components is called *term weighting.* A term weighting function depends on three important factors as follows.

a) **Term Frequency:** The frequency of a term in a document approximates the *aboutness* of the document, e.g. *information* and *retrieval* are highly frequent words in this thesis. Assigning a higher weight to these terms enables this document to be retrieved at top ranks for a query containing the terms *information* and *retrieval.* Using the absolute value of the frequency of a term as the term weight does not produce effective results primarily because if a document has one matching query term with very high term frequency, then that document is not necessarily more relevant than another document which has two matching query terms with less frequencies (Singhal, 1997). For an example query *relativity theory*, if D_1 has only one matching term *theory* with 20 occurrences, and D_2 has two matching terms *relativity* and *theory* with frequencies of 3 and 5 respectively, it is more likely the case that the latter is more relevant than the former, because the term *theory* in D_1 may refer to some theory other than the relativity theory. The term frequency function thus has to ensure that documents with a higher number of query term matches are ranked higher (better) than documents with a lower number of query term matches, a characteristic commonly known as the *coordination ranking* in IR literature (Hiemstra, 2000).

   Some commonly used term frequency functions used to ensure coordination ranking in VSM are as follows:

   - the *logarithmic tf*: $1 + \log(tf)$ (Buckley et al., 1993; Singhal et al., 1996), which is designed primarily to down-weight the contributions of terms with very high frequencies in a document.

- the *augmented tf*: $\frac{1}{2} + \frac{tf}{2\max(tf)}$, which normalizes the term frequency values within a range of $[\frac{1}{2}, 1]$ (Salton and Buckley, 1988).

It is easy to see that both the logarithmic and the augmented tf ensure that D_2 is ranked higher than D_1. The score assigned to D_1 by logarithmic tf is $1+log(20) = 1 + 2.99 = 3.99$, whereas the score assigned to D_2 is $1 + log(3) + 1 + log(5) = 2+1.09+1.60 = 4.69$, which is higher than that of D_1. The augmented tf in turn scores D_1 as $0.5 + 20/(2 \times 20) = 1$, and D_2 as $0.5 + 3/(2 \times 5) + 0.5 + 5/(2 \times 5) = 1 + 0.3 + 0.5 = 1.8$, which is higher than that of D_1.

b) **Inverse Document Frequency:** The mere presence of a word within a document should not be an indicator of its importance, e.g. common words in English such as "the", "of" etc. may occur in almost all documents of a collection, and hence can play no role in distinguishing a relevant document from a non-relevant one. In practice, a pre-configured list of such words, commonly known as *stopwords*, are filtered from documents before they are entered into the index. Non-stopwords, i.e. words not belonging to the stopword list, yet occurring in a large number of documents in the collection, should also not play a pivotal role in distinguishing documents. Hence, a measure, which is inversely proportional to the *document frequency* (the number of documents in which a term occurs), is used as a factor to weigh the importance of a term. This measure is referred to as the *inverse document frequency* (*idf*) (Salton and Buckley, 1988). It is important to note that *idf* is a feature of the collection rather than of individual documents. As an example the terms *information* and *retrieval* should be the distinguishing factor in retrieving this thesis from a collection of theses on other subjects such as machine learning or machine translation. However, if the collection comprises of theses on IR, then these terms are assigned a low *idf* value. The most commonly used idf measure is the logarithmic idf defined as $idf(t) = \log(\frac{N}{df(t)})$, where N is the total number of documents in the collection and $df(t)$ is the number of documents in which t occurs (Sparck-Jones, 1973).

c) **Document Length:** Long documents tend to have a higher term frequency for the constituent terms, as a result of which an IR system tends to retrieve longer documents at higher ranks due their higher tf values which arise mainly due to the large length of a document rather than due to its informativeness. Moreover, long documents comprising of a high number of terms are associated with a higher likelihood for more query term matches, as a result of which they tend to be retrieved at higher ranks due to the effect of coordination ranking.

   An approach to limiting the impact of length related factors in document ranking, is to use *length normalization* to negate this bias towards retrieving longer documents. A common method to do this is cosine normalization which involves reducing the length of each document vector to unity by dividing the components with the magnitude of the vector reducing each document vector to unity so that the dot product of a document and the query (cf. Equation 2.1) yields the value of the cosine of the angle between them. VSM term-weighting with the tf, idf components normalized with the cosine measure was shown not to perform well for large document collections in early TREC evaluations (Singhal, 1997). However, VSM was significantly improved in later TREC evaluations by the introduction of the *pivoted length normalization* technique, which is a document length normalization method (Singhal et al., 1996), the working principle of which involves favouring shorter documents (documents shorter in length than a threshold length, say l_t) by boosting their similarity values and down weighting those of the longer documents (documents with length $>= l_t$). The value of l_t has to be computed empirically using a training set of queries with relevance judgements.

Since both the term frequency (*tf*) and the inverse document frequency (*idf*) are essential in deriving a effective term weight for retrieval, it is a standard practice in VSM to combine these components simply by multiplying them together. This combination is usually known as the tf-idf weighting.

The major criticism against the VSM is that the model itself does not propose a theoretically sound principle for determining the term weighting components, such as which tf function to use, whether to use cosine normalization or the pivoted length normalization etc. Neither does the VSM model theoretically justify the multiplicative combination of tf-idf weighting. The VSM was subsequently followed by the development of more theoretically motivated IR models which we review next.

Probabilistic Model

The main principle behind the probabilistic model of IR is that it estimates the posterior probability of a document d being relevant, given the query q, i.e. $P(d = R|q)$ for each document d in the collection, and then simply ranks the documents in decreasing order of these probabilities. This is known as the *probabibility ranking principle* (PRP) (Robertson, 1977). The basic version of the probabilistic model uses a binary independence model (BIM) between terms, i.e. it assumes that the terms are pairwise independent. Note that the VSM also implicitly assumes this while mapping each term as an orthogonal axis in the term space. The limitation of BIM is that it relies on the boolean presence of a term in a document, and does not use the term frequencies or document length information. The BM25 weighting model extends the BIM by incorporating these information (Robertson et al., 1994; Sparck-Jones et al., 2000). More specifically, the BM25 model scores a document d by accumulating the *idf* values of the query terms multiplied by the factor of frequency of each term and the document length, as shown in Equation 2.2.

$$sim_{BM25}(d,q) = \sum_{t \in q} \log \frac{N}{df(t)} \times \frac{(k_1+1)tf(t,d)}{k_1(1-b+b\frac{L_d}{L_{avg}})+tf(t,d)} \tag{2.2}$$

In Equation 2.2, $tf(t,d)$ is the term frequency of a term t in document d, L_d is the length of document d, and L_{ave} is the average length of documents computed over the collection. The tuning parameters k_1 and b serve the following purpose.

- $k_1 = 0$ totally ignores the term frequency contribution, and the similarity only depends on the idf factor. A very high value of k_1 favours the tf factor more in comparison to the idf factor. The reasonable value of k_1, is between 1 and 1.2 which is often found to strike a balance between the two contributions (Robertson et al., 1994).
- b ($0 \leq b \leq 1$) controls the degree of length normalization. $b = 1$ ensures full length normalization, wheras $b = 0$ imples no length normalization. Again, a reasonable choice of b is often found empirically to be 0.75, which suggests that length normalization is an important factor, but on average giving too much imporatnce to this factor can be ineffective since long relevant documents tend to be over-penalized in case of full length normalization.

Language Modelling

The language modelling (LM) approach to IR, similar to BM25, is motivated by the PRP. The main difference is that instead of computing the probability estimates that a document is relevant to a given query, as in a probabilistic model, LM estimates the probability of generating a query from a document. The working principle of LM is that a query term is assumed to be generated by a uniform sampling process from a document (which thus corresponds to the tf contribution in term-weighting) or from the collection itself (which in turn corresponds to the idf factor). This is analogous to the process of query formation by a real-life user, in that the user would typically constitute a query by recollecting *important* terms that are likely to be contained in a document that is in turn likely to be relevant to the information need, or the user The derivation of the LM approach to IR starts with a formulation of the expression for the PRP basis of ranking where documents are sorted by the decreasing values of the posterior probabilities of relevance with respect to a query, which in the case of LM is represented by the probability of generating a document

d given a query q, i.e. $P(d|q)$.

$$P(d|q) = \frac{P(q|d)P(d)}{\sum_{d' \in C} P(q|d')P(d')} \propto P(q|d)P(d) \tag{2.3}$$

Equation 2.3 is obtained by appying Bayes theorem, and ignoring the constant factor for the denominator. To find an expression for the right hand side of Equation 2.3, note that the probability of sampling the query q from a document d, assuming a unigram sampling model, is then given by

$$P(q|d) = \prod_{t \in q} \lambda P_{MLE}(t|d) + (1 - \lambda)P_{coll}(t) \tag{2.4}$$

The notations of Equation 2.4 are explained as follows.

- P_{MLE} is the maximum likelihood estimate of generating a query term t from d and is given by Equation 2.5.

  $$P_{MLE}(t|d) = \frac{tf(t,d)}{L_d} \tag{2.5}$$

  .

- $P_{coll}(t)$ is the probability of generating the term t from the collection. This is typically given by the ratio of the number of documents in which t occurs to the total value of $df(t')$ for each t' in the collection, as shown in Equation 2.6.

  $$P_{coll}(t) = \frac{df(t)}{\sum_{t'=1}^{n} df(t')} \tag{2.6}$$

  Sometimes, *collection frequency*, a measure similar in nature to the document frequency, is used for computing $P_{coll}(t)$. The collection frequency of a term t, denoted by $cf(t)$, is the number of times the term t occurs in the collection. This measure is normalized by the total number of terms in the collection, i.e. the collection size. The reciprocal of the normalized collection frequency is the inverse collection frequency (icf).

In Equation (2.4), tf and cf denote the term and collection frequencies respectively. $P(q_i)$ can also be obtained by smoothing with document frequencies instead of collection frequencies, the difference between the two being insignificant (Hiemstra, 2000).

- λ is the probability of a binary indicator random variable, say X, whose value forces a selection between the two events of either choosing t from d (if $X = 0$), or choosing t from the collection (if $X = 1$). The probability of choosing a term from the document d, i.e. $P(X = 0)$, is denoted by λ. The probability of the complementary event is thus $P(X = 1) = 1 - \lambda$. λ thus balances the tf and the idf components, playing a role similar to that of k_1 in BM25.

  The parameter λ is also known as the smoothing parameter because by adding the collection or the document frequency component it ensures that the probability of generating a term t from a document d, i.e. $P(t|d)$ is never zero even if the term t does not exist in d or in other words $P_{MLE}(t|d) = 0$.

The document length factor is taken care of in LM by the prior probability of a document, i.e. $P(d)$ in Equation 2.3. Instead of assuming uniform priors, this may be a function of the document length of d, namely L_d (Hiemstra, 2000).

Hiemstra and Kraaij (2005) showed that an LM approach with non-uniform document length priors outperformed the BM25 retrieval model by 8.19% on the TREC-7 dataset. For the experiments described in this thesis, the initial retrieval results are obtained by the use of LM with non-uniform document length priors as described in (Hiemstra, 2000).

2.1.4 Relevance Feedback

The initial retrieval results, obtained after the matching step, can be improved by applying *relevance feedback* (see Figure 2.1), a process which involves modifying the search query or system parameters, and then carrying out further retrieval runs. The feedback can be explicitly obtained from the searcher such as asking him to mark the relevant documents. An IR system then uses the known relevant documents to refine the search with adjusted parameters aiming to eventually return a modified ranked list containing more relevant documents at better ranks.

Generally speaking, relevance feedback encompasses two activities:

- *Query term re-weighting*, which involves re-weighting the terms of the query. The objective of term re-weighting is to increase the weight of terms that are likely to be relevant and down-weight the ones which are likely to be non-relevant to the query.

- *Query expansion* (QE), where additional terms, i.e. the ones which do not already appear in the query, are added from the relevant documents. Typically, QE is accompanied by a re-weighting of the query terms, i.e. both the original and the additional ones.

In practice, users are not always keen to provide feedback to a search system. Moreover, obtaining relevance feedback from real users is also not possible in an automatic IR experiment framework. Consequently, methods of implicitly obtaining feedback are common in practice. Implicit feedback can either be obtained through user interaction events such as clicks and subsequent document visiting times, or by assuming that a certain number of top ranked documents from an initial retrieval step are relevant. This is known as pseudo relevance feedback (PRF), which is a simple and often effective technique to improve on the initial retrieval output in the absence of explicit or implicit user feedback. In Figure 2.1, PRF is shown by the arrow from the *feedback* box going back to the *query* box. The case of explicit user feedback is shown by the other arrow, which goes back to the *information problem* box, implying that the searcher interacts with the retrieval systen, and often modifies the information need itself. In this thesis, we mainly concentrate on PRF.

Query term re-weighting

A classic example of term reweighting for PRF is Rocchio expansion (Rocchio, 1971) of the vector space model (VSM) (See Section 2.1.3). Recall that in VSM, both the query and the documents in the collection are represented as vectors. The pseudo-relevant document set is thus a set of vectors. The objective of the Rocchio method is then to shift the query vector towards the centroid of these vectors and away from the centroid of the non-relevant ones (the set of documents in the initially retrieved set of documents complementary to the set of the pseudo-relevant ones). This shifting of the vector is realized by reweighting its components.

The query modification algorithm as proposed by Rocchio is shown in Equation 2.7. The parameters α, β, and γ are the weights attached to the original query vector q, the set of judged relevant documents in the feedback step (R), and the complementary set of non-relevant documents (NR) respectively. It has been shown that setting $\alpha = \beta = \gamma = 8$ works well in practice on the TREC dataset (Mitra et al., 1998).

$$q' = \alpha q + \frac{\beta}{|R|} \sum_{d \in R} d - \frac{\gamma}{|NR|} \sum_{d \in NR} d \tag{2.7}$$

For the probabilistic model, the most commonly used method to rank potential expansion terms is based on the Robertson selection value (RSV) (Robertson, 1990; Robertson et al., 1994), which is defined as follows:

$$RSV(t) = \log \frac{(r + .5)(N - R - n + r + .5)}{(n - r + .5)(R - r + .5)} \tag{2.8}$$

In Equation 2.8, r is the number of known relevant documents in which the term t occurs, N is the total number of documents in the collection, n is the total number of documents in which term t occurs, and R is the number of known relevant documents. Equation 2.8 is employed to reweight a query term t in the PRF step. The objective is to put more emphasis on terms with high idf values occuring frequently in the (pseudo-)relevant documents.

Hiemstra (2000) proposes using the Expectation-Maximization (EM) algorithm to compute optimal retrieval settings for the importance of query terms i.e. the λ_i values associated to each query term q_i (which during the initial retrieval is set to $\lambda\ \forall i$). In this approach, the expectations are computed using an initial probability estimate and then the probability estimates are refined to maximize the computed expectations.

$$\begin{aligned} m_i &= \sum_{j=1}^{R} \frac{\lambda_i^{(p)} P(t_i|D_j)}{\lambda_i^{(p)} P(t_i|D_j) + (1-\lambda)P(t_i)} \quad (\mathrm{E-step}) \\ \lambda_i^{(p+1)} &= \frac{m_i}{R} \quad (\mathrm{M-step}) \end{aligned} \tag{2.9}$$

Equation 2.9 shows that the expectations at the p^{th} iteration are computed using the probability values of the p^{th} iteration, namely the $\lambda_i^{(p)}$ values. In the M-step, the probabilities for the next iteration are recomputed using the expectation value. This cycle repeats till a pre-configured number of maximum iterations or until the probabibility values converge.

More recent works on query term reweighting are co-occurence based. The key idea is that if a term in a pseudo-relevant document co-occurs more with a query term, it is assigned a higher weight as compared to a term with a lower number of co-occurrences. Xu and Croft (1996) proposed Local Context Analysis (LCA) which involves decomposing the feedback documents into fixed length word windows and then ranking the terms by a scoring function which depends on the co-occurrence of a word with the query term, the co-occurrence being computed within the fixed word length windows. In contrast to Rocchio's method, LCA also uses the *idf* information of a word to boost the co-occurrence score of rarely occurring terms compared to the commonly occuring ones. The additional query terms in LCA are assigned weights proportional to the co-occurrence-based scoring function.

Lavrenko and Croft (2001) establishes the co-occurrence principle of LCA theoretically by proposing the relevance model (RLM). In the relevance model, it is assumed that the terms in the (pseudo)-relevant documents as well as the query

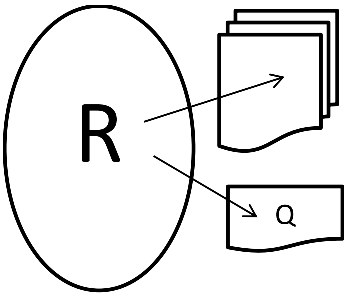


Figure 2.2: Relevance model.

terms are sampled from the same generative model, which in this case is a hypothetical model of relevance. If the documents relevant to a given query are known, it is easy to estimate the relevance model using the maximum likelihood estimate of the probability of generating a term from the relevance model, The observable variables in the model are the generated query terms from the relevance model. Thus, the estimation of the probability of a word w being generated from the relevance model is approximated by the conditional probability of observing w given the observed query terms. This is illustrated in Figure 2.2. The relevance model, represented by the oval on the left hand side of the figure labelled "R", is shown to generate the set of relevant documents and the query represented by the directed arrows going from the oval on the left hand side to the documents and the query.

Given a query $q = \{q_i\}_{i=1}^{n}$ of n terms, the probability of generating a word w from an underlying relevance model R is thus approximately estimated as follows.

$$P(w|\mathrm{R}) \approx P(w|q_1, \ldots, q_n) = \prod_{i=1}^{n} P(w|q_i) \tag{2.10}$$

Assuming that the query terms are conditionally sampled from multinomial document models $\{D_j\}_{j=1}^{R}$, where R is the number of top ranked documents obtained

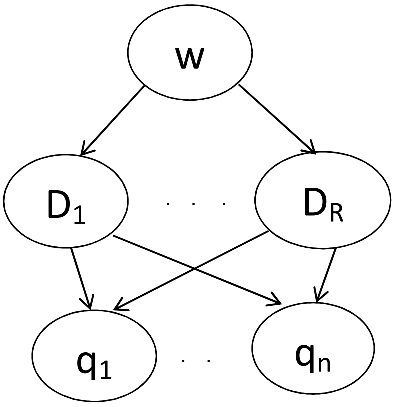


Figure 2.3: Dependence graph for relevance model.

after initial retrieval, as shown in Figure 2.3, we obtain

$$\begin{aligned} P(w|q_i) &= \sum_{j=1}^{R} P(w|D_j)P(D_j|q_i) \\ &= \sum_{j=1}^{R} \frac{P(w|D_j)P(q_i|D_j)P(D_j)}{P(q_i)} \\ &\propto \sum_{j=1}^{R} P(w|D_j)P(q_i|D_j) \end{aligned} \tag{2.11}$$

The last step of Equation (2.11) has been obtained by discarding the uniform priors for $P(q_i)$ and $P(D_j)$. Equation (2.11) has an intuitive explanation in the sense that the likelihood of generating a word w from the relevance model R will increase if the numerator $P(w|D_j)P(q_i|D_j)$ increases, or in other words w co-occurs frequently with a query term q_i in a pseudo-relevant document D_j. The RLM thus utilizes co-occurrence of a non-query term with the given query terms to boost the retrieval scores of documents, which otherwise would have had a lower language model similarity score due to vocabulary mismatch.

We will revisit Equation 2.11 to develop a generalized version of the relevance model while exploring the research question RQ-3.

Query expansion

Query expansion (QE) is a popular technique used to bridge the vocabulary gap between the terms in the query and the documents. QE techniques work by adding terms to the user's original query so as to enrich it to better describe the information need by including additional terms which might have been used in the relevant documents (Rocchio, 1971), or which augment the terms in the original query such as synonyms (Berger and Lafferty, 1999). If good expansion terms are selected then the retrieval system can retrieve additional relevant documents or improve the rank of documents already retrieved. Query expansion techniques aim to predict the most suitable candidate words to be added to the query so as to increase retrieval effectiveness.

An example of a vocabulary gap is when a user queries for "*atomic power*" whereas most documents in the collection relevant to this particular information need contain the words "*nuclear energy*". Addition of the words *nuclear* and *energy* to the original query can result in these potentially relevant documents to be retrieved.

A standard QE approach is typically term-based, i.e. a subset of terms occurring in relevant documents are chosen based on some term scoring function aiming to select the good expansion terms. The various different retrieval models for IR have corresponding different recommended term scoring functions for QE.

The term scoring function for VSM, which works well in practice in combination with Rocchio's term reweighting, uses term occurrence statistics alone as advocated by (Buckley et al., 1994), where terms occurring in a larger number of pseudo-relevant documents are added to the query. The score assigned to a term t is in this approach is shown in Equation 2.12, where r is the number of pseudo-relevant documents that the term occurs in.

$$Occ(t) = r \tag{2.12}$$

Such a simple scoring function does not distinguish terms by their collection statistics and might end up adding too many common terms (because these terms are also abundant in the relevant documents), thus not increasing IR effectiveness significantly. Scoring functions thus need to be augmented by incorporating the idf factor (Robertson, 1990; Robertson et al., 1994). In fact, Equation 2.8 in addition to reweighting the terms appearing in a query, can also be used to select additional expansion terms with high values of $RSV(t)$.

Expansion terms in LM feedback are chosen by the odds of generating a term from the set of top ranked pseudo-relevant documents to that of generating it from the collection (Ponte, 1998).

$$LM(t) = \sum_{j=1}^{R} \frac{P(t|D_j)}{P(t)} \tag{2.13}$$

2.1.5 Limitations of Pseudo-Relevance Feedback

In this section, we review some of the limitations and risks associated with PRF. Despite these limitations, PRF on average improves the retrieval effectiveness over a set of queries. The limitations are discussed here because our work presented in the subsequent chapters of this thesis is motivated towards addressing some of these issues.

The first limitation of PRF is that it is highly parameter sensitive. The two main parametrs of PRF are: a) the number of top ranked documents assumed to be relevant, i.e. R, and b) the number of expansion terms, say T. A judicious choice has to be made while setting these parameters. Too many expansion terms can result in query drift, a phenomenon in which the information need expressed by the expanded query is very different from that expressed in the initial query (Mitra et al., 1998). Note that too many expansion terms may even have a negative effect on relevance feedback in general, and not only PRF in particular. This query drift can be visualized by imagining the modified query vector drifting further away from the centroid of the relevant documents. This may result in degraded retrieval quality after the feedback step. A high value of R, may increase the risk of falsely assuming more number of non-relevant documents as relevant. Extracting terms from these non-relevant documents may further introduce query drift (Wilkinson et al., 1995). It has been found that PRF degrades performance for a significant proportion of queries in a set of queries, particularly if most of the top ranked pseudo-relevant documents are actually not relevant to the query (Billerbeck and Zobel, 2004).

Some research even questions the usefulness of PRF in general (Billerbeck and Zobel, 2004). Many approaches have been proposed to increase the overall IR performance of PRF, for example by

- adapting the number of feedback terms and documents per topic (Ogilvie et al., 2009);
- selecting only good feedback terms after classifying terms into two classes, namely good and bad (Cao et al., 2008; Leveling and Jones, 2010); or
- increasing diversity of terms in pseudo-relevant documents by skipping feedback documents (Sakai et al., 2005).

The research questions RQ-1 and RQ-3, introduced in Section 1.2, seek to improve PRF effectiveness by addressing the limitation of partial relevance of documents in PRF (Wilkinson et al., 1995).

2.2 An Overview of Topic Modelling

This section builds up the necessary background for exploring research questions RQ-3 and RQ-4 (see Section 1.2). To recapitulate, RQ-3 explores whether discovering the topical structure of the pseudo-relevant set of documents can benefit the PRF process. The underlying hypothesis pertaining to RQ-3 is that the subtle aspects of the information need of a query manifest themselves as topics in the top ranked documents retrieved in response to the query. A discovery of this topic distribution may potentially improve PRF effectiveness by i) better predicting term associations with the query, and ii) providing a more uniform and comprehensive coverage of topics (query aspects) in the PRF. Moreover, the topic distribution in the pseudo-relevant set of documents can also help in providing a topic based access to information by visualization and navigation through these topics, as explored in research question RQ-4.

This section therefore provides an introduction to the topic modelling techniques in general, which is then followed by a survey of its applications in IR.

2.2.1 Topic Modelling

Intuitively speaking, topic modelling can be defined as a classification problem in which each term in a set of documents is assigned a membership class, the membership classes commonly known as the topics. Generally speaking, in contrast to the discriminative approach to the classification problem, where the output obtained from a classifier for a given test point is a class label, topic modelling techniques usually involve the generative approach, where the output form a classifier is the posterior probability of the class membership values. In particular, in the case of topic modelling these posterior probabilities of class (topic) membership values are computed for every word, the advantage of which is that a word can in theory belong to multiple topics with varying membership values.

It is expected that related terms, i.e. terms representing similar concepts, are categorized into the same topic. In general, most topic modelling algorithms use the common fundamental principle of discovering relatedness between terms through term co-occurrences with the hyopthesis that if two terms are highly related, they will co-occur frequently. We now provide a brief review of topic modelling techniques developed over the years.

Latent Semantic Analysis (LSA)

An initial attempt towards the construction of a topic model by automatically inferring latent relationships between terms by utilizing term co-occurrences from the term-document matrix of a collection, was latent semantic analysis (LSA) (Deerwester et al., 1990).

The intention in LSA is to represent the documents and the queries in a lower dimensional term space to capture term dependencies. The motivation for representing the vectors in a reduced dimensional space is as follows. Recall that in the vector space model (VSM) (see Section 2.1.3), both documents and queries are represented as vectors in a term space, with the assumption that each term is independent of the others (in mathematical terms, each term corresponds to a orthogonal dimensions in the term-space). Such a term space however, fails to capture the term dependencies such as the one between the terms *nuclear* and *atomic*. These dependencies can be however captured if the document and the query vectors are represented in a lower dimensional term space. In a lower dimensional term space, the cosine similarity between a document containing the term *atomic* and a query containing the term *nuclear* is higher in comparison to the similarity between them in the non-reduced term space. This is due to the fact that in a reduced dimensional space since *nuclear* and *atomic* are likely to belong to the same topical class due to a high likelihood of co-occurence, the separate dimensions *nuclear* and *atomic* should be compressed into one dimension representing a single concept. This leads to a non-zero similarity score between a document vector containing the word *nuclear* and a query vector containing the word *atomic*.

The working principle of LSA involves application of singular value decomposition (SVD), which is an orthogonal linear transformation technique, for reducing the rank of the term-document matrix. The objective of the SVD is to transform the original vectors into a reduced dimensional space such that the variances of the projection of the original vectors onto the reduced dimensions are maximized (Bishop, 2006, Chap. 12). It can be mathematically shown that maximizing the variances is in fact identical to finding a suitable set of orthogonal basis vectors in the reduced dimension such that the total projection error of the original vectors onto the reduced dimensional space is minimized.

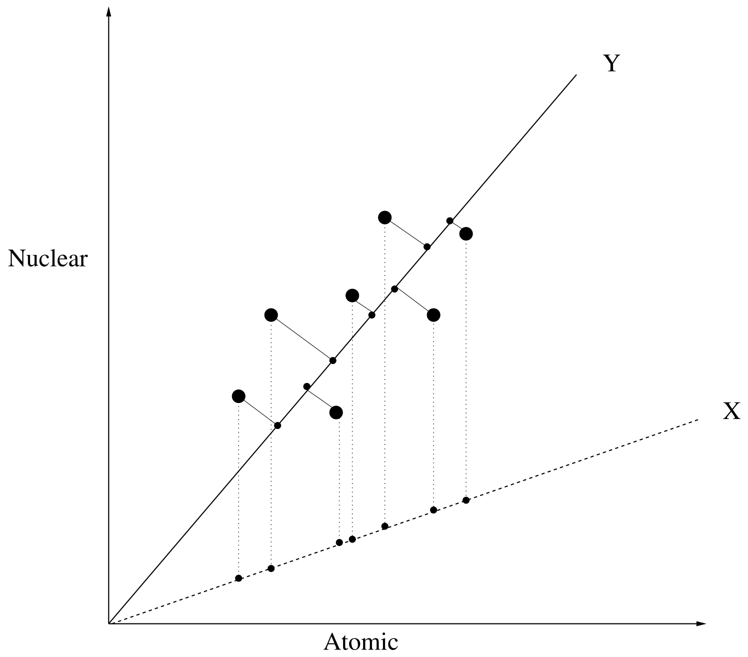


Figure 2.4: Illustration of the working principle of LSA on an example term-space of two dimensions.

We illustrate the working principle of LSA with a simple example. Figure **??** shows a sample term-space comprising of two terms, namely *atomic* and *nuclear*. The figure shows a few sample vectors containing both the terms *atomic* and *nuclear*. It can be clearly seen that these two dimensions are highly correlated. Intuitively speaking, one may visualize representing these vectors by their projections on a single dimension (a line). The question then is to determine the optimal line on which the vectors are projected so that the sum of variations of the projected values are maximized. The figure shows that the line Y is more preferable than the line X, in the sense that projected points on the line Y are farther apart than they are in X, implying that the sum of the variances of the projections on Y is higher than that in X. In this example, document or query vectors can thus be represented by their projections on the line Y which in the context of IR represents a concept rather than the two constituent terms *nuclear* and *atomic*.

We now discuss the limitations of LSA. Some of these are as follows:

- While LSA is able to capture dependencies between terms, it fails to explicitly capture the distribution of topics in a document.

- The SVD transformed term-document matrix can have negative values. While computation of cosine similarities is applicable also for vectors with negative components, it is somehow difficult to find a natural interpretation of these vectors with negative tf components in terms of the corresponding document compositions.

Probabilistic Latent Semantic Analysis (PLSA)

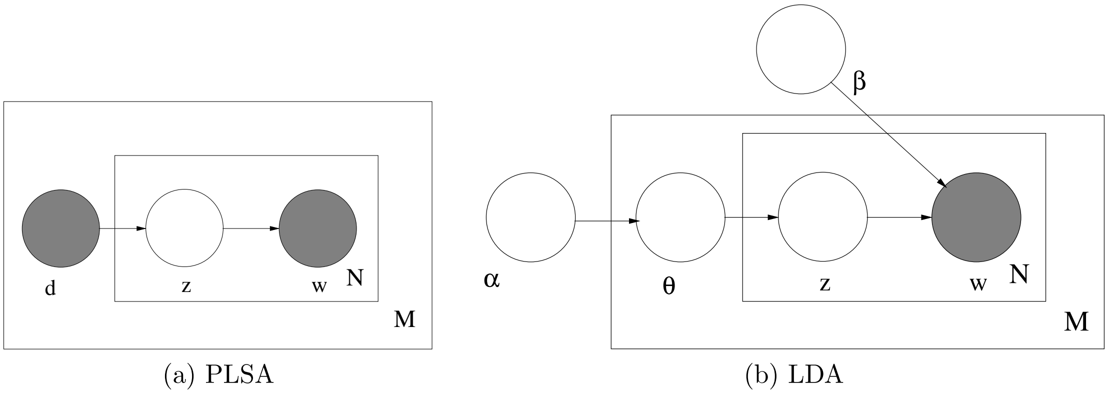
To overcome these limitations of LSA, probabilistic latent semantic analysis (PLSA) was proposed. PLSA is a probabilistic technique for topic modelling which treats each document as a mixture of multinomial distributions (Hofmann, 1999). Given a collection of M documents $\{D_1, \dots D_M\}$, where each document D_i is comprised of words drawn from the vocabulary $\{w_1, \dots w_N\}$, each word associated to a topic $z \in Z = \{z_1, \dots z_K\}$, PLA estimates a parametric generative model with the help of expectation maximization (EM) algorithm. The generative process, shown in Figure 2.5a and in Equation 2.14, works as follows.

- Select a document d with probability $P(d)$ (the prior probability of selecting a document).
- Select a topic class z with probability $P(z|d)$.
- Generate a word w in D with probability $P(w|z)$.

$$P(d, w) = P(d) \sum_z P(w|z) P(z|d) \tag{2.14}$$

Latent Dirichlet Allocation (LDA)

One of the major problems of PLSA is that it involves a large number of parameters In fact, the number of parameters grows linearly with the number of documents.



(a) PLSA

(b) LDA

Figure 2.5: Comparison between PLSA and LDA.

This is because the parameters for a k-topic pLSI model are k multinomial distributions of size N (each multinomial distribution representing topic-word mapping) and M mixtures over the k hidden topics (each mixture representing a document). The total number of parameters in PLSA is $KN + KM = K(N + M)$. This linear growth in parameters suggests that the model is prone to overfitting (Blei et al., 2003).

Latent Dirichlet allocation (LDA) overcomes this parameter explosion by introducing Dirichlet priors to the multinomials (Blei et al., 2003). LDA, similar to PLSA, assumes that each document is a mixture of multinomial topic distributions. However, the distribution of the topics themselves is assumed to follow a conjugate Dirichlet prior. The additional parameters introduced in the conjugate Dirichlet prior act as hyperparameters to control the multinomials for each document. In LDA, it thus suffices to estimate the hyperparameters instead of estimating each multinomial mixture for each document individually. The number of parameters in LDA is thus $K + KN$, i.e. in addition to the KN parameters for the topic-word mapping, LDA involves infering only K additional parameters for the Dirichlet prior of the topics, in contrast to PLSA where an additional KM parameters for the M documents need to be estimated.

The generative process in LDA, shown in Figure 2.5b works as follows.

- Choose a multinomial distribution $\theta^{(i)}$ with Dirichlet prior α for the i^{th} document, where $i = 1 \ldots M$ and $\theta^{(i)} \in R^K$.

- Choose a multinomial distribution $\phi^{(k)}$ with Dirichlet prior β, where $k = 1 \ldots K$ and $\phi^{(k)} \in R^V$.

- Choose the k^{th} topic in i^{th} document viz. z_{ik}, following the multinomial distribution $\theta^{(i)}$.

- The j^{th} word in i^{th} document is generated by following the multinomial distribution $\phi^{(z_{ik})}$.

The advantages of LDA over PLSA are:

- the presence of the Dirichlet priors for the multinomials tends to smooth out the distribution of words over topics; and

- fewer parameters avoid the problem of over-fitting.

LDA Inference

LDA inferencing involves estimating the parameters θ and ϕ, i.e. the document-topic and the term-topic associations respectively. Unfortunately, there is no closed form solution of the LDA corpus generation probability, and hence approximate inferencing techniques are used for estimating the distributions. Various inference techniques have been proposed for estimating the probabilities including the variational Bayes (VB) (Blei et al., 2003), expectation propagation (EP) (Minka and Lafferty, 2002) and Gibbs sampling (Griffiths and Steyvers, 2004). Gibbs sampling for infering LDA has been shown to be computationally faster and also outeprforming the other two, i.e. VB and EP, in approximating the posterior more accurately (Griffiths and Steyvers, 2004).

Gibbs Sampling for LDA

We now briefly introduce the series of steps of Gibbs sampling is applied to infer the posterior probabilities in the particular case of LDA. Below we list the computational steps of Gibbs sampling to estimate the topic-word (ϕ) and the document-topic (θ)

relationships, which in turn are applied for PRF and topic visualization for our work involving research questions RQ-3 and RQ-4, respectively.

Let the number of words in the i^{th} document d_i assigned to the j^{th} topic be $n_j^{(d_i)}$, and the number of instances of word w assigned to the j^{th} topic z_j be $n_w^{(z_j)}$. Instead of explicitly representing θ or ϕ as parameters to be estimated, the Gibbs sampling approach to LDA inferencing considers the posterior distribution of the assignments of words over topics, namely $P(w|z)$. Generally speaking, Gibbs sampling involves estimating a multivariate distribution after a number of iterations by randomly sampling from a conditional univariate distribution, where all the random variables but one are assigned fixed values (Griffiths and Steyvers, 2004; Geman and Geman, 1987). This general principle of Gibbs sampling, when applied to LDA in particular, involves computing the conditional distribution $P(z_i|z_{-i}, w)$, i.e. the current topic variable z_i conditioned on all the rest of the topic variables excluding z_i (denoted by z_{-i}). For LDA, this is given by

$$P(z_i = j|z_{-i}, w) \propto \frac{n_{-i,j}^{(w_i)} + \beta}{\sum_{k \neq i} n_{-i,j}^{(w_k)} + V\beta} \cdot \frac{n_{-i,j}^{(d_i)} + \alpha}{\sum_{z_k \neq j} n_{-i,k}^{(d_i)} + K\alpha} \tag{2.15}$$

In Equation 2.15, the n_{-i} values denote the counts not including the current assignment of z_i. The first ratio in Equation 2.15 expresses the probability of w_i under topic j, and the second ratio expresses the probability of topic j in document d_i. The z_i variables are initialized to values in $\{1, 2, ...K\}$ randomly. The sampling process is then repeated for a series of iterative steps each time finding a new state by sampling each z_i from the conditional distribution specified in Equation 2.15. After a sufficient number of iterations, which is typically around 1000, the estimates of θ and ϕ are obtained using the current assignments of z_is as shown in Equation (2.16) and (2.17). For more details on LDA inference by Gibbs sampling, the reader is referred to (Griffiths and Steyvers, 2004).

$$\hat{\theta}_j^{(d_i)} = \frac{n_j^{(d_i)} + \alpha}{\sum_{j'=1}^{K} n_{j'}^{(d_i)} + K\alpha}, \; i = 1 \ldots M, j = 1 \ldots K \tag{2.16}$$

$$\hat{\phi}_w^{(z_j)} = \frac{n_w^{(z_j)} + \beta}{\sum_{w'=1}^{V} n_{w'}^{(z_j)} + V\beta}, \ j = 1 \dots K, w = 1 \dots V \tag{2.17}$$

Using the estimates of $\hat{\theta}$ and $\hat{\phi}$, the probability of generating a word w from the i^{th} document d_i is obtained by marginalizing over the latent topic variables z_js as shown in Equation (2.18).

$$\begin{aligned} P_{LDA}(w|d_i, \hat{\theta}, \hat{\phi}) &= \sum_{j=1}^{K} P(w|z_j, \hat{\phi}) P(z_j|d_i, \hat{\theta}) \\ &= \sum_{j=1}^{K} \frac{(n_w^{(z_j)} + \beta)(n_j^{(d_i)} + \alpha)}{(\sum_{w'=1}^{V} n_{w'}^{(z_j)} + V\beta)(\sum_{j'=1}^{K} n_{j'}^{(d_i)} + K\alpha)} \end{aligned} \tag{2.18}$$

In the context of our work described in this thesis, we use the closed form approximation of $P_{LDA}(w, d)$ in Equation 2.18 to smooth the relevance model (cf. Section 2.1.4) in relation to the research question RQ-3, and also use the word-topic and the document-topic mappings θ and ϕ for the topic visualizations relating to research question RQ-4.

2.2.2 Applications of Topic Modelling in IR

After an introduction to the topic modelling in general, we now turn our attention in this section to applications of topic modelling in the particular domain of IR. We also highlight the major differences of the existing work with our work pertaining to the research questions RQ-3 and RQ-4.

The unigram document models in LM based retrieval have been extended to cluster-based document models, where it is assumed that a word in addition to being generated from a document d or the collection, as in a standard LM (Hiemstra, 2000), can also be generated from a cluster of documents containing d, say C_d , that is the cluster of documents on topics similar to d (Liu and Croft, 2004). Equation 2.4 can thus be extended to

$$P(q|d) = \prod_{t \in q} \lambda P_{MLE}(t|d) + \mu P_{MLE}(t|C_d) + (1 - \lambda - \mu) P_{coll}(t) \tag{2.19}$$

In Equation 2.19, the smoothing in language modelling (LM) retrieval is performed with the help of the cluster model in addition to the collection model. This method thus groups together documents which share topics. A limitation of clustering is that a document can be only belong to a single cluster.

The use of LDA in LM retrieval was investiaged in (Wei and Croft, 2006). In contrast to using unigram document language models of Equation 2.4, Wei and Croft (2006) employed Equation 2.18 as the term sampling model for a document d in LM retrieval. The authors call this approach the "LDA based document models" (LBDM). LBDM involves estimating LDA over the whole collection of documents by Gibbs sampling, and then linearly combining the standard LM term weighting with LDA-based term weighting as shown in Equation 2.20. The reason to use linear combination was due to the fact that LDA itself may be too coarse to be used as the only representation for IR. In fact, optimal retrieval effectiveness on ad-hoc search is reported with setting the proportion of LDA to 0.3, i.e. setting $\mu = 0.3$ in Equation 2.20, and a complementary proportion of 0.7 for standard LM weighting (Wei and Croft, 2006).

$$P(q|d) = \prod_{t \in q} (1-\mu) P_{LDA}(t|d, \hat{\theta}, \hat{\phi}) + (1-\mu)\big(\lambda P_{MLE}(t|d) + (1-\lambda) P_{coll}(t)\big) + \quad (2.20)$$

Our work relating to the third research question RQ-3 overcomes the coarseness of the topic representation limitation by restricting LDA to only the top ranked pseudo-relevant set of documents. This also makes the estimation a lot faster since our method does not require the LDA estimation to be conducted on a whole collection of documents, which is typically much larger in the case of IR than the typical collection sizes used in LDA estimations for other application domains of LDA (Blei et al., 2003). Another major difference to (Wei and Croft, 2006) is that we do not linearly combine LDA document model scores with the unigram language model scores. We rather rely on the KL divergence between the estimated relevance model and the document language model to re-rank each document similar to the principle of relevance modelling. Thus, our work does not require an extra parameter for the linear combination, i.e. the parameter μ in Equation 2.20, which makes optimization easier.

Zhou and Wade (2009) applied LDA in the feedback step to re-rank documents retrieved in the initial step. LDA estimation was based on the pseudo-relevant document space, instead of the whole collection as in (Wei and Croft, 2006), by using a linear combination of the initial retrieval LM score and the KL divergence between a document model and the constructed LDA model. Although Zhou and Wade (2009) provide empirical evidence of the effectiveness of the LDA technique in reranking initial retrieval results, their method in general lacks a theoretical justification. Our work is an attempt to provide a formal justification to the same principle.

Recent extensions to the RLM involving inference of query term dependencies by training (hierarchical) Markov random fields (MRF) were proposed in (Metzler and Croft, 2007; Lang et al., 2010). These MRF models require a training phase to learn the model parameters. A retrieval evaluation metric, such as MAP, is used directly as the objective function to be maximized in the learning phase. However this in turn, makes such models dependent on the availability of a set of training queries with manual relevance assessments.

Our work towards RQ-3 does not require a separate training phase and hence is not restricted to work in the presence of a training set of queries with relevance assessments. Moreover, our model is motivated from information needs in queries rather than term dependencies. Another major difference is that the Markov random field models have not been tested on very long queries such as patent prior art search, whereas our proposed model is evaluated on both short and very long queries.

2.3 Summary

This chapter provided a brief overview of IR, introducing the retrieval models such as the VSM, BM25 and the LM. We also introduced the (pseudo-) relevance feedback approaches for the various retrieval models. Out of the different feedback methods discussed, the relevance model (RLM) is of particular interest to us because in our work related to research question RQ-3, we propose an extension to the RLM by the use of topic modelling. Moreover, since topic modelling forms a core part of our research work involving research questions RQ-3 and RQ-4, this chapter also provides an introduction to topic modelling in general and its applications to IR in particular. The topic modelling technique, which will be of particular interest to us throughout the course of this thesis, is LDA. LDA is an unsupervised model which can estimate the topic distribution in a collection of documents to more accurately than its predecessors such as the PLSA and LSA. In our work involving the research questions RQ-3 and RQ-4, we apply LDA for improving the retrieval effectiveness of ad-hoc search through PRF, and for topic visualization in a search interface, respectively.

With this background, we are now ready to move onto the next chapter where we describe the framework of the experimental investigations carried out in the subsequent chapters of this thesis.

Chapter 3

Evaluation Framework

An important factor in evaluating IR methods is in the judicious choice of the evaluation framework used to test the proposed methods. The evaluation framework in the context of our study reported in this thesis needs to include test collections with i) short keyword type queries, and ii) very long queries where each query is comparable in length to that of each document in the collection. These correspond to the first two research questions, investigating document and query segmentation, respectively.

To recapitulate, in research question RQ-1, we seek answer to whether document segmentation can improve retrieval quality. The dataset used to test the document segmentation method is the TREC dataset, which is a standard IR test collection comprising of news articles. The hypothesis that documents as a whole are seldom relevant is in general true for the TREC dataset.

In the second research question we seek to explore how retrieval with long queries can be improved. A conventional document collection such as the TREC dataset, where queries are typically very short comprising of a few keywords, is thus not suitable for setting up the evaluation framework for RQ-2. A suitable test collection to explore this research question is the patent document collection, namely the CLEF-IP 2010 dataset.

In the third research question, namely RQ-3, we attempt to combine the document and query segmentation techniques under a single model. Hence, our experiments on RQ-3 uses both the TREC and the CLEF-IP datasets.

In the fourth research question, RQ-4, which seeks to explore techniques of providing a topic-based information access to the users, we use the CLEF-IP 2010 dataset. The reason we chose the CLEF-IP dataset is that since the patent documents and queries typically comprise of a mixture of topics, it is particularly interesting to see the effects of visualization of the topics in the retrieved documents and the query. We believe that it would be convenient for a patent examiner in validating or invalidating prior art claims by visualizing the constituent topics in the retrieved documents and the query.

This chapter is organized as follows. It starts with a brief review on the standard IR evaluation methodology by introducing the concepts of document collections, query test sets and standard evaluation metrics. We then describe setting up of the evaluation framework for our experimental investigations described in the subsequent chapters of this thesis. In particular, we describe the characteristics of the datasets, tools and other resources used for our experiments.

3.1 IR Evaluation

Chapter 2 introduced the component stages of a standard IR system, techniques for document/query representation and the comparison between these using retrieval models ranging from the simple tf-idf weighting to more involved techniques such as the BM25 and LM term weightings. These choices available in construction of an IR system make it a highly empirical discipline requiring careful and thorough evaluation of retrieval effectiveness using representative test collections. In this section, we introduce the notion of test collection and evaluation metrics.

Evaluation of IR is more challenging than it might appear at a first glance. One may imagine an IR task to be somewhat analogous to a binary classification problem in which the documents retrieved in response to a query have to be classified to either of the two classes, namely *relevant* or *non-relevant*. Hence one may be inclined to believe that an IR system can be evaluated by a simple metric such as the ratio of the number of relevant documents returned to the number of non-relevant ones, as can be done in a binary classification problem. In fact, this measure can be applied to evaluate non-ranking retrieval models, such as the Boolean retrieval model introduced in Section 2.1.3. A careful analysis however reveals that such an evaluation methodology is inaccurate for the ranked retrieval models. The reason is as follows. The rank at which a relevant document is returned, is of utmost importance as far as user satisfiablity is concerned because users typically look only at a few top documents in the ranked list (Baeza-Yates et al., 2005). Since the eventual objective of IR evaluation is to approximate level of user satisfaction with behaviour of the IR system as accurately as possible, an IR evaluation method has to take into consideration the ranks at which relevant documents are retrieved.

Formal laboratory evaluation of an IR system typically follows the Cranfield paradigm (Cleverdon, 1960, 1991). The Cranfield paradigm involves the creation and use of standard test collections for evaluating effectiveness of IR systems. Automatic evaluation involves comparing the documents as returned by an IR system with a set of manually evaluated relevant documents.

The standard components of an IR test collection are:

- A collection of documents typically comprising of text in a domain in which the IR system is intended to be used, a set of test user search queries typical of expected user behavious, and corresponding relevance judgements, which list the relevant documents for each query.

- One or more suitable evaluation measures for quantification of retrieval effectiveness.

- A statistical methodology that determines whether the observed differences in performance between the methods investigated are statistically significant. (Hull, 1993).

We first introduce the components of a standard IR test collection. This is followed by a discussion of automatic measurement of the effectiveness of an IR system, and then we briefly describe the statistical significance testing methodology used in IR evaluation.

3.1.1 Test collection components

Document collection The document collection for an IR test collection is a static set of documents typical of the search task to be evaluated, e.g. web content, news articles, medical reports etc.

In order to make the retrieval task tractable on one hand and challenging on the other, the documents in standard collections, such as the TREC, are usually of varied lengths, varied writing styles, varied levels of editing, varied timeframes and a varied vocabulary (Harman, 1993).

Query collection The topics in an IR evaluation test set should mimic a real user's need and should reflect typical query bbehaviour of the target users of the IR system. Standard test collections, such as the TREC, comprise of queries queries from diverse domains in order to ensure a fair and comprehensive comparison between different IR systems. Moreover, queries in standard test collections represent information needs with variable granularities ranging from very specific, e.g. *osteoporosis*, to more general ones, e.g. *bone disease*.

The performance of an IR system under evaluation needs to be averaged over a set of queries in order to ensure statistical reliability of the results. As a rule of thumb, 25 information needs has usually been found to be a sufficient minimum (Buckley and Voorhees, 2000).

Relevance judgements To accurately measure the effectiveness of a retrieval system in response to a query, ideally the relevance of every document in the collection should be known.

However, this is not achievable in practice for practical document collections due to the impossible manual effort that would be required. The challenge is then to approximate the set of complete relevant documents as accurately as possible. To this effect, an incomplete set of relevance judgements is obtained by a process called *pooling*, where the main idea is as follows. A pool of documents is constructed by taking the union of top ranked documents from the retrieval runs which are to be evaluated. Assuming that each retrieval run returns a finite number of documents in response to the query, and that there is sufficient amount of overlap between the documents retrieved, the number of documents that need to be judged is kept manageable (Harman, 1993). The assumption behind the working principle of pooling is that any relevant documents which are not retrieved within top ranked documents by any of the retrieval systems, will not have a significant impact on the measured retrieval effectiveness of an IR system or its performance relative to other systems.

After constructing the pool of documents, human accessors examine each document of the pool in turn. The human relevance assessments are typically made on a scale of relevant, partially relevant and non-relevant.

3.1.2 Evaluation metrics

The intention of an IR evaluation metric is to measure satisfaction of a user's information need in user satisfaction, for the purposes of laboratory evaluation, it is assumed that relevance alone should be the focus of IR evaluation.

Precision and Recall

Under the relevance ranking paradigm, the effectiveness of an IR system is measured by its ability to retrieve all and only relevant information. These two aspects of IR quality correspond to: i) precision, which measures how much relevant content is there in the retrieved set, and ii) recall, which measures how much relevant content have been retrieved from the collection. These concepts can be seen clearly in the contingency table shown in Table 3.1.

Table 3.1: Contingency table for precision-recall

	Relevant	Non-relevant
Retrieved	True positive (tp)	False positive (fp)
Non-retrieved	False negative (fn)	True negatives (tn)

The definitions of precision and recall are thus

$$P = \frac{tp}{tp + fp} \tag{3.1}$$

$$R = \frac{tp}{tp + fn} \tag{3.2}$$

It is easy to apply the above definitions of precision and recall to retrieval systems returning sets of documents, such as the Boolean retrieval model introduced in Section 2.1.3. However, the major problem with Boolean retrieval, as already pointed out in Section 2.1.3, is that it does not conform to the user prefered mode of information access in the form of a ranked list of documents sorted by relevance (Baeza-Yates et al., 2005). It is thus important to extend these definitions of precision and recall to ranked lists.

The set-based definition of precision, as shown in Equation 3.2 can be applied to ranked lists by cutting the ranked list to sets of top k documents. Hence the quality of a ranked list of documents is often measured by the simple metric P@k denoting precision at top k documents, or the number of relevant documents found in the top k documents. However, this metric fails to distinguish the quality of retrieval runs by the ranks of relevant documents, e.g. a ranked list of documents with relevant documents at ranks 2, 3, 4 should be rated better than a one with relevant documents at 7, 8, 9, although P@10 for both is 3/10. Furthermore, P@k does not take into account the recall factor, which in fact can be ignored for cases such as the web search.

Mean average precision (MAP)

A common metric of IR quality measurement, which combines the two aspects of precision on ranked lists and recall, is the mean average precision (MAP). For a single information need, average precision (AP) is the average of the precision values obtained for the top set of k documents, after each relevant document is retrieved. This value, when averaged over a query set, is called the mean average precision (MAP). The fact that a non-zero component, i.e. the precision at k^{th} rank, is added for every relevant document retrieved at rank k, tends to favour retrieval results with higher recall and relevant documents retrieved at lower (better) ranks. The mathematical expression for AP for a single query is shown in Equation 3.3, and its average over the set of queries, i.e. MAP, is shown in Equation 3.4. Let, for example, the relevant set of documents for a query $q \in Q$, be $\{d_1, \ldots, d_m\}$, m being the total number of relevant documents (not necessarily retrieved) for the query q, and R_k denoting the ranked list of documents from d_k to d_1.

$$AP(q) = \frac{1}{m} \sum_{k=1}^{m} \mathrm{P@R_k} \tag{3.3}$$

$$MAP(Q) = \frac{1}{|Q|} \sum_{q \in Q} AP(q) \tag{3.4}$$

Coming back to our earlier example, introduced in the previous section, AP of the ranked list $\{2, 3, 4\}$ is $1/3(1/2 + 2/3 + 3/4) = 0.638$, whereas AP of the ranked list $\{7, 8, 9\}$ is $1/3(1/7 + 2/8 + 3/9) = 0.242$. This clearly shows that MAP prefers ranked lists with more relevant documents at early ranks.

F-measure

Although MAP was designed to address both aspects of retrieval quality, namely precision and recall, yet it can be argued that MAP is more biased towards precision than recall, as can be seen by the progressively smaller amounts of each contribution added to the AP value for increasing ranks of relevant documents. A solution is to use a weighted harmonice mean of the precision and recall through the metric, called the F-measure metric, shown in Equation 3.5.

$$F = \frac{1}{\alpha\frac{1}{P} + (1-\alpha)\frac{1}{R}} \quad \text{where } \alpha \in [0,1] \tag{3.5}$$

Setting α to 0.5 in Equation 3.5 equally balances precision and recall. This metric however, cannot be applied to ranked lists since it is a set-based measure. Moreover, it is also not sensible to apply the F-measure on top k cut-off sets because in most retrieval domains, such as the patent search, recall is expected to be low within the top k retrieved documents.

Patent retrieval evaluation score (PRES)

A recall oriented metric, recently devised for patent retrieval, is the patent retrieval evaluation score (PRES), which overcomes the excessive precision biasness of MAP (Magdy and Jones, 2010b). This metric is inspired by another recall oriented metric R_{norm}. We illustrate the working principle of PRES in Figure 3.1. The figure plots the retrieved ranks along the x-axis and the absolute recall values in the y-axis. Retrieval of a relevant document steps up the recall. In the best case, the relevant documents can all be retrieved at the top ranks, shown by the left most plot marked *best*. A parameter N denotes the maximum number of documents in the ranked list that are to be checked manually for relevance. The plots are therefore cut-off at this point, as shown by the vertical red line. A ranked list in PRES is simply evaluated by how close the actual plot is to the best case plot, by computing the ratio $\frac{A}{B}$, as shown in the figure.

The standard tool for IR evaluation given retrieval results for a set of topics and the corresponding relevance assessments is *trec_eval*[1]. It generates output on performance metrics such as mean average precision and recall at different cut-off

[1] `http://trec.nist.gov/trec_eval/`

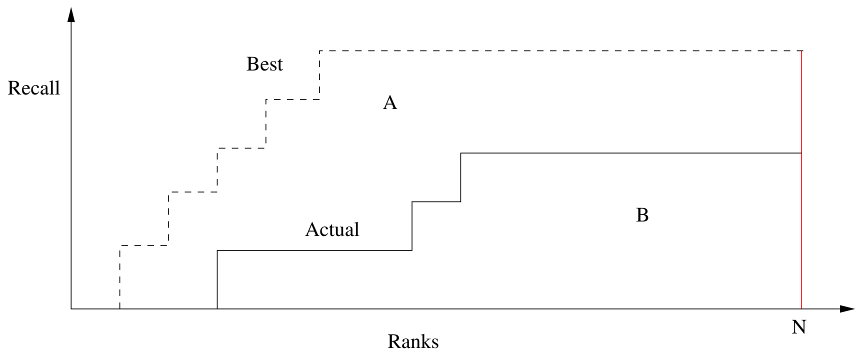


Figure 3.1: A graphical example for PRES measurement.

points. We use *trec_eval* to compute the standard metrics such as MAP, precision at fixed cut-off points such as 5 or 10 for our experimental investigations. Following the recommendation of (Magdy and Jones, 2010b), we use PRES to evaluate our investigations with respect to recall.

3.1.3 Significance tests

It is not possible to conclude from the percentage improvements of one method over the another that whether the improvements are genuinely due to the superiority of one method over the other, or is a case of random fluctuations in performance for each query, as might be the case because one retrieval model may suit a particular type of query, say queries with broad information need whereas the other may perform well for queries with a more specific information need. The retrieval effectiveness in IR thus has to be measured by statistical significance tests.

A reasonable amount of sample points is required in a significance testing to disprove the null hypothesis H0, which for IR is representative of the fact that one method is not better than the other. In the case of IR, the sample points may refer to the average precision (AP) values for $|Q|$ individual queries and the null hypothesis H0 is that there is no difference between method A and method B (Hull, 1993). The idea is to show that, given the data, the null hypothesis is incorrect, because it leads to an implausible low probability. Rejecting H0 implies accepting the alternative hypothesis H1. The alternative hypothesis, H1, for the retrieval experiments will be that either method A consistently outperforms method B, or method B consistently outperforms method A. Two methods A and B are distinguishable if either the left tail[2] , or the right tail of the distribution confirms H1, i.e. A is better than B or B is better than A respectively.

The following paired tests are used in the SMART retrieval experiments (Salton and McGill, 1984).

- The paired *t-test* assumes that errors are normally distributed. H0 follows the Student's t distribution with $|Q| - 1$ degrees of freedom.
- The paired *Wilcoxon's signed ranks test* is a non-parametric test that assumes that errors come from a continuous distribution that is symmetric around 0. This test uses the ranks of the absolute differences instead of the differences themselves.
- The paired *Wilcoxon sign test* is a non-parametric test which only uses the sign of the differences for each sample point. The test statistic follows a binomial distribution.

We do not use the t-test because the t-test assumes that the errors are normally distributed. However, it is not reasonable to assume that precision and recall are normally distributed since they are discrete measurements (Hull, 1993). Thus, for the experiments in this thesis, we employ the Wilcoxon signed ranks test, which we simply refer to as Wilcoxon test henceforth.

3.2 Evaluation Test Collections

In this section, we introduce the two datasets, namely i) the TREC collection - volumes 4 and 5, comprised of news collection used for exploring research questions

[2] The tail of a distribution refers to the region under the extreme ends of a distribution where the probability mass is usually low. In a Gaussian distribution for example, the left (right) tail represents the small area towards the left (right) end under the bell curve.

RQ-1 and RQ-3; and ii) the CLEF-IP 2010 collection comprising of patent articles, used for investigating research questions RQ-2, RQ-3 and RQ-4.

3.2.1 Document Collections

TREC document collection.

The research questions RQ-1 and RQ-3 examine the utilization of document segmentation for improving the retrieval performance in ad-hoc IR. Hence, for this purpose, we use the standard TREC dataset[3], which has been extensively used for ad-hoc IR experiments over the years.

The TREC collection was compiled by the NIST[4]. The document collection used for our experiments is available in disks 4 and 5. It comprises of a total of 528542 newspaper and newswire data from four different sources, namely the Federal Register, Los Angeles Times, Foreign Broadcast Information Services and the Financial Times. The volumes 4 and 5 of the data collection correspond to the TREC 6,7,8 and the robust track query sets. Table 3.2 outlines the document collection characteristics, and also lists the query set names.

Table 3.2: Document collection statistics

Collection Name	# Documents	Type	Test Topic Set
TREC vol 4 & 5	528542	American news articles	TREC-6, TREC-7, TREC-8, TREC Robust
CLEF-IP 2010	1327489	European patents	CLEF-IP 2010

Each TREC document has beginning and end markers, and a unique "DOCNO" field containing a unique document identifier. The documents are uniformly formatted into an SGML-like structure, as can be seen in Figure 3.2 (Harman, 1993).

The query set TREC-6 is chosen as the training set for choosing the parameters, whereas the other datasets are used for testing, the parameters being set to the optimal values obtained on TREC-6. This way of splitting up the query set into

[3] `http://trec.nist.gov/`

[4] `http://www.nist.gov/index.html`

```
    <DOC>
    <DOCNO> W5J880406-0090 </DOCNO>
    <HL> AT&T Unveils Services to Upgrade Phone Networks Under Global
Plan </HL>
    <AUTHOR> Janet Guyon (WSJ Stafi) </AUTHOR>
    <DATELINE> NEW YORK </DATELINE>
    <TEXT>
    American Telephone & Telegraph Co. introduced the first of a new genera-
tion of phone services with broad implications for computer and communications
equipment markets. AT&T said it is the first national long-distance carrier to
announce prices for specific services under a world-wide standardization plan to
upgrade phone networks.
.
.
.
    </TEXT>
    </DOC>
```

Figure 3.2: An excerpt from a TREC document.

separate training (50 out of 250 queries) and test (remaining 200 queries) sets ensures that there is a less chance of overfitting due to parameter selection (Bishop, 2006, Chap. 1). The decision to choose TREC-6 as the training set was arbitrary, rather than due to any specific characteristics of this particular query set.

CLEF-IP document collection.

The research quesion RQ-2 involves exploring query segmentation to improve retrieval effectiveness for very long queries. The TREC queries are too short to gain any benefits and rather inappropriate to test our proposed methods of segmenting a query. It is therefore more appropriate to test our method on genuinely large queries. An instance of such a collection is a patent search collection in which the queries being new patent claims are comparable in length to the existing patent articles of the collection. For our exeriments in particular, we chose the patent collection from CLEF-IP 2010, which is an evaluation campaign for evaluating patent search[5]. The document collection in CLEF-IP 2010 comprises of patents filed with the European Patent Office (EPO) in three languages, namely English, German and French. We restrict our investigation to the English subset of the collection (68% of the full collection), and without loss of generality, refer this subset as the CLEF-IP

[5] http://www.ir-facility.org/clef-ip

```
<patent-document ucid=“EP-1158672-A2” country=“EP” doc-number=“1158672”
kind=“A2” lang=“EN” family-id=“27343456” status=“new” date-produced=“20090516”
date=“20011128”>
<bibliographic-data>...</bibliographic-data>
<abstract load-source=“ep” status“new” lang=“EN”>
<p>
A longitudinally coupled resonator type surface acoustic wave filter (1) includes a piezoelectric substrate (2) and first (6), second (7) and third (8) IDTs provided on the piezoelectric substrate and arranged in a surface wave propagating direction such that the second IDT (7) is interposed between the first and the third IDTs (6,8). ...
<img id=“img-00000001” orientation=“unknown” wi=“109” img-format=“tif”
img-content=“ad” file=“00000001.tif” inline=“no” he=“86”/>
</p>
</abstract>
<description load-source=“ep” status=“new” lang=“EN”>
<heading>BACKGROUND OF THE INVENTION</heading>
<heading>1. Field of the Invention</heading>
<p num=“0001”>
The present invention relates to a longitudinally coupled resonator type surface acoustic wave filter, and more particularly, to a longitudinally coupled resonator type surface acoustic wave filter, in which at least three IDTs are arranged in a surface wave propagating direction.
</p>
...
</description>
<claims load-source=“ep” status=“new” lang=“EN”>
<claim num=“1”>
<claim-text>
A longitudinally coupled resonator type surface acoustic wave filter comprising:
<claim-text>a piezoelectric substrate (2); and</claim-text>...
</claim>
</claims>
<copyright>...</copyright>
</patent-document>
```

Figure 3.3: An excerpt from a CLEF-IP 2010 patent document.

2010 henceforth.

Each patent document in the CLEF-IP 2010 collection is a structured document consisting of several sections. A document is structured with XML tags which correspond to the sections in the patent document and some additional metadata such as dates, addresses, agencies, document versions, etc. The most important section of a patent document is the *claims* section. Each patent contains at least one claim. The claim section defines the scope of protection granted by the patent and the specific novel aspects of the invention that need to be protected. Other sections present in a CLEF-IP patent document are as follows:

- *title* comprising of a few keywords represents the title of the invention.

Table 3.3: Query set characteristics

Qry set	Fields	# Queries	Avg. qry length	Avg. # rel. docs
TREC 6	title	50	2.48	92.22
TREC 7	title	50	2.42	93.48
TREC 8	title	50	2.38	94.56
TREC Robust	title	100	2.88	37.20
TREC Robust	title, description, narrative	100	17.88	37.20
CLEF-IP 2010	title, abstract, claims, description	50	9278.24	37.98

- *abstract* summarizes the key contributions of the claimed invention in a single paragraph.
- *description* encapsulates the detailed description of the invented techniques.
- *citation* points to other works reated to this invention.

A sample patent document from the CLEF-IP 2010 collection is shown in Figure 3.3.

3.2.2 Query sets

After describing the document collection, we now describe the query set details in this section.

TREC queries

Table 3.3 gives details of the TREC topics used as queries for our experiments. A TREC query has three different fields namely the *title* which typically comprises of a few keywords, the *description* which comprises of a few natural language sentences which describe the information need in more detail, and the *narrative* which explicitly describes the required criteria that a relevant document must possess. Retrieval experiments can be performed using any combinations of fields from a TREC topic. The standard combinations, reported in many IR investigations, involve using the title only (T), or the title and the description together (TD).

A sample TREC query is shown in Figure 3.4. In this example query, the user information need is to retrieve documents on polio disease data, information on the

<top>
<num> Number: 302
<title> Poliomyelitis and Post-Polio
<desc> Description:
Is the disease of Poliomyelitis (polio) under control in the world?
<narr> Narrative:
Relevant documents should contain data or outbreaks of the polio disease (large or small scale), medical protection against the disease, reports on what has been labeled as "post-polio" problems. Of interest would be location of the cases, how severe, as well as what is being done in the "post-polio" area.
</top>

Figure 3.4: A TREC query.

outbreaks of the disease, medical protection against the disease etc. as explicitly expressed in the narrative of the query. However in practice, it is unlikely for a user to enter such a detailed query. The information need of the user is often represented by a short set of keywords, of which the title field is intended to be representative. The challenge of the retrieval systems is then to retrieve relevant content for the query. Since the title of a query is the most common type of queries encountered in real-life search scenarios (Baeza-Yates et al., 2005), it is a standard practice to use the T queries for IR experiments. Our test set of queries thus also use the T field alone discarding the description and narrative fields. The TREC Robust track topics are included in our experimental investigations because these are particularly interesting for testing PRF methodologies since these are known instances of queries which are difficult to improve with query expansion (Voorhees, 2004). A reason for this difficulty can possibly be attributed to the fact that the average number of relevant documents for the TREC Robust queries is much less as compared to the TREC 6, 7, and 8 query sets (see Table 3.3).

Another reason to choose the TREC Robust topics is that these queries have already been analyzed for common causes of failures and some of the queries have been categorized into failure classes with increasing levels of difficulty and natural language understanding (Harman and Buckley, 2004). It is thus worthwhile to benchmark the retrieval effectiveness of a new PRF method on these difficult queries. Table 3.4 summarizes the analyzed failure classes.

Table 3.4: Buckley's failure analysis on the TREC Robust topic-set.

Category number	Class Description	# queries
2	General technical failures such as stemming	2
3	Systems all emphasize one aspect, miss another required term	3
4	Systems all emphasize one aspect, miss another aspect	7
5	Some systems emphasize one aspect, some another, need both	5
6	Systems all emphasize some irrelevant aspect, missing point of topic	1
7	Need outside expansion of "general" term	3
8	Need query analysis to determine relationship between query terms	1
9	Systems missed difficult aspect	6

CLEF-IP queries

In general, a patent topic is much longer than the topics in other standard IR tasks, such as the ad-hoc search. In particular, a topic in the CLEF-IP 2010 collection is similar in structure to a document in the collection. As a matter of fact, the only difference between a document in the collection and a query in this case is that whereas a document is a granted patent, a query is a patent submitted to the patent office.

A topic in the CLEF-IP 2010 dataset comprises of the *title*, *abstract*, *claim* and *description* sections. The *abstract* summarizes the invention, each *claim* field describes a novel invention, and the *description* field provides technical details regarding the invention. The objective of patent prior art search is to find all patents relevant to the query, which in this case is a new patent application, potentially invalidating the novelty in the claims of the new patent application.

3.3 Framework for Experimental Investigation

In this section, we decsribe the tools and resources used for conducting our experiments reported in this thesis. All our IR experiments are conducted within the framework of the SMART system extended to include language modelling (LM) retrieval model. The following section provides an overview of the extended SMART sytem used for the experimental Investigations in the subsequent chapters of this thesis.

3.3.1 An overview of the SMART system.

The SMART[6] retrieval engine is an open-source IR engine implemented in C. It was originally designed for retrieving using the VSM, but it provides a general framework to implement term weightings for other retrieval models. In particular, we implemented the LM IR model in SMART using the method described in (Hiemstra, 2000)[7]. SMART supports the following text indexing functionalities.

- **Tokenization:** The text inside specified tags is tokenized into individual words and other special tokens such as the hyphens, underscores etc.
- **Stopword removal:** Frequently occuring words such as *the*, *of* etc., known as stopwords, are removed from the list of tokens. The SMART system uses a pre-defined list of 571 stopwords. For CLEF-IP 2010, in addition to using a standard list of stopwords[8], we also removed formulae, numeric references, chemical symbols and patent jargon such as *method*, *system*, *device* etc.
- **Stemming:** Various morphological variations of a word can be normalized to the same stem. Simple rules of suffix stripping are usually used for this process. For our experiments, we used the default stemmer in SMART, which is a variant of the Lovin's stemmer (Lovins, 1968).
- **Phrase formulation:** Optionally, a pre-defined list of phrases can be used as additional vocabulary, i.e. these phrases can be indexed as separate entities (Singhal, 1997). For our experiments on the TREC data, we used the standard phrase list of SMART, which comprises of $150,000$ most frequent phrases extracted from the TREC documents (Singhal, 1997). Our experiments on the CLEF-IP 2010 data does not use phrases as indexing units because the phrase list available in SMART has been constructed from the TREC documents.

[6] `ftp://ftp.cs.cornell.edu/pub/smart`

[7] `http://www.computing.dcu.ie/~dganguly/smart.tar.gz`

[8] `http://members.unine.ch/jacques.savoy/clef/`

- **Weighting:** The raw term (word and phrase) frequency count vectors are first created directly as the output of the indexing step. For our experiments, these are then re-weighted by the LM term weights with Jelineck Mercer smoothing (Hiemstra, 2000, see Equation2.4). All retrieval experiments described in this thesis employ LM as the initial retrieval step with λ set to 0.4. The value of λ was optimized on the TREC-6 training set.

- **Feedback:** The default feedback mechanism in SMART is the Rocchio term re-weighting (Rocchio, 1971). We implemented the LM score-based term selection in SMART (Ponte, 1998). In addition, we also implemented the RLM-based PRF in SMART (Lavrenko and Croft, 2001).

Chapter 4

Sentence based Query Expansion

This chapter seeks to answer the first research question RQ-1 introduced in Chapter 1, which is " *Can additional terms in close proximity to query terms from retrieved documents enrich the information need of the query and improve the retrieval effectiveness of ad-hoc IR?* "

In Chapter 2, we have seen that traditional pseudo-relevance feedback (PRF) in IR typically does not restrict the choice of feedback terms to particular segments in documents. The underyling risk in the standard approaches is that addition of terms which are topically not related to the query terms are likely to introduce a significant query drift, i.e. change the underlying information need expressed in the original query, as a result of which the documents retrieved with the query expanded with such terms are less likely to improve the retrieval effectiveness.

An intuitive approach to address this problem is to ensure that the terms which are topically related to the query terms are selected for query expansion. The next obvious question is then to determine a method to identify such terms. The term selection scores, reviewed in Section 2.1.4, attempt to choose terms based on a combination of measures such as how often does the term occur in the relevant documents, and how rare are they in the collection, with the assumption that relatively rare terms occuring frequently in the relevant documents are good candidates for expansion. A limitation of such methods is that these do not really capture the topical relatedness of an expansion term with a query term.

Methods such as the local context analysis (LCA) and the relevance model (RLM) also take into account how frequently a term co-occurrs with a query term to predict how much related the term is to a query term (cf. Section 2.1.4). The limitation in these cases is that the co-occurrences alone are computed over full documents, which is likely to fail in capturing topical relations between terms specially if the documents are comprised of multiple topics.

We undertake a simple approach of restricting the choice of feedback terms to regions of documents which are maximally similar to the given query. The hypothesis is that terms in retrieved documents that are in close proximity to the query terms, are topically related to the query terms. The unit of proximity chosen for our experimental investigations is the sentence, with the assumption that a sentence characteristically represents natural semantic relationships between its constituent terms (Luhn, 1958).

This chapter is organized as follows. We start with a description of our proposed method and then present the evaluation results of our proposed method. This is followed by a detailed post-hoc analysis. Finally, this chapter ends with a conclusive summary.

4.1 Background and Motivation

Background

Standard PRF methods do not take into consideration the topical structure of the assumed relevant documents. For example, a long document is often comprised of multiple topics all of which may not be relevant to the information need expressed in a query (Wilkinson et al., 1995). Expansion terms are typically extracted from the whole document, as seen by the arrow from the box *retrieved documents* to the rounded box *feedback* of Figure 2.1. This may add a lot of nosiy terms, not associated with the core concepts of the information need to the original query,

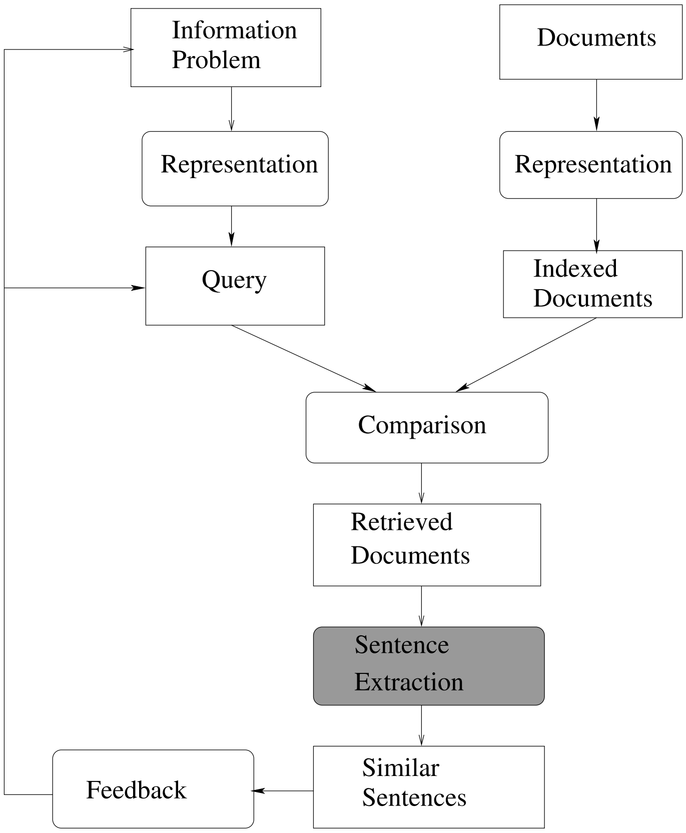


Figure 4.1: Schematic representation of document segmentation.

leading to inefficient query expansion or even a degradation of retrieval effectiveness (Mitra et al., 1998; Billerbeck and Zobel, 2004).

A potential means of addressing this problem is to decompose the retrieved documents into smaller units, and then performing the feedback process using these smaller segments instead of the whole documents. This step is broadly referred to in this thesis as the *document segmentation*. Document segmentation is motivated by the reported result that a feedback document as a whole is seldom relevant (Wilkinson et al., 1995) and that the non-relevant parts of a document can add noise in the feedback step, which in turn can harm the retrieval effectiveness in the feedback step (Terra and Warren, 2005).

Previous research has shown that decomposition of the pseudo-relevant documents into smaller units and a judicious choice of these smaller units can reduce the risk of PRF significantly. For example, the LCA method introduced in Section 2.1.4, uses fixed length word windows to compute the co-occurences so as to reduce the risk of choosing expansion terms from segments of documents unrelated to the query (Xu and Croft, 1996). A more recent work theoretically establishes this principle of local co-occurrences by down-weighting non-proximal co-occurrence with the help of counts of terms propagated by a Gaussian kernel function (Lv and Zhai, 2010). Mitra et al. (1998) used local term correlation weighted *idf* scores summed over fixed length windows to re-rank a subset of top ranked documents, and then assume the re-ranked set as pseudo-relevant.

An attempt to use shorter context for PRF instead of full documents can be found in Lam-Adesina and Jones (2001) where document summaries are extracted based on sentence significance scores, which are a linear combination of scores derived from significant words found by clustering, the overlap of title terms and document, sentence position, and a length normalization factor. Research also provides evidence that summarization improves the accuracy and speed of user relevance judgments (Tombros and Sanderson, 1998).

Motivation

The above arguments suggest that decomposing the retrieved documents into semantically coherent segments can potentially result in improved retrieval. With reference to the standard IR process, as shown in Figure 2.1, this involves inserting an extra processing step, namely that of identifying topically related terms to the query terms, as shown with the gray coloured box in Figure 4.1. To address this limitation, we propose a *sentence based query expansion* technique which restricts the choice of expansion terms to relevant sentences in a document.

4.2 Sentence Based Query Expansion

In this section, we first describe the details of our proposed methodology, and then follow it up with a comparison of our method to other standard PRF methods. Finally, we describe what makes SBQE potentially better than the other PRF methods.

4.2.1 Method Description

The conventional feedback strategy in ad-hoc IR is to assign scores to terms contained in pseudo-relevant documents using a term scoring function, and then to add the top scoring terms to the original query. To reiterate, the limitation of this conventional feedback strategy is that documents as a whole are assumed to be relevant. In this chapter, we propose a simple approach of document segmentation to restrict PRF to parts of documents most similar to the query. The unit of segmentation that we decide to use are sentences within a pseudo-relevant document assuming that these sentences potentially represent short focused syntactically coherent relevant pieces of information to be used for PRF.

Sentence based query expansion (SBQE) alleviates the limitation of partial relevance of feedback documents by segmenting each document into sentences and adding the most similar sentences to the query thus restricting the choice of feedback terms at sub-document level. We add sentences instead of terms extracted from sentences because a sentence can provide semantic context to the expanded query.

The steps of our proposed method is enumerated below.

1. Initialize a sorted set S to NULL. (This is used to store sentences ordered by decreasing similarities).
2. For the i^{th} document $D_i \in R$, where $R = \{D_1, \ldots, D_{|R|}\}$ is the pseudo-relevant set of documents, repeat step 3, incrementing i at each step.
3. For each query sentence in the query do steps 3(a) and 3(b).
   (a) For each sentence in d, compute its cosine similarity with the query and store the sentence-query similarities in S ordered by their decreasing values.
   (b) Add the first $m_i = \min(\lfloor \frac{1-m}{|R|-1}(i-1) + m \rfloor, |S|)$ sentences from the set S to the query.

We now explain the rationale behind each step as follows. In Step 3(a), we use cosine similarity to measure how similar a sentence vector is to the query vector. The reason for choosing cosine similarity as the similarity measure is that it favours shorter texts (Wilkinson et al., 1995). We prefer to choose short sentences similar to the query instead of long ones with the assumption that addition of short relevant sentences to the original query can potentially improve on retrieval effectiveness without introducing too much query drift (see for example Mitra et al., 1998).

Step 3(b) ensures that the number of sentences used for expansion gradually decreases in a linear fashion as we traverse down the list of pseudo-relevant documents, i.e. to say we add the most sentences (m) from the top ranked document and gradually decrease the number of sentences to be added from the subsequent pseudo-relevant documents. Note that the assumption here is that there is some correlation between the rank of a retrieved document and its likely relevance, which may not always be necessarily true. The motivation behind adding different number of sentences comes from the fact that the pseudo-relevant documents are not all equally relevant to the query and hence the PRF contribution from a document is weighted by its rank.

This methodology is also used in RLM feedback introduced in Section 2.1.4. The equation for RLM (Equations 2.10 and 2.11) is reproduced here for the convenience of reading.

$$\begin{aligned} P(w|\mathrm{R}) &\approx \prod_{i=1}^{n} P(w|q_i) \\ &\propto \prod_{i=1}^{n} P(w|q_i) \sum_{j=1}^{R} P(w|D_j)P(q_i|D_j) \\ &\propto \sum_{j=1}^{R} P(w|D_j) \prod_{i=1}^{n} P(q_i|D_j) = \sum_{j=1}^{R} P(w|D_j) \underbrace{P(Q|D_j)} \end{aligned} \tag{4.1}$$

Equation 4.1 shows that the co-occurrence of a term w with a query term q_i is weighted by $P(w|D_j)P(q_i|D_j)$. It is easy to see that $P(q_i|D_j)$ is the probability of generating the query term q_i from the document D_j, which in other words, is the LM similarity of the query term with the document. The relevance model computation in Equation 4.1 is thus directly proportional to the similarity of a document with the query (as seen by the under-bracketed expression in Equation 4.1), implying that the top ranked document plays the most significant role in RLM feedback with progressively decreasing contributions for the subsequent documents in the ranked list. Our method achieves the same objective as well by the use of Step 3(b).

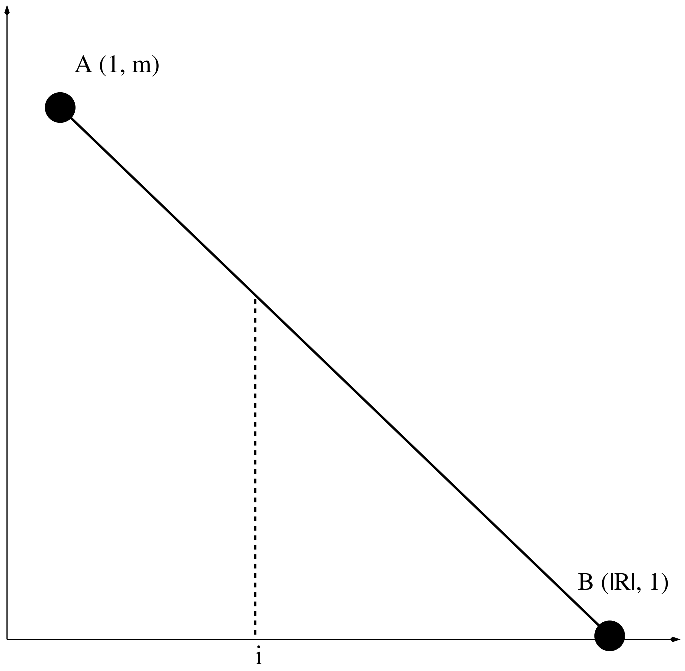


Figure 4.2: Determining how many sentences to add for the i^{th} document.

The parameter of our method is the number of sentences to add for the top ranked document, denoted by m. We determine m_i, the number of sentences to add from the i^{th} pseudo relevant document, as a decreasing linear function with increasing i thus ensuring different importance of feedback documents. Figure 4.2 shows how we compute the value of m_i. Along the x-axis of the figure we plot the number of pseudo-relevant documents while along the y-axis we plot the number of sentences to be added. The point $(1, m)$, labelled A, thus represents the number of sentences to add for the first document, i.e. when $x = 1$, we add m sentences. Similarily, we add only one sentence from the $|R|^{th}$ document as shown by the point B, labelled $(R, 1)$. For any intermediate i, we would want to compute the value of m_i, i.e. to compute the height of the dotted line shown in the figure. The slope of the line AB is given by $\frac{1-m}{|R|-1}$. Using this value for the slope of AB, the equation of

Table 4.1: Differences between the standard term-based and our proposed sentence-based query expansion method for PRF.

Feature	Term based QE	SBQE
QE components	Term-based	Sentence-based
Candidate scoring	Term score/RSV	Sentence similarity
Number of terms	Few terms (5-20)	Many terms ($>$ 100)
Extraction	Terms from feedback documents or segments	Sentences from the whole document
Working Methodology	On the whole set of feedback documents	On one document at a time
Differentiation between feedback documents	Not done	More sentences are selected from a top ranked document as compared to a lower ranked one
idf factor of terms	Used	Not used

the line AB is therefore given by $m_i = \frac{1-m}{|R|-1}(i-1)+m$. Finally, we use the floor function $\lfloor \cdot \rfloor$ to ensure that the number of sentences to add from the i^{th} document is an integer.

4.2.2 Relation to other PRF methods

Our proposed method can be related to the above mentioned existing works in the following ways:

- It utilises the co-occurrence information of LCA and relevance model (RLM) in a different way. The difference is explained as follows. A word may co-occur with a query term in a document, but they may be placed far apart. The proximity between the two cannot be handled by LCA and RLM. Recent work by Lv and Zhai (2010) attempted to address this issue by generalizing the RLM, in a method called PRLM, where non-proximal co-occurrence is down-weighted by using propagated counts of terms using a Gaussian kernel. The difference between our work and LCA and (P)RLM is that co-occurrence of terms is not computed explicitly, since we rely on the intrinsic relationship of a document word with a query term as defined by the proximity of natural sentence boundaries.

- Our method utilizes shorter context as explored in Lam-Adesina and Jones (2001) and Järvelin (2009), but differs from these approaches in the sense that these methods follow the traditional term selection approach over the set of extracted shorter segments whereas we do not need to employ any additional term selection method from the shorter segments (sentences). The reason we do not employ an additional term selection step is that term selection does not take into account the term proximity, which we hypothesize is important for identifying terms topically related to the information need.
- In our method we also do not need to tune additional parameters such as the window size for passages as in Allan (1995), which makes optimization easier.
- Existing work on sentence retrieval consider sentences as the retrieval units instead of documents (Murdock, 2006; Losada, 2010). The difference between this and ours is that our goal is not to retrieve sentences, but rather use sentence selection as an intermediate step to help PRF.

Table 4.1 summarizes the major differences between term-based QE and SBQE.

4.2.3 Justification of SBQE

It is of utmost importance to ensure that the terms added to the initial query for expansion are topically related to the concept of the original query terms. The key hypothesis behind SBQE is that term proximity can play a part in identifying terms which are topically related to the query terms, with the assumption that such terms occur in close proximity to the query terms. The proximity unit chosen for our investigation is the sentence, because we believe that sentences define the natural semantic boundaries between the terms of a document. Adding sentences can hence capture the useful context information, which is often missed if only isolated terms are added to the query, such as in the standard PRF approaches (Ponte, 1998; Lavrenko and Croft, 2001).

Our method is different from the other PRF extensions which attempts to restrict the choice of terms to selected regions of documents, such as (Lam-Adesina and Jones, 2001; Järvelin, 2009). These methods however, do eventually employ a term selection at the level of sub-documents rather than whole documents, thus not addressing the term proximity. We not only restrict the choice of terms to the sentences of documents most similar to the query but also ensure that we preserve the importance of term proximity by adding the full sentences.

It is not only the presence of a query term that SBQE feedback utilizes, but it also tries to reproduce the distribution of the query terms through evidences in the top R document texts as accurately as possible, explained as follows. For a query term (say q_i), greater the number of sentences in which q_i occurs in, greater is the number of times these sentences will be selected for addition to the query, which in turn implies that greater is the number of times this particular term will be added to the expanded query. In contrast to this, for another query term (say q_j), which does not occur frequently in the top ranked documents, fewer sentences containing this term will be selected for addition, which in turn implies that q_j will have a low frequency count in the expanded query. The term frequency of q_i in the expanded query will thus be higher than that of q_j, thus assigning more weight to q_i than q_j in the second retrieval step.

This difference in the relative frequencies of terms plays a very important in the LM retrieval model, because the relative importance of a more frequent term q_i is increased, which means that a match in term q_i is more important than a match of term q_j which occurs less frequently in the query (see Equation 2.4).

4.3 Evaluation

To evaluate the effectiveness of SBQE we used the TREC ad-hoc document collection and title fields of the TREC 6, 7, 8 and TREC Robust query sets (cf. Section 3.2.1). We used TREC-6 topics for training the parameters R (the number of top ranked pseudo-relevant documents) and T (the number of expansion terms). TREC 7,8 and Robust topic sets were used as the test sets.

4.3.1 System Description

In common with all the experiments described in this thesis, we used the LM implementation of SMART for indexing and retrieval (cf. Section 3.3). SBQE involves choosing expansion terms from sentences of documents most similar to the query. Sentence boundary detection is a part of the original SMART implementation. Our feedback module thus uses this implementation to constitute a list of sentences for every pseudo-relevant document, compute the similarity of each with the query and choose the top most ones for query expansion.

To compare SBQE with the existing feedback approaches in LM, we selected two baselines. The first baseline used was the LM term based query expansion, hereafter referred to as LM (Ponte, 1998). Ponte advocates adding the top LM scored (see Equation 2.13 of Section 2.1.4) T terms from R top ranked documents to the original query (Ponte, 1998). The LM term score prefers terms which are frequent in the set of pseudo-relevant documents (R) and infrequent in the whole collection.

The second baseline used was the RLM (Relevance Model). RLM involves estimating a relevance model and reordering the initially retrieved documents by KL-divergence from the estimated relevance model (Lavrenko and Croft, 2001). Query expansion with additional query terms and a subsequent retrieval with the expanded query was also performed on the reranked RLM results.

4.3.2 Parameter Sensitivity

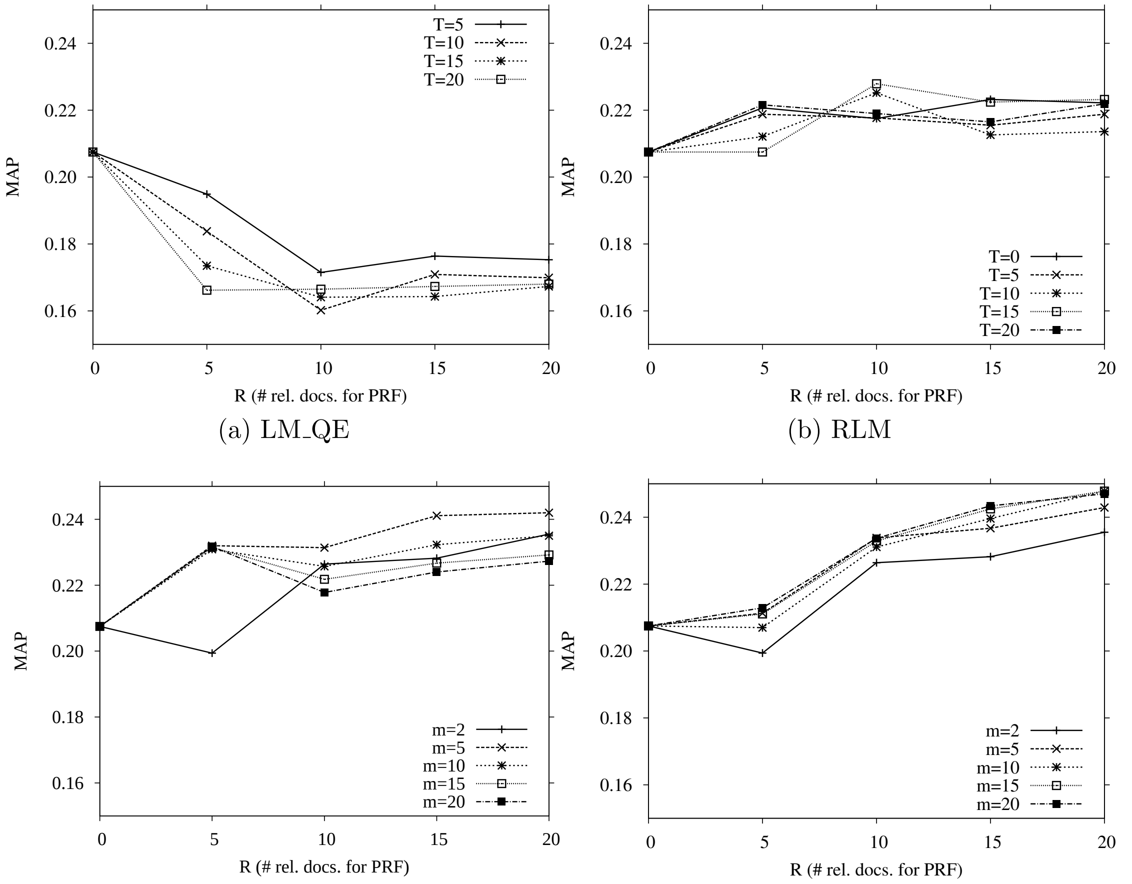
One of the parameters to vary for both LM and SBQE is the number of documents to be used for the PRF which we refer to as R. We vary both R and T (the number of terms to add for LM) in the range of [5, 50] in steps of 5. The other parameter to vary for SBQE is m, which is the number of sentences to add to the query from the top ranked document. To see whether adding a variable number of sentences is beneficial, we also experimented with a version of SBQE, where we fix the number of sentences to be added from each pseudo-relevant document to the constant m, instead of decreasing the value of m linearly as proposed in the SBQE algorithm 4.2. The different approaches for PRF that we experimented with are summarized below:

- LM_QE: LM score based query expansion (Ponte, 1998).
- RLM: Relevance Modeling feedback without QE ($T = 0$) or with query expansion ($T > 0$) (Lavrenko and Croft, 2001).
- SBQE$_{\text{cns}}$: Sentence based QE with constant number of sentences, i.e. where we add an equal number of sentences from each pseudo-relevant document. The objective is to see whether PRF contribution from a document proportional to its rank helps improve the feedback quality.
- SBQE: Sentence based QE with a progressively decreasing number of sentences.

Figure 4.3 shows the effect of varying the parameters for the different PRF approaches. Fugure 4.3a shows that LM_QE performs very poorly. It can be seen that for all combinations of (R, T), the MAP decreases as compared to the initial baseline. The best MAP we obtain with LM_QE, namely 0.1949 (using 5 documents and 5 terms), is in fact worse than the initial retrieval MAP 0.2075 as seen in Figures 4.3a.

Figure 4.3b shows that the RLM performs best when additional 10 terms are added on top of the RLM reranked results using R=15 documents. The runs labelled as RLM in the subsequent experiments of this chapter uses the same settings of RLM.

SBQE$_{\text{cns}}$ performs slightly worse than its variable sentenece counterpart SBQE (see Figures 4.3c and 4.3d). SBQE is also more robust than SBQE$_{\text{cns}}$ with respect to parameters, as can be seen by the less number of intersecting iso-m lines. Both



(a) LM_QE

(b) RLM

(c) SBQE$_{\text{cns}}$

(d) SBQE

Figure 4.3: Parameter sensitivity of LM_QE, RLM, SBQE$_{\text{cns}}$ and SBQE on TREC-6 topics used as the training set.

the LM and the RLM graphs are more parameter sensitive than SBQE, as can be seen from the larger average distances between iso-T points and a higher number of intersections of the iso-R lines. Furthermore, in case of SBQE, there is no noticeable degradation in MAP with an increase in m, as seen by a more or less steady increase in the MAP values in Figure 4.3d.

4.3.3 SBQE Results

In this section, we compare the different PRF approaches on the test set, i.e. the TREC 6-8 and the Robust topics. Since SBQE with a variable number of sentences outperforms its counterpart which uses a constant number of sentences, the subsequent SBQE experiments on the test data set are conducted using this version

Table 4.2: Comparative evaluation of LM_QE, RLM and SBQE on TREC topics (TREC-6 topics were used for parameter training).

TREC	Topics	MAP			
		LM	LM_QE	RLM	SBQE
TREC-6	301-350	0.2075	0.2061 (-0.67%)	0.2279 (9.83%)	**0.2481**$^{+*}$ (19.56%)
TREC-7	351-400	0.1614	0.1673 (3.65%)	0.1714 (6.19%)	**0.1963**$^{+*}$ (21.62%)
TREC-8	401-450	0.2409	0.2302 (-4.44%)	0.2612 (8.42%)	**0.2891**$^{+*}$ (20.01%)
TREC-Robust	601-700	0.2618	0.2796 (6.79%)	0.3236 (23.60%)	**0.3540**$^{+*}$ (35.21%)

only.

In Table 4.2 we report the MAPs obtained via all three approaches for the 400 TREC topics, repeating the results for the 50 TREC-6 topics used as the training set. Alongside the MAP values the table also reports the percentage changes in MAPs computed with reference to the initial retrieval MAPs for the corresponding approach (which is not shown in the table for brevity).

It can be observed that SBQE outperforms both LM and RLM on these test topics. The statistically significant [1] improvements in MAP with SBQE over LM and RLM are shown with a $^+$ and * respectively.

The most interesting observation is the 35.6% improvement in MAP for the Robust track topics, which are topics known to be difficult to improve with query expansion (Voorhees, 2004). The best performing TREC Robust track runs in 2004 used external resources to improve retrieval effectiveness (Kwok et al., 2004; Amati et al., 2004). Our method produces results close to these without relying on the availability of external resources.

4.4 Analysis of SBQE

We hypothesize that the queries for which the initial retrieval average precision (AP) is low, have in fact the highest scope of improvement in the AP value. It is particularly interesting to see the effect of PRF on these queries. To this end,

[1] Throughout the rest of this thesis, *significance* would refer to statistical significance measure by Wilcoxon test with 95% confidence measure.

we categorized the queries by their initial retrieval AP values. Categorizing queries this way has in an approximate sense an effect of grouping the queries by their difficulty levels with the hypothesis that the most difficult queries are arguably those ones for which the initial retrieval AP is very low, i.e. within the range $[0, 0.1]$. The intention of the per-group analysis is to see how the performance of the PRF methods compare for each group, specially the group representing queries which have the lowest initial retrieval AP and thus the highest scope of improvement. The per-group analysis is followed by a term frequency analysis of the expanded query, where we show that leaving out less frequent terms from the expanded query degrades SBQE performance, which in turn shows that adding whole sentences is vital for SBQE. We then provide a run-time (i.e. the total execution time over a set of queries) comparison of SBQE with the baselines. Finally, this section concludes with a comparison with true relevance feedback, where we show that SBQE can in fact add a higher number of relevant terms to the expanded query than the other baseline approaches.

4.4.1 Query Drift Analysis

PRF is associated with the implicit risk of degrading retrieval effectiveness for many queries, since not all top ranked pseudo-relevant documents are relevant to the query. Hence, terms selected from these non-relevant documents when added to the query may cause the expanded query to drift away from the intended focus of the query causing it to favour retrieval of non-relevant documents (Billerbeck and Zobel, 2004). A feedback method which benefits the queries with reasonably low initial AP is particularly desirable since these queries have a huge scope of improvement as the initial retrieval resultset for these queries mostly comprise non-relevant documents. On the other hand these queries are particularly susceptible to query drift firstly due to the presence of non-relevant content in the initial retrieval resultset and secondly due to the lack of topical focus and coherence between the top ranked documents, as a result of which the terms selected for query expansion are likely to lack focus as

Table 4.3: Effect of initial retrieval average precision on LM_QE, RLM and SBQE.

Initial retrieval (LM)	% Queries improved			% change in AP		
precision interval	LM_QE	RLM	SBQE	LM_QE	RLM	SBQE
[0 – 0.1)	45.8	51.7	**53.9**	+48.6	+42.5	**+75.0**
[0.1 – 0.2)	51.9	57.6	**74.0**	+18.1	+34.8	**+64.0**
[0.2 – 0.3)	58.1	67.7	**83.8**	+1.7	+20.1	**+37.1**
[0.3 – 0.4)	39.2	64.2	**82.6**	−4.7	+12.1	**+27.5**
[0.4 – 0.5)	45.4	50.0	**83.3**	−9.4	−1.1	**+23.5**
[0.5 – 1]	38.7	**67.7**	64.3	−8.6	**+3.4**	+1.3

well. We hypothesize that SBQE can prove beneficial for these queries since it relies on identifying terms topically related to the query terms based on the proximity evidence rather than relying on the term scores.

To see how initial retrieval precision can affect SBQE, we categorized the topics of TREC 6-8 into classes defined by a range over the initial retrieval APs. Five equal length intervals were chosen as $\{[i, i + 0.1)\}$ where $i \in \{0, 0.1 \dots 0.4\}$. Since there are not many topics with initial retrieval AP over 0.5, the last interval is chosen as $[0.5, 1]$ so as to maintain a balance in the number of queries in each bin for meaningful comparisons. Thus the first bin contains the topics for which the initial retrieval AP is between 0 and 0.1, the second bin consists of topics for which the it is between 0.1 and 0.2 and so on. For each bin, the AP is computed by considering only the queries of that current bin.

In Table 4.3, we report statistics computed for each query class for the three expansion techniques. We observe that SBQE results in the highest percentage of query improvement for every class except the last one. This also suggests that RLM has a tendency to improve retrieval effectiveness only for queries with high initial retrieval average precision. Moreover, SBQE results in the highest percentage gains in AP values for each class. Particulary interesting is the 75% increase in AP for the first query class which suggests that SBQE is able to increase the retrieval peformance of those queries which are most in need of improvement.

The improvement achieved by SBQE for the queries with low initial retrieval AP conforms to our hypothesis that it is very important to choose topically related expansion terms for these queries. Proximity of a term with a query term turns out to be useful in predicting the topical relatedness of that term with the information need concept expressed in the query.

4.4.2 Feedback effect on TREC Robust topics

The TREC Robust track explored retrieval for a challenging set of topics from the TREC ad hoc tasks (Voorhees, 2004).

As pointed out in Section 3.2.2, a subset of 28topics from TREC Robust topics were categorized as *hard* based on Buckley's failure analysis (Harman and Buckley, 2004). It has been shown that the average precision of these queries in general is difficult to improve by application of PRF. It is particularly interesting to see the performance gains achieved by SBQE on these difficult queries. Consequently, we report the performance of each PRF method on the *hard* topics (categories 2-9) (cf. Table 3.4).

Our results for individual groups of topics are shown in Table 4.4.

From the results of Table 4.4 we can see that SBQE outperforms LM_QE and RLM for most topic category type. The categories improved are primarily the ones where failure occurs due to missing one or more aspects of the query such as difficulty categories 4, 5, and 7 (see Table 3.4 or Table 4.4). In addition, SBQE also works particularly well for category 2 queries, for which IR systems are prone to general technical failures, such as stemmming. We further observe that SBQE fails only for query 6, as against failure of LM_QE for 5 topic categories (2, 5, 6, 7, 8) and failure of RLM for 4 (2, 6, 7, 9), proving that SBQE is a more robust PRF method than the other two.

To see that SBQE is able to add more important and relevant terms for query expansion in comparison to LM_QE and RLM, we take a sample query from each category and report some terms added by SBQE, but not by LM_QE and RLM. For topic 445 - "women clergy" belonging to category 3, true feedback adds terms like *stipend*, *church*, *priest*, *ordain*, *bishop*, *England* etc. The description of the

Table 4.4: Revisiting Buckley's failure analysis for LM_QE, RLM and SBQE.

Topic	MAP			
Category	LM	LM_QE	RLM	SBQE
2: General technical failures such as stemming.	0.2111	0.1275 (-39.6%)	0.0877 (-58.4%)	**0.2685(+27.1%)**
3: Systems all emphasize one aspect, miss another required term.	0.0835	0.1518 (+81.8%)	**0.1891(+126.3%)**	0.1693 (+102.6%)
4: Systems all emphasize one aspect, miss another aspect.	0.0939	0.1360 (+44.9%)	0.1508 (+60.6%)	**0.1518(+61.6%)**
5: Some systems emphasize one aspect, some another, need both.	0.2330	0.2323 (-0.3%)	0.2618 (+12.3%)	**0.2840(+22.0%)**
6: Systems all emphasize some irrelevant aspect, missing point of topic.	0.0617	0.0146 (-76.33%)	**0.0372(-29.3%)**	0.0184 (-70.1%)
7: Need outside expansion of "general" term.	0.0527	0.0339 (-35.65%)	0.0372 (-29.32%)	**0.0553(+4.9%)**
8: Need query analysis to determine relationship between query terms.	0.2295	0.1654 (-27.93%)	**0.2881(+25.53%)**	0.2622 (+14.2%)
9: Systems missed difficult aspect.	0.0481	**0.0618(+28.41%)**	0.0421 (-12.42%)	0.0547 (+13.6%)

topic reads "What other countries besides the United States are considering or have approved women as clergy persons". While LM_QE and RLM add the terms *church*, *priest* and *ordain*, SBQE in addition to these ones adds terms such as (*bishop*, 7), (*England*, 10), (*stipend*, 7), (*ordain*, 11) where the numbers beside the terms indicate their occurrence frequencies in the expanded query. As per the description of this topic, *England* is indeed a sensible term to add. A look at topic 435 - "curbing population growth" belonging to category 4, reveals that term based LM feedback adds terms like *billion*, *statistics*, *number*, while it misses terms representing the other aspect of relevance (the aspect of contraceptive awareness in rural areas to prevent population growth - emphasized by terms like *rural*, *contraceptive* etc.), which are added by SBQE.

We thus conclude that SBQE is able to add a higher number of thematically related terms, which are likely to be relevant, to the initial query. The fact that

SBQE outperforms other approaches for the failure class 4 and 5 shows that SBQE in fact is able to extract related terms from each relevant topic from multi-topical documents.

4.4.3 Term frequency analysis of expanded query

One argument against SBQE is that it often adds a large number of terms to the query. Although the results show a significant increase in MAP as compared to the baseline methods, it can be argued that a more careful additiion of a smaller number of expansion terms may in fact increase the retrieval effectiveness further.

However, recall that SBQE does not attempt to reduce the number of terms since a part of its motivation is the hypothesis that whole sentences provide valuable semantic context to the expanded query. To provide empirical justification to this hypothesis, we report an experiment where we intentionally reduce the length of an SBQE expanded query in order to see the effect on retrieval effectiveness. The experiments are conducted on the TREC-8 dataset.

In these experiments, firstly we set the frequency of each term to 1, thus reducing the expanded query to a uniform distribution where every term is equally likely to occur. The objective is to test whether the term frequency values of the expanded query play a part in the working principle of SBQE (see Section 4.2.3 for a discussion on this).

Next, we seek an answer to the question of whether all terms that we added to the query are indeed useful for retrieval or could we filter out some of the rarely occurring terms from the expanded query. We therefore remove terms falling below a frequency cut-off threshold of 10, 2 and 1, i.e. to say, we remove those terms which have a frequency less than or equal to 10, 2 and 1 respectively from the expanded query.

Table 4.5 reports the observations and clearly shows that the frequencies indeed play a vital role because retrieval effectiveness decreases either when we set the term frequencies to one ignoring the evidence we collected from each feedback document,

Table 4.5: Term frequency variations on the expanded TREC-8 topics

Terms	MAP
All terms	0.289
$tf(t_i) \leftarrow 1$ (Frequencies set to 1)	0.181
Terms with frequency > 1	0.280
Terms with frequency > 2	0.273
Terms with frequency > 10	0.248

or when we leave out some of the terms. Since we add a large number of terms to the original query, the expanded query at a first glance might intuitively suggest a potential for query drift. However, the observation which needs to be made here is that a vast majority of the terms are of low frequency. *Important* terms are those which have maximal evidence of occurrence in the feedback documents in proximity to the original query terms (the notion of proximity in SBQE refers to the natural sentence boundaries). However, frequency alone is not the only criterion for the *goodness* of a term. The low frequency terms are also beneficial for the feedback step as suggested by the fact that simply cutting off the terms based on frequency has a negative effect on retrieval quality.

4.4.4 Run-time Comparisons

In this section, we compare the average run-times for the three PRF approaches, namely LM_QE, RLM and SBQE. Table 4.6 shows the average number of terms and the run-times for the three PRF methods. It may occur that using more than 400 terms for retrieval can lead to poor retrieval performance in terms of runtime efficiency. However, a careful observation reveals that the run-time complexity of RLM is $O(VR)$, where V is the number of unique terms in R pseudo-relevant documents (see the description of RLM in Section 2.1.4). Thus, RLM can be viewed as a massive query expansion technique where the original query is replaced by a probability distribution over the vocabulary of V terms.

SBQE, on the other hand, involves iterating over R feedback documents, splitting these up into sentences and sorting them on the basis of similarities with the query.

Let the average number of sentences in a feedback document be S. Thus the run-time complexity for feedback is $O(RS \log(RS))$. Clearly, the average number of sentences in a feedback document is much less than the number of terms in the vocabulary of R documents. This explains why SBQE is computationally faster than RLM.

Table 4.6: Run-time comparisons of LM term-based expansion, RLM and SBQE

Method	Avg. # terms	Time (s)
LM	7.38	7
RLM	2.38	209
SBQE	465.88	91

4.4.5 Comparison with True Relevance Feedback

To see if SBQE is indeed able to add the *important* query terms to the original query we run true relevance feedback (TRF) experiments which involve selecting terms by the LM term selection values as was done in our standard PRF experiments, the only difference being that we use only the true relevant out of the top R documents of the initial ranked list for feedback. While we do not expect that SBQE can outperform TRF, this experiment was designed with the purpose of testing how close the performance of SBQE can get to the ideal scenario. Our main aim was to define a gold-standard for the feedback terms by restricting the LM term selection value to the set of true relevant documents with the assumption that the terms hence selected for feedback provide an evidence of *good* feedback terms. An overlap between the terms obtained by SBQE and the *good* terms found this way from TRF can be a measure of the effectiveness of SBQE.

We do the TRF experiments for TREC 6-8 topic sets. Since the maximum value of R, the number of pseudo-relevant documents used for PRF experiments reported in Section 4.3.3 is 20, we use the same set of documents for the TRF. The difference is that in TRF, we filter out the non-relevant documents from this set, retaining only the relevant ones. The cardinality of intersection of the set of terms obtained

Table 4.7: Intersection of PRF terms with the gold-standard TRF terms

Topic set	TRF		LM		SBQE	
	MAP	$\|T_{TRF}\|$	MAP	$\|T_{TRF} \cap T_{LM}\|$	MAP	$\|T_{TRF} \cap T_{SBQE}\|$
TREC-6	0.409	1353	0.195	316 (23.3%)	0.248	**901 (66.6%)**
TREC-7	0.422	1244	0.163	311 (25.0%)	0.196	**933 (75.0%)**
TREC-8	0.376	1234	0.213	317 (25.7%)	0.289	**977 (79.1%)**

by a PRF approach with the set of terms obtained by the TRF method indicates the effectiveness of the former. Note that our intention here is not to compare the retrieval effectiveness directly; rather we seek to explore how close in performance can a feedback method get to the TRF as far as selection of expansion terms is concerned.

In Table 4.7 we report the intersection of the set of terms obtained by LM and SBQE with TRF terms. We denote by T_X the set of terms for a particular set of topics obtained by method X (X is either TRF, LM or $SBQE$). We also re-report the MAPs from Table 4.2 for convenience of reading. We observe from Table 4.7 that SBQE is able to add more *important* terms due to the higher degree of overlap with TRF terms.

4.4.6 True Relevance Feedback with SBQE

In the previous section, we have shown that SBQE is able to achieve a higher overlap of expansion terms with those obtained with term based true relevance feedback in comparison to the baseline approaches. The next interesting question is how does SBQE perform when the set of documents used for the feedback is comprised of *true* relevant documents. We expect that since SBQE outperforms the baseline approaches with pseudo-relevant documents, it also is likely to outperform the baselines with true relevant documents. The results are reported in Table 4.8. It can be seen that TRF with SBQE produces significantly better results than the LM term based expansion (except for the TREC-7 topic set where the results are statistically indistinguishable), which is suggestive of the fact that SBQE is a more

Table 4.8: True relevance feedback with SBQE.

Topic set	LM	SBQE
TREC-6	0.4093	**0.4924***
TREC-7	**0.4224**	0.4202
TREC-8	0.3762	**0.4735***

robust feedback technique than its term based counterpart.

4.5 Summary and Conclusions

The results of SBQE, i.e. our proposed approach of segmenting the feedback documents into sentences and using the sentences as units for query expansion, show that SBQE is effective and has the following desirable qualities:

- It can be successfully applied to significantly improve retrieval effectiveness as compared to standard term based query expansion methods and relevance language model based feedback.
- The improvement margin is the most for queries having low initial retrieval average precision, which are the queries badly in need of the improvement.
- It works very well on the TREC Robust track topic set, which was specifically designed to test the robustness of PRF in IR, even without the use of any external resources.
- It is able to find a higher number of *useful* terms for query expansion.

The hypothesis behind the working principle of SBQE is that the a sentence characteristically represent natural semantic relationships between its constituent terms. Exploiting these semantic relationships between the terms through proximity helps in choosing expansion terms which are topically related to the concept of the initial information need, the addition of which in turn helps to enrich the topics or aspects of the initial information need. Our method however, does not explicitly attempt to model the latent topics manifested in the top ranked retrieved documents due to the effect of the multiple fine-grained aspects of a query. Term proximity alone may not be sufficient to determine topically related terms, and approaches of statistically modelling topics such as those reviewed in Section 2.2.1, may be more effective in further improving the PRF performance. With this objective, we explore research question RQ-3 attempting to integrate a topic modelling approach within the standard framework of PRF. This work of ours is described in Chapter 6.

However, before moving onto Chapter 6, we first explore the effectiveness of PRF for very long queries such as the patent prior art search queries. The multi-topical nature of these queries can potentially create problems for retrieval since the information need is not focused. The next chapter explores methods to address this problem.

Chapter 5

Query Segmentation

In the previous chapter, we explored use of document segments at the level of sentences with the aim of improving pseudo-relevance feedback (PRF), hypothesizing that such smaller units of text such as the sentences can capture the natural semantic relationships between their constituent terms. In this chapter, we explore the complementary approach of query segmentation for improving PRF. Recall that the second research question RQ-2, introduced in Chapter 1, involves exploring segmentation of very long queries (such as those encountered in patent prior art search), for improving retrieval effectiveness. Typically, very long queries do not focus on one particular topic, but rather are associated with multiple topics, each of which may relate to its own set of relevant contents. The intention of an IR system for such long queries is to retrieve documents relevant to each topic of interest in the query. In the case of patent queries in prior art search this means finding documents related to each topic or claim expressed in the patent query. Typically, these very long queries in patent search are not focused to a single information need because these were not written keeping any information need in mind. Due to the lack of focus in the such long queries, traditional retrieval methods may not be effective and PRF methods have been shown to actually degrade average retrieval effectiveness (Magdy and Jones, 2010a).

This chapter proposes a method to utilize topical segmentation of queries by decomposing a query into sub-topics, which are focused on separate individual facets of the information need. Our research objective in this chapter is to explore on whether transforming these long queries into segments in this way can in fact make the applications of standard IR methods more effective. The remainder of the chapter is organized as follows. First in Section 5.1, we motivate the objective of the study by highlighting the limitations of standard IR methods for long queries. We then describe the details of our proposed method in Section 5.2. Next, we provide details of the experimental setup for evaluationg our proposed method in Section 5.3 followed by presentation of the results in Section 5.4. This is followed by presentation of a detailed analysis for our proposed method in Section 5.5. Finally, Section 5.6 concludes the chapter summarizing the observations.

5.1 Motivation

In contrast to short queries comprising a few words such as in ad-hoc IR where in information need can be multi-faceted, such multiple aspects of the information need can often be explicitly expressed, such as in the claims of patent prior art search queries. Each *claim* field of a patent query typically expresses an individual information need for the prior art related to that claim. An excerpt from a patent document (a query in the CLEF-IP 2010 is structurally and characteristically identical to a document) is shown in Fig 3.3.

The standard IR architecture, shown in Figure 2.1, does not have provisions for addressing each aspect of complete statement of multiple related facets of an information need. In particular, for very long patent search queries, where the aim is to retrieve prior art for each claim, a standard IR system is unlikely to work well in practice for the following reasons.

- The expository content of the query patent can in some sense *confuse* a retrieval model in the *mathcing* (retreival) phase, i.e. in the computation of the similarity of a document to the query. Consider for example the part of a

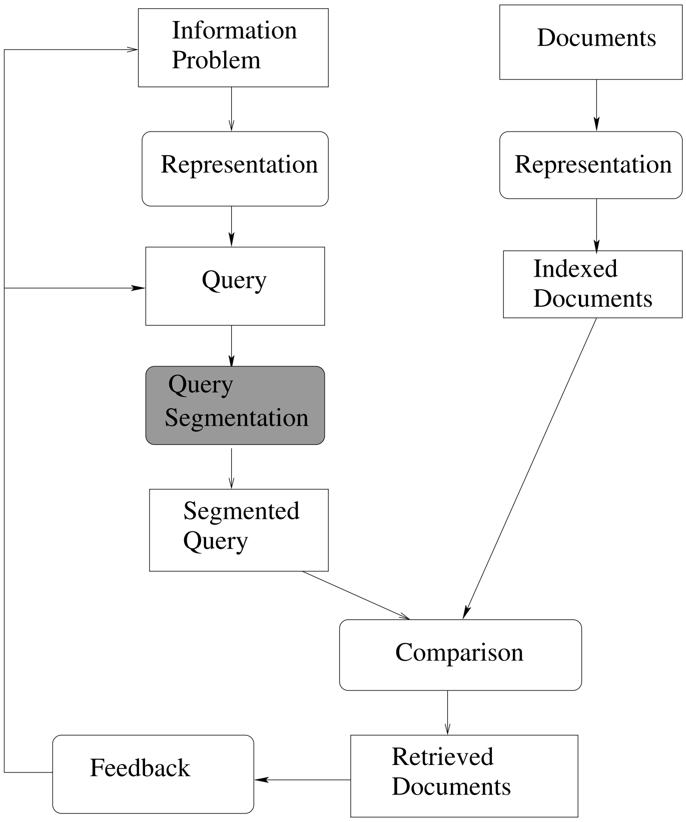


Figure 5.1: Extending standard IR with pre-retrieval query segmentation.

CLEF-IP 2010 patent query, say Q, as shown in Figure 5.2. The query shows two consecutive paragraphs, the first on methods of forming pits through corrosion and the second on growing Group-III nitride on these pits. Existing prior art related to both these methods should be relevant for a patent examiner. Now let us assume that there exist documents D_1 and D_2, which respectively are prior arts to these two methods. The similarity values of the whole query with these documents, i.e. $sim(Q, D_1)$ and $sim(Q, D_2)$ respectively, can be low because of the presence of a large number of non-overlapping terms, e.g. second paragraph of Q not matching with D_1 and the first paragraph of Q not matching with D_2, and other paragraphs matching with a different set of documents, as a result of which these documents can be retrieved at better ranks than D_1 or D_2. Contrast this to the case when we split the query into two segments Q_1 and Q_2, each focusing on one particular topic, which are respectively the pit formation and semiconductor growing, in this particular example of ours. Since Q_1 and Q_2 comprise of more focused information needs, the retrieved ranks of D_1 for the query Q_1 and that of D_2 for the query Q_2 are better than the ranks at which D_1 and D_2 are retrieved for query Q. Clearly, splitting up the queries in this way thus potentially ensures a more meaningful matching between a document in the collection with a particular aspect of the query.

- The PRF process itself can suffer because of the presence of documents on different topics within the top ranked set. Returning back to our example, adding terms related to neither pit formation nor nitride growing may contribute to an increase in the *confusion* in the matching phase for the query Q during the subsequent retrieval phase after feedback. This is because adding these expansion terms further defocuses Q in terms of both the topics, namely pit formation and nitride growing. Similar to the earlier reasoning, it can be argued that this may contribute in further degrading the ranks of the relevant documents D_1 and D_2, since the additional terms can retrieve more documents related to neither of these topics at top ranks, thus pushing D_1 and D_2 further down the ranked list.

> . . . Therefore, when the portion of the uppermost layer containing lattice defects is subjected to treatment by use of a solution or vapor that can corrode the portion more easily than it can corrode the portion of the uppermost layer containing no lattice defects, pits having the shape of an inverted hexagonal cone and having center axes coinciding with threading dislocation lines are formed. The vertexes of the pits correspond to the end points of the threading dislocations.

> After the pits are formed as described above, when a second Group III nitride compound semiconductor layer is grown through vertical and lateral epitaxial overgrowth around nuclei as seeds for crystal growth which are . . .

Figure 5.2: Output of TextTiling for the description of a CLEF-IP 2010 query.

To address these limitations, highlighted above, the standard IR architecture can be extended by inserting an extra processing step, namely *query segmentation* which works as follows. The query representation process can be extended to include an additional layer of processing to obtain a set of segmented queries. The comparison process then computes the similarities between each query segment and the indexed documents. This method is shown in Figure 5.1. In Section 5.2, we investigate the details of the query segmentation approach schematically described in Figure 5.1.

5.2 Retrieval with Query Segments

In this section, we explore research question RQ-2 by investigating segmentation of multi-topic queries.

Every sub-topic in a patent query expresses a particular aspect of the claimed invention. The prior art search task requires existing patent documents to be retrieved for each such aspect. Using a full patent claim as a query is therefore associated with a risk of not focusing on one particular aspect of the information need thus leading to an ineffective document-query matching, as illustrated in the example given in Section 5.1. The example also illustrated that expanding the query further contributes to a degradation of the specificity of the information need, hurting retrieval effectiveness further.

To alleviate these issues, we propose to use each of the sub-topics or segments of a whole patent as separate queries to produce individual sub-queries to be given as inputs to the retrieval system, and then to merge the retrieval results from each of the individual sub-queries to construct a final ranked list for the whole original query. We hypothesize that using each sub-topic as a separate sub-query should enable a retrieval system to identify relevant documents from the collection in a more effective way, and also that it will allow the PRF algorithm to work effectively since it can be applied to a more focused set of pseudo-relevant documents, than in the case when using a single multi-faceted patent document as the query.

5.2.1 Method Description

The details of our proposed method of retrieval using query segments as follows. The rationale behind each step is then explained following the description of the algorithm.

1. Segment each patent query Q into the constituent fields: title (Q_t), abstract (Q_a), description (Q_d), and claim (Q_c).
2. Segment each Q_d into $\eta(Q_d)$ segments Q_d^1, $Q_d^2 \ldots Q_d^{\eta(Q_d)}$.
3. Remove the unit frequency terms from each query segment.
4. Run retrieval on each query segment: Q_t, Q_a, Q_c and $\{Q_d^i\}$, $(i = 1 \ldots \eta(Q_d))$. Let the list of documents retrieved for a segment q be $L(q)$.
5. Interleave one document from each $L(q)$, eliminating duplicates while interleaving, in a round-robin manner to construct the initial retrieval ranked list for the whole query Q.
6. Expand query using R pseudo-relevant documents from each initial ranked list $L(q)$ by adding T terms to each query segment q to obtain the expanded query segment q'.
7. Perform retrieval to obtain feedback ranked lists $L(q')$ on each expanded query segment q', and build up the feedback retrieval result for the original query Q in the exact same manner as Step 5.

In step 2, we segment the patent query into coherent topical units. The intention of text segmentation is to decompose a text into blocks, where the content of each block is focused on a particular topic. The segmentation method that we use in our experiments is TextTiling (Hearst, 1997). TextTiling is an automatic text segmentation process which involves splitting up a document into coherent sub-topics, by selecting the valleys in the smoothed plot of cosine similarities between adjacent blocks of sentences as potential topic shift points (Hearst, 1997). A sample TextTiling output is shown in Figure 5.2 illustrating two consecutive segments obtained from the description text of a CLEF-IP 2010 query. While the first segment talks about forming pits by corrosion, the second elaborates on growing a layer of Group-III nitride on the pits thus formed. So it makes more sense to use these two segments as separate queries and retrieve relevant contents from each of these and then merge the separate result lists.

The unit frequency terms are removed in step 3 in accordance to the observation that doing so improves retrieval effectiveness (Magdy and Jones, 2010a) (see also Section 3.2 for more details on the additional query pre-processing step conducted for the CLEF-IP queries).

In step 5, we use interleaving as opposed to the more standard fusion techniques such as COMBSUM etc. (Fox and Shaw, 1994). This is because COMBSUM is particularly useful for merging results retrieved by different retrieval algorithms executed against the same query, but in our case it is the queries which are different and not the retrieval algorithm. More precisely speaking, every query segment is a sub-topic or one specific aspect of the whole information need and we expect that the relevant set should comprise of documents from each of these query segment. This is what we do by the one-way interleaving or choosing documents in a round-robin manner from the ranked lists retrieved against each query segment. Thus, in the merged result set we end up with documents from each sub-topic.

The intuitive reason why COMBSUM is not a good fusion candidate is explained as follows. In COMBSUM, the document similarity score assigned to a document D from among a list of $\{L_1, \cdots L_N\}$ retrieval results is as follows (Fox and Shaw, 1994).

$$COMBSUM(D, Q) = \sum_{i=1}^{N} score(D, L_i) \tag{5.1}$$

COMBSUM is thus expected to work well when the individual lists L_is are ranked retrieval resultlists obtained by different approaches using the same query Q, since in this case COMBSUM looks to collect multiple evidences of the relevance of a document D with different retrieval approaches. A document D which has high similarity values in all the lists is thus given a high score in the fused resultlist. In our case however, the lists L_is are resultlists obtained by retrieving documents with different query segments. A document D is thus usually expected to be present in one particular L_i while the values of $score(D, L_j)$ for the other lists L_j ($j \neq i$) are expected to be 0. In such a case, COMBSUM score assignments may be biased towards one or a few resultlists, which could particularly be the lists L_is with relatively higher values of the scores $score(D, L_i)$s. As a result of this, the final merged resultlist may only contain documents from the query segments pertaining to these lists, thus missing relevant documents from the other query segments.

The major fallacy in the argument fo applying COMBSUM in our case is in the assumption that scores from the individual resultlists can simply be summed up. This hardly makes sense because the resultlists do not comprise of documents retrieved against the same query. In fact, our experimental investigation does show that COMBSUM is not an effective merging technique in this case thus supporting our hypothesis.

A work related to our proposed approach can be traced back to (Takaki et al., 2004), which describes decomposing a patent query into sub-topics and forming the final retrieval results by fusing individual retrieval results for the decomposed queries by a weighted combined summation of similarities. The major difference between our work and that of (Takaki et al., 2004) is that our motivation for query segmentation is driven by an effort to adapt query expansion (QE) for patent prior art search whereas QE was not addressed in (Takaki et al., 2004). This is a very important issue because most reported work on PRF for patent prior art search particularly on the CLEf-IP dataset report a negative PRF effect on retrieval effectiveness (Magdy and Jones, 2010a). Through our work, we hypothesize that by making use of more focused query segments, we may potentially exploit PRF to our advantage so as to improve IR effectiveness. In addition to this major difference, there exist other subtle differences between (Takaki et al., 2004) and our approach, listed as follows:

- The existing work involved segmenting query patents into sub-topics and extracting keywords from each of these sub-topics for retrieval, whereas we use the full text of each of the segments as individual sub-queries conforming to more recent findings suggesting the use of full patent text as queries (Xue and Croft, 2009; Wanagiri and Adriani, 2010).

- The existing work used a standard fusion technique of weighted COMBSUM (Fox and Shaw, 1994), whereas we show that a 1-way interleaving of the individual result-lists produces superior results to COMBSUM in our case.
- We do not distinguish between the relative importance of the individual subtopics by specificity measures as was done in (Takaki et al., 2004).

5.3 Experimental Setup

This section describes the experiments performed to evaluate our approach to patent retrieval using query segments. The evaluation of the segmented query retrieval methodology is conducted on the patent document corpus of CLEF-IP 2010 (see Section 3.2.1 for a detailed description of the dataset). Since the queries in patent prior art search are very long in comparison to standard ad-hoc search queries, a suitable query formulation technique is necessary for transforming the very long documents into queries which can be provided as input to a retrieval system for obtaining the resultlist of retrieved documents. Section 5.3.1 reviews some existing query formulation methodlogies for patent prior art search and describes the query formulation strategy we adapt in particular for our subsequent experimental investigations. The next section provides a description of the baselines and the parameters for the experimental investigations reported in this chapter.

5.3.1 Query formulation

Patent examiners typically formulate queries for (in)validating patent claims manually. This manual process often involves selecting high frequency terms from the text of a given patent claim (the query). Some early work on automatic keyword extraction to form a reduced query modelled on this real-life methodology of patent examiners includes that of (Takaki, 2005; Itoh et al., 2003). More recent work by Xue and Croft (2009) advocates the use of full patent text as the query to reduce the burden on patent examiners and concludes with the observation that usage of the whole patent text with raw term frequencies gives the best mean average precision (MAP). Recent work in the CLEF-IP [1] task has shown that best retrieval results are obtained when terms are used from all the fields of the query patents (Wanagiri and Adriani, 2010). The recent trends thus favour using full patent claim texts as queries to a patent document search system. One recent work demonstrates that a patent query formulated by extracting terms with frequencies higher than one (or in other words, removing terms which have frequency of one) outperforms the retrieval effectiveness obtained with full queries (i.e. when no terms are removed) (Magdy and Jones, 2010a).

The crucial observation which can be made is that while on one hand there is evidence which suggests that a full patent claim is more effective than short *keyword-style* queries, on the other hand there is empirical evidence that a reduced query (with the unit frequency terms removed) yields better retrieval results than the full patent claim text. This observation in turn leads to the conclusion that neither of the two extremes i.e. the (keyword queries) or the full text query is suitable for patent search, but rather an approach which is in between the two extremes is likely to be effective. Consequently, for all the experients on the CLEF-IP datatset reported in this thesis, we directly apply the fast and simple yet effective strategy of removing the unit frequency terms from the query text as prescribed in (Magdy and Jones, 2010a).

We now describe the experimental and then present the results, first for query segmentation based retrieval alone, followed by an application of PRF on these segmented queries.

[1] `http://www.ir-facility.org/clef-ip`

5.3.2 Baselines

The objective of our experiments is two-fold:

i) To explore whether decomposing a query into segments and retrieving with the individual segments can perform better than retrieving with the whole query. Note that while Takaki et al. (2004) already showed that constituting separate queries by extracting terms from each segment of the query text contributes to an increase in IR effectiveness, in our case we use the full text in each segment as separate queries following the work described in (Magdy and Jones, 2010a). Consequently, it is interesting to see whether we can achieve an increase in performance for our particular case.

ii) To investigate whether PRF can perform better on the individual query segments as compared to a whole query.

Keeping these two objectives in mind, the baselines are chosen as follows.

For the first objective, our baseline, which we call WHOLE, is a reproduction of the methodology of the second best performing run of CLEF-IP 2010, which is statistically indistinguishable than the best run (Magdy et al., 2011). We chose the second best performing run rather than the best one since the latter involves a series of complex processing steps which are very difficult to reproduce and are not significant for the purposes of our investigation (see (Magdy et al., 2011) for a more detailed comparison between the two approaches). For the second objective, i.e. to see whether PRF is improved for more focused query segments than on the whole queries, we have two baselines described as follows.

- WHOLE_PRF: PRF on the retrieval run WHOLE to measure the relative gains in the effectiveness of PRF when whole patents are used as queries.
- SEG: The initial retrieval results obtained by merging the resultlists retrieved against each query segment (the result of executing the method outlined in

Section 5.2 without executing the query expansion step, i.e. steps 6 and 7) in order to test the effectiveness of PRF on segmented retrieval.

5.3.3 Segmented Retrieval Implementation

The patent queries were segmented by applying TextTiling (Hearst, 1997). Our experiments used the TextTiling implementation obtained from the *MorphAdorner* package [2], where the default segmentation package is TextTiling. The other segmentation alternative which MorphAdorner provides is the C99 algorithm (Choi, 2000). A manual inspection of the segmentation outputs by these two methods revealed that the segments obtained by C99 are too short to encapsulate an information need in comparison to those obtained by TextTiling. The retrieval effectiveness of the segmented retrieval method (the evaluation metric values reported for the run named "SEG_RR" in Table 5.1), was in fact worse for C99. Hence, TextTiling was thus chosen as the segmentation algorithm for all our subsequent experiments on segmented retrieval.

TextTiling is not a part of the SMART system. Query splitting is therefore conducted as a pre-processing step before inputting each query segment to SMART individually. The fusion module which uses the COMBSUM (Fox and Shaw, 1994) and the round-robin techniques was implemented in SMART. Specifically speaking, this module takes as input a list of retrieval results stored in the SMART file format, and then combines them by applying the COMBSUM or round-robin methods to produce an output file also stored in the SMART format.

5.3.4 Parameters

TextTiling has two parameters: i) the *window size* and ii) the *step size*. Since sentence length can vary considerably, Hearst (1997) suggests to decompose the text into fixed length blocks of token streams. The parameter *window size* refers to the size of such token streams, the default value of which is 20 in the MorphAdorner

[2] `http://morphadorner.northwestern.edu/`

Table 5.1: Segmented vs. whole query retrieval.

Run Name	Parameters		Evaluation metric		
	Segmented	Fusion method	PRES	MAP	Recall
WHOLE	No	N/A	0.4413	0.0899	0.5310
SEG_COMBSUM	Yes	COMBSUM	0.1545	0.0308	0.1759
SEG_RR	Yes	Round-robin	**0.4949**	**0.0947**	**0.5982**

package as prescribed in (Hearst, 1997). The token streams or windows are then grouped together into blocks or pseudo-paragraphs. Blocks can be merged together if the inter-block similarities are high. The second parameter in TextTiling, namely the *step size*, refers to the size of these fixed length blocks. The default value of the block size is 10. For our experiments, we used these default parameter settings for TextTiling since TextTiling, in general, has been shown to work best with these parameter settings (Hearst, 1997).

As a QE technique, we use the LM score based QE as proposed by Ponte and Croft (Ponte, 1998) on the whole query and on the respective query segments. The two parameters for QE are the number of pseudo-relevant documensts, R, and the number of terms added for expansion, T. We varied R and T within a range of $[5, 20]$. The best settings for these parameters was found out to be $(R, T) = (10, 10)$, i.e. when we use 10 terms from top 10 documents. There was no separate training set used for training these parameters.

5.4 Results

In this section, we first report the retrieval results obtained by using segmented queries and then explore the use of PRF on these initial results.

5.4.1 Query Segmentation Results

In this section, we report the results of executing the method described in Section 5.2.1 without the QE step. It can be seen from Table 5.1 that the method of retrieving by separate query segments works well in conjunction with the one-way interleaving of documents returned for each query segment. By comparison, combination of documents by the standard fusion technique produces very poor results. The most likely reason for this observation is due to the fact that the standard fusion techniques have been devised to merge retrieval results obtained for the same query by different retrieval techniques. However, in our case we obtain the query segments by applying TextTiling to the full query description, which draws boundaries at sharp valleys of plotted cosine similarities between consecutive blocks of sentences. Thus the query segments, comprising of the textual contents of the output of TextTiling, are minimally similar to each other. The documents retrieved for each of the individual segments are mostly expected to be non overlapping, and hence not conducive to be fused by the standard technique of COMBSUM.

5.4.2 PRF Results

In this section we report the post feedback results both on whole queries and segmented queries. Table 5.2 presents the results, in this we include the whole and the segmented runs from Table 5.1 for the sake of continuity. Both the segmented runs reported in this table use 1-way interleaving. Also, recall that the columns R and T denote the number of pseudo relevant documents, and terms added for QE respectively.

Table 5.2: Pseudo Relevance Feedback on segmented retrieval.

Run Name	Parameters				Evaluation metric		
	Segmented	PRF	R	T	PRES	MAP	Recall
WHOLE	No	No	-	-	0.4413	0.0899	0.5310
WHOLE_PRF	Yes	Yes	10	10	0.4415	0.0889	0.5333
SEG	Yes	No	-	-	0.4949	0.0947	0.5982
SEG_PRF	Yes	Yes	10	10	**0.5033**	**0.1025**	**0.6166**

The table shows that the relative gains from QE are higher if it is performed on each of the segments separately, and the results then merged, as is evident from comparing the results of SEG_PRF and WHOLE_PRF. WHOLE_PRF results in almost negligible gains in PRES and average recall, and a very slight decrease of MAP. A possible reason for this very small change in the results confirms our hypothesis (see Section 5.1) that documents on different topics within the top ranked set contribute to a further decrease in the specificity of the overall information need. However, for the segmented case, since the queries are much shorter and focused to a precise information need, PRF in this case play a pivotal role in improving retrieval results. This can be verified from the fact that SEG_PRF retrieves a significantly larger number of relevant documents, as can be seen from the 3.1% relative increase in recall compared to the run SEG.

5.5 Analysis of Retrieval Segments

We have already seen that the proposed method of segmented retrieval produces overall better retrieval performance. It is particularly interesting to see the relative gains in retrieval effectiveness obtained for each individual query segment and to see how the aggregation of these per segment results can influence the overall retrieval performance. This section thus reports and analyzes the per segment retrieval performance for our proposed method. We first investigate the ranks of the relevant documents retrieved in each query segment. Next, we investigate the relative gains in feedback effectiveness for each retrieval segment.

5.5.1 Per Segment Ranks of Relevant Documents

Let us assume that we need to retrieve N documents for every original patent query and let τ be the average number of query segments over the set of whole patent queries. Thus, the expected number of documents we pick up from each list to construct the final retrieved set for the whole query, is $c = N/\tau$. The potential worst case of the segmented retrieval algorithm can arise when the retrieved sets of documents for each query segment do not intersect, and all the relevant documents have been retrieved at ranks beyond c.

For the CLEF-IP task, $N = 1000$ and from the output of TextTiling on the query set we find that $\tau = 17.66$, i.e. on average we end up decomposing every whole patent query into almost 18 segments. The expected farthest position in the ranked lists we need to visit during the interleaving process, starting from their tops, is thus $1000/17.66 \approx 57$. It is thus easy to see that our proposed algorithm can work well if all the query segments retrieve a high number of relevant documents within the top 57 positions. Hence it is interesting to see the number of relevant documents retrieved within the cut-off rank of 57.

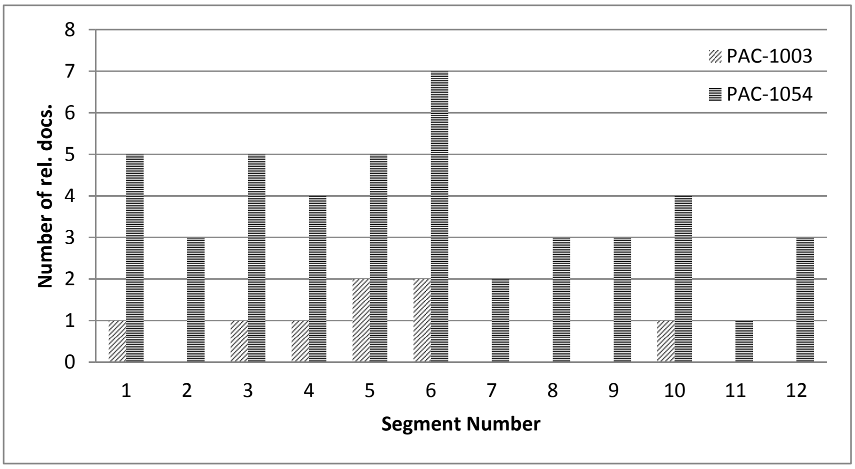


Figure 5.3: Per segment analysis of the best (PAC-1054) and the worst (PAC-1003) performing query.

Figure 5.3 shows the number of relevant documents retrieved within a cut-off value of 57 for two query instances described as follows.

i) PAC-1054: the query producing the maximum relative gain in PRES, i.e. producing maximum relative difference of PRES when SEG and WHOLE are compared.

ii) PAC-1003: the query with the maximum relative loss in PRES when SEG and WHOLE are compared.

The reason why query PAC-1054 is able to achieve good performance can be seen from the fact that the individual segments retrieve many relevant documents within

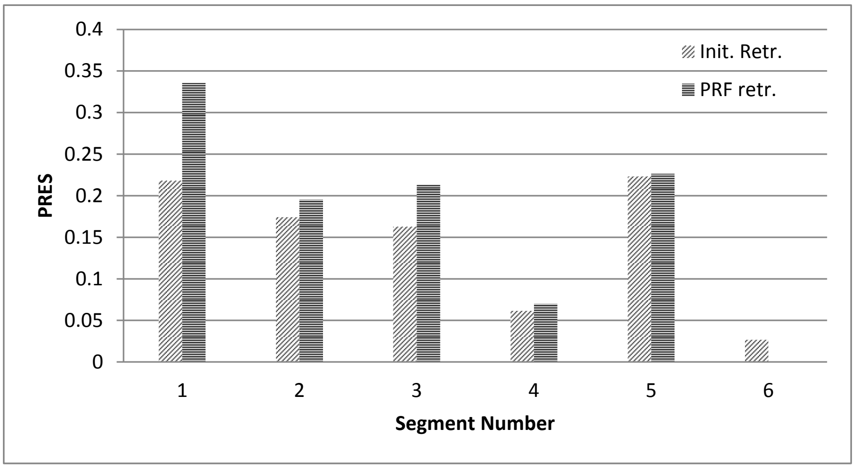


Figure 5.4: Feedback effect on each query segment for the query with the maximum gain in retrieval effectiveness through PRF, namely the query PAC-1038.

the average rank cut-off. In fact, the results show that the overall improvement in retrieval effectiveness of SEG as compared to WHOLE is in fact achieved by the cumulative improvements obtained on each query segment.

5.5.2 Per Segment PRF Performance

In Section 5.5.1, we compared the number of relevant documents retrieved within the top ranks of each query segment to those retrieved within the top ranks for the whole query. In this section, we investigate the relative performance gains achieved by PRF. We thus compare the performance of feedback between the whole queries and the segmented queries, i.e. we compare the relative retrieval effectiveness gain achieved by SEG_PRF over SEG and that of WHOLE_PRF over WHOLE.

In order to compare the maximum benefit in retrieval effectiveness achieved by the two PRF methods, i.e. one on the segmented and the other on the whole queries, we take a look into the queries with the maximum gain in feedback effectiveness achieved by the two methods. The best performing query in terms of relative PRES gain (from SEG to SEG_PRF) is the query named PAC-1038 having a 59.9% increase in PRES. The best performing query, involving PRF on whole queries, is the query named PAC-1036 with a relative gain (from WHOLE to WHOLE_PRF) in PRES of only 1.48%. The huge difference in the relative gains suggests that PRF in the case of segmented queries is much more effective than when applied to the whole queries, which in turn provides empirical justification to the hypothesis that multi-topical queries are not suitable for PRF.

To see whether we achieve a uniform performance gain over the query streams, we plot the PRES values for the initial retrieval alongside the PRES obtained after application of PRF in Figure 5.4. The figure shows that all query segments (except the one numbered 6) register an increase in PRES. It is thus seen firstly that the we obtain consistent increments in retrieval effectiveness for each query segment, and secondly that these consistent increments for each separate query segment contribute to a very large overall increase of 59.9% increase in PRES.

In order to see the feedback effects per query (or per query segment for the segmented retrieval), we categorize every query (segment) into bins of initial retrieval metric ranges. This analysis is similar to the analysis presented in Section 4.4, where in order to see the PRF effect on individual query groups, we categorized the queries by the initial retrieval average precision (AP) values obtained for them. It is particularly interesting to see the PRF effect on queries with low AP values, say in the range of $[0, 0.1)$ because these are the queries which can be considered as difficult or hard for the initial retrieval stage implying that these also have the highest scope for improvement in the feedback step. Thus, this way of categorizing the queries allows us to look at the performance over a group of queries having an initial retrieval measure of very poor $(0 - 0.2)$, poor $(0.2 - 0.4)$, average $(0.4 - 0.6)$, good $(0.6 - 0.8)$ or excellent $(0.8 - 1.0)$. For example, if the initial retrieval AP for 5 queries are 0.15, 0.23, 0.25, 0.68 and 0.52, we place the first query in bucket-1, the next two in bucket-2, the next one in bucket-4 and the last one in bucket-3. We categorize for the other metrics PRES and recall in an identical manner.

To provide a comparison between unsegmented and segmented feedback, Table 5.3 shows the number of queries belonging to each category, the number of queries improved in each category and the average relative gains in the three metrics for the runs WHOLE_PRF and SEG_PRF. From the table, we can see that

Table 5.3: PRF on whole vs. segmented queries

Run	Interval	PRES		MAP		Recall	
name	range	# qries	#improved	# qries	#improved	# qries	#improved
	[0.0,0.2)	13	3 (+0.18%)	41	20 (-1.54%)	12	0 (+0.00%)
	[0.2,0.4)	8	6 **(+0.25%)**	8	4 (-0.95%)	5	0 (+0.00%)
WHOLE_PRF	[0.4,0.6)	11	5 (+0.03%)	0	0 (+0.00%)	8	0 (+0.00%)
	[0.6,0.8)	11	4 (-0.01%)	1	0 (+0.00%)	10	1 **(+1.68%)**
	[0.8,1.0]	7	6 (+0.07%)	0	0 (+0.00%)	15	0 (+0.00%)
	[0.0,0.2)	472	235**(+14.04%)**	775	433**(+157.77%)**	357	81**(+25.09%)**
	[0.2,0.4)	328	213 (+3.55%)	9	2 (-1.71%)	391	79 (+3.33%)
SEG_PRF	[0.4,0.6)	54	46 (+1.33%)	0	0 (+0.00%)	103	7 (+0.31%)
	[0.6,0.8)	35	27 (+0.23%)	0	0 (+0.00%)	38	0 (+0.00%)
	[0.8,1.0]	0	0 (+0.00%)	105	47 (-17.78%)	0	0 (+0.00%)

WHOLE_PRF results in a very slight increase of PRES in each query group, whereas the method SEG_PRF yields a considerable increase in percentage gain of PRES, MAP and recall for the segment group $[0, 0.2)$. The next group, i.e. $[0.2, 0.4)$, also registers a significant increase of PRES. In addition to PRES, the average gain in MAP and recall are huge for the query group $[0, 0.2)$ as can be seen from the first row of SEG_PRF. The huge improvements in PRES, MAP and recall for the first group shows that feedback in this case is improving the retrieval effectiveness of query segments for which the initial retrieval results are poor, which in turn shows that the segmented PRF is able to increase retrieval effectiveness significantly for the queries for which retrieval is difficult during the initial retrieval stage. This observation verifies our hypothesis that query expansion can be successfully applied to patent search if the queries are decomposed into shorter and unambiguous segments. The selection of the expansion terms in query expansion inherently involves computation of a term score based on the similarity with the query. The very long queries can result in the choice of irrelevant terms for expansion, whereas the short focused queries enable a more accurate term score computation thus leading to a more judicious choice of the expansion terms.

5.6 Summary

This chapter addressed the research question RQ-2 introduced in Chapter 1, which explores whether segmentation of very long queries improve IR effectiveness. In our proposed method, we first segment the patent queries by the text segmentation method TextTiling (Hearst, 1997), retrieve against each of these segments and finally merge the results in a round-robin way.

Retrieval using query segments results in significantly improved retrieval quality. The experimental results show gains in terms of all the three evaluation metrics, namely PRES, recall and MAP for retrieval with segmented queries in comparison to retrieval with whole queries. This confirms the hypothesis that patent prior art search task performs well when more focused sub-queries with precise information needs are used for retrieval. Merged retrieval with separate query segments shows that the relevant documents in fact pertain to one or more of the fine-grained aspects in a query. Query expansion is also shown to perform well on segmented retrieval, which proves that the shorter and more focused queries proves beneficial in increasing the effectiveness of PRF.

In the current and the preceding chapters, we have explored segmentation for documents and queries. The results have shown that both document and query segmentation can improve retrieval effectiveness. Segmentation of the documents into sentences makes use of term proximity to predict term relatedness and helps us to restrict the choice of feedback terms to relevant parts of documents, whereas segmentation of the long queries serves to focus on each fine-grained aspect of the queries during retrieval. Instead of having these two methodologies as stand-alone techniques, it would be more useful to combine the ideas under a unifying framework. The next chapter describes our work in this direction.

Chapter 6

Topical Relevance Model: Topical Segmentation of Pseudo-Relevant Documents and Queries

In the work presented in Chapters 4 and 5 respectively on query expansion based on topically related terms predicted by term proximity within sentences, and segmentation of long queries into topically coherent units, it was observed that topically related terms in both pseudo-relevant documents and the queries can be beneficial to improve the effectiveness of the expansion methods in PRF. While Chapter 4 showed that the use of sentence based segmentation proves effective for the standard ad-hoc search, Chapter 5 revealed that both sentence and sub-topic based segmentation of very long queries are beneficial for effective retrieval.

The preceding two chapters explored the first two research questions RQ-1 and RQ-2, namely *"Can additional terms in close proximity to query terms from retrieved documents enrich the information need of the query and improve the retrieval effectiveness of ad-hoc IR?"* and *"Can segmentation of very long queries into topically coherent segments be utilized to improve IR effectiveness?".* The work in this chapter in an attempt to address these two complementary questions in a single framework seeks to investigate the third research question RQ-3 introduced in

Chapter 1, namely *"Can topic modelling prove beneficial in improving the retrieval effectiveness for both short and long queries thus unifying the solutions of RQ-1 and RQ-2?."* To this end, firstly, we aim to develop a model combining the segmentation of documents and queries into one framework. Secondly, we explore topic based inference of a probability distribution of words to topics within a document instead of content-based segmentation which assigns a particular region (a sentence or a paragraph) of a document to a topic.

The chapter is organized as follows. In Section 6.1, we start with a brief motivation for development of a topic-based model before describing the details of our model. This is followed by Section 6.2 which first gives a schematic description of our proposed method as a generalization of the relevance model, Section 6.3 presents a detailed evaluation of our proposed model on two different query types, i.e. keyword type short queries and much longer queries as encountered in associative document search. This is followed by Section 6.4, where we analyze the performance and investigate the robustness of the model for various query types with different amounts of specificities in the information need. Finally, Section 6.5 concludes the chapter with a summary and outlook.

6.1 Motivation

From the results of our experiments using segmentation for long queries as presented in Chapter 5, we noted that retrieval using topically coherent query segments followed by a merging of the result-lists demonstrates that it can be beneficial to split very long queries into a sequence of more focused topical statements. The methodology presented in Chapter 5, although effective, is associated with the following drawbacks.

1. Retrieval needs to be done separately for each sub-topic, the results of which then need to be merged. An obvious disadvantage of the query segmentation prior to the retrieval phase, is that retrieval needs to be executed for each such

query segment, which is inefficient in practice.

2. In Chapter 5 we adopted a simple technique of merging the individual result-lists, obtained by retrieving against each query segment, in a round-robin technique. The drawback of the round-robin merging policy is that it assumes equal importance for all the query segments which may not be true in practice. For example, returning back to our example in Section 5.1, the relative importance of the topic of nitride growing may be higher than that of pit formation in invalidating the query patent. Assigning relative importance to sub-topics while merging the results would involve another level of optimization of the weights assigned to sub-queries using specificity measures, as recommended in (Takaki et al., 2004).

3. Two non-adjacent query segments, which represent blocks of text such as paragraphs in the query text, can be on the same topic. However, a text segmentation method is unable to detect this, as a result of which, the retrieval results obtained for these two non-adjacent segments may have a high topical overlap. An ideal case would be to merge these two segments into one, which is not possible in contiguous segmentation methods such as TextTiling used in the experimental investigations in Chapter 5. This can however be achieved through topic-based segmentation.

4. Individual facets of an information need are assumed to be represented by consecutive blocks of text (the output of text segmentation), which need not necessarily be true. Each individual segment may in fact be a mixture of two or more topic classes. This cannot be represented by contiguous segmentation methods.

This motivates us to develop a retrieval model with the following inherent characteristics to address the above issues. The first problem can be addressed by postponing the segmentation step to the feedback stage, i.e. after the initial retrieval result has been obtained. This way of postponing the segmentation in the feedback stage makes the whole process more efficient in terms of run-time.

To overcome the second problem, the model should take into account the different aspects of relevance with different weights in the sense that the weights, instead of being optimized separately, should be the intrinsic outcome of the model itself. For example, one may use a weighted round-robin technique to merge the result-lists obtained by retrieving against individual query segments. However, the weight for each segment has to be optimized separately and it is in fact very difficult to predict the relative importance of each query segment and optimize the weight of each segment without explicitly modelling the word-topic relationships. Instead, we will see that our proposed model can make use of the word-topic mappings to predict the relative importance of topics in the query. These values can then act as weights to assign more importance to one topic than the others during the PRF.

To address the last three issues, we need to use a segmentation method more general than the contiguous segmentation. More precisely speaking, we apply the more flexible and general method of *topic-based segmentation* in comparison to contiguous segmentation. In topic-based segmentation, instead of mapping regions to topics, we map words to topics. The imposed probability distribution of terms over a set of topics ensures that a term can in theory belong to multiple topics with different probabilities. This behaviour is more pronounced for terms which are associated with multiple meanings. Let us illustrate this with a simple example. The word *bank* can either be related to the topic *finance* (i.e. when the word *bank* is used in the sense of financial institution), or to the topic *nature* (i.e. when it is used in the sense of land alongside a river). In a collection of documents, the word *bank* can appear in both senses. For each sense, it can co-occur with other words in the same topic, e.g. when used in the former sense, it can co-occur with words such as *money*, *credit* etc., whereas for the latter sense it can co-occur with words such as *river*, *land* etc. The word bank will thus have considerable probabilities of membership in the clases *finance* and *nature*. It is not possible to model these term-topic relationships

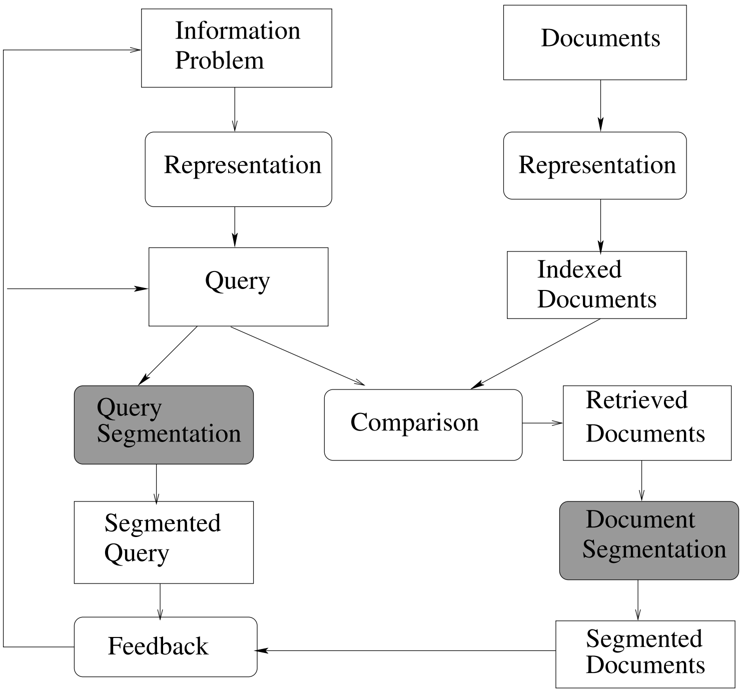


Figure 6.1: Feedback using a combination of document and query segmentation.

in contiguous segmentation methods.

The previous two chapters have shown that predicting term relationships with natural sentence boundaries and using content based segmentation for the long queries can improve PRF quality. However, none of the methods proposed in the previous two chapters made use of topic-based segmentation by statistically modelling the topic-term distributions. The aim of this chapter is then to combine these two in a single feedback model by making use of the topical distribution. The schematic view of such a model, which is described in more details in the subsequent sections of this chapter, is shown in Figure 6.1. Note that feedback in the model essentially involves matching the topics in the documents with those of the query instead of trying to match whole documents with the query. This enables the model to focus on individual topics in the documents and the query similar to the objectives of the research questions RQ-1 and RQ-2 which make use of segments or sub-parts of documents and the query respectively.

6.2 Topical Relevance Model: A Generalization of the Relevance Model

This section presents our proposed topical relevance model (TRLM) as a generalization to the relevance model. We introduced the relevance model (RLM) in Chapter 2.1.4. To recapitulate, the RLM involves estimating a posterior distribution for generating the set of relevant documents. The observable variables in the RLM are the query terms and the set of pseudo-relevant documents. The model then estimates the probability of generating the unobserved set of relevant documents given that they can be generated effectively using the same model as for the observed query terms. The RLM utilizes the co-occurrences of a query term with a term occurring in the pseudo-relevant documents.

A more general approach however would assume that the relevance model itself is a mixture model of topics which in turn generates terms in the relevant documents. This approach can model the fact that there can be several multiple facets associated with the query generated to express an information need. Latent topic nodes can represent more such fine grained facets.

Two generalization approaches are presented one for the standard keyword type queries and the other for the very long query types encountered in associative document search. To cater for the different characteristics of the two types of queries, we propose two variants of our model: one with the assumption that terms in a query are generated by sampling from a number of relevance models each of which can relate to a specific aspect of the potential information need; and the other with the assumption that each relevance model belonging to a particular topic generates its own set of query terms. Simply speaking, in the first variant the query itself is not topically segmented because the queries are too short. In the second variant however, the query itself is topically segmented. We call the two variants the unifaceted topical relevance model (uTRLM) and the multifaceted topical relevance model (mTRLM), respectively. The naming convention has been adopted with respect to the query. For short queries, we assume that each query term belongs to the same topic class (hence the name uni-faceted); whereas for the long ones it is assumed that query terms can belong to different topic classes (hence the name multi-faceted) implying that a document is relevant if it satisfies the information need expressed in one or more parts of the query, e.g. a prior art is relevant for one or more claims in a patent query.

6.2.1 Uni-faceted Topical Relevance Model

The relevance model (RLM), introduced in Section 2.1.4, has an oversimplified assumption in that all the relevant documents are generated from a single generative model. Under such a scheme it is difficult to model the observed fact that retrieved documents in fact tend to form clusters of topics (Xu and Croft, 1996). While this multi-topical nature of the retrieved set of documents might be hard to explain through the standard relevance model, it can be easily modelled through our proposed topic-based generalization of the RLM. The multiplicity of the topics in the retrieved set of documents may then be realized by the relevance model being a mixture model of relevance from various topics, where each such topical relevance model generates its own set of words in the relevant documents. It can be imagined in an ideal scenario that each topic in the retrieved set of documents manifests from one particular aspect of the query.

Returning back to the example query introduced in Section 1.1, a generalized relevance model will be able to explain the various topics on polio disease in general, its outbreaks, medical protection against the disease, post-polio problems etc. as being generated by the mixture model of topical relevance. This generalized model may thus be able to provide a better estimation of relevance at the level of topics, associating a subset of topics to a subset of potential information need aspects of the query. Let us now take a closer look at the proposed model.

The working principle of the uTRLM is depicted schematically in Figure 6.2. Let R represent the underlying relevance model that we are trying to estimate. In

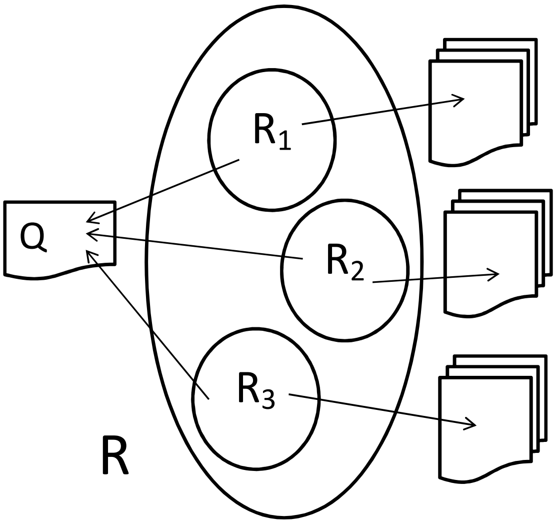


Figure 6.2: A Uni-faceted topical relevance model (uTRLM).

the standard RLM, it is assumed that words from both the relevant documents and the query are sampled from R, as shown in Figure 2.2. In contrast to this, the unifaceted topical relevance model (uTRLM) assumes that a query expresses a single overall information need, which however in turn encapsulates a set of potential sub-information needs. This is shown in Figure 6.2, where each sub-information need is encapsulated within the more global and general information need R. This is particularly true when the query is broad and is comprised of a wide range of underspecified information needs. The uTRLM thus assumes that the relevance model is a mixture model, where each model $\mathrm{R_i}$ generates words in potentially relevant documents addressing a particular topic.

The unifaceted relevance model is a more generalized treatment of the first research question RQ-1, which investigated whether terms in close proximity to query terms from retrieved documents enrich the information need of the query and improve IR effectiveness. Similar to SBQE described in Chapter 4 to investigate RQ-1, the working principle of the uTRLM also involves choosing topically related terms for feedback from the pseudo-relevant set of documents. However, the differences are that firstly in contrast to predicting term relatedness on the basis of proximity within sentence boundaries, uTRLM explicitly models the latent topics in the pseudo-relevant set of documents in order to infer the topical relatedness between terms. Secondly, uTRLM is a generative model whereas SBQE is not. Thirdly, SBQE is associated with query expansion and hence is a two-step retrieval process involving a subsequent retrieval with the expanded query following the initial retrieval, whereas the basic working principle of the uTRLM on the other hand lies in reranking the result-list of documents obtained after the initial retrieval, the reranking being performed on the basis of how close the document language models are to the estimated relevance model. It is however possible to extend uTRLM to include query expansion.

6.2.2 Multi-faceted Topical Relevance Model

Another generalization which can be made to the RLM is for the case when a query explicitly conveys multiple distinct information needs. For example, the queries in patent prior art search, explored in Chapter 5 fall under this category since the objective is to retrieve relevant documents (prior art) on each claim. In Chapter 5, we proposed a method of segmenting a query into non-overlapping blocks of text and then using each block as a separate query for retrieval and then finally merging the results. The results showed that such an approach of retrieving with segmented increases retrieval effectiveness. Consequently, this illustrates that such long queries are comprised of a number of different information needs, in the ideal case each of these relates to one query segment. However, this is a limitation of our earlier method which our proposed topical relevance model tries to address. In the context of our proposed mode, it is reasonable to assume that each topical relevance model generates its own set of relevant documents and its own subset of query terms pertaining to one topic. In the ideal case, each claim of the patent query can be mapped to a distinct topic. This is shown in Figure 6.5, which shows that $\mathrm{R_i}$ generates words in relevant documents pertaining to the i^{th} topic along with a subset $\{Q_i\}$s of query terms associated with the same topic.

The multifaceted topical relevance model (mTRLM) is a more generalized treatment of the second research question, namely RQ-2 which investigated whether it

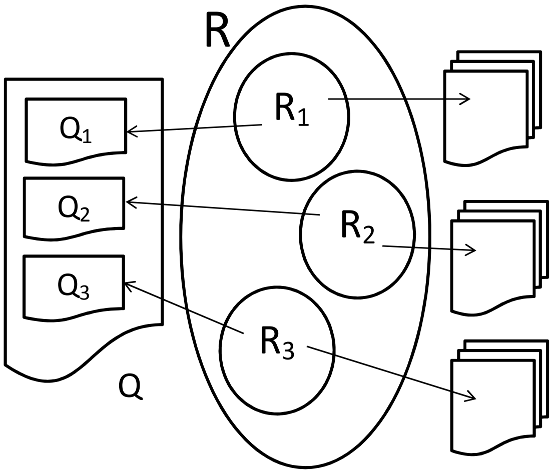


Figure 6.3: A multifaceted topical relevance model (mTRLM).

is beneficial to segment very long queries for improving the retrieval effectiveness. Similar to the segmented query retrieval method, as described in Chapter 5, the mTRLM segments the top ranked pseudo-relevant documents into multiple regions of topics. mTRLM thus has the similar objectives as PRF on segmented reteieval methodlogy. The difference are highlighted as follows. Firstly, retrieval with query segments use segmentation of the queries only, whereas the mTRLM employs topical segmentation of both the documents and the queries. Secondly, retrieval with segmented queries involves retrieval against each query segment followed by a merging of the results. The number of retrieval steps is thus identical to the number of query segments. In contrast, the mTRLM involves only a single retrieval step. Thirdly, the approach in Chapter 5 identified topically coherent blocks of text in the query by predicting topic shift points through TextTiling. TextTiling however does not explicitly model topics and hence it is not possible to merge two topically similar non-adjacent segments into one unit. Explicitly modelling the topics enables mTRLM to address this limitation.

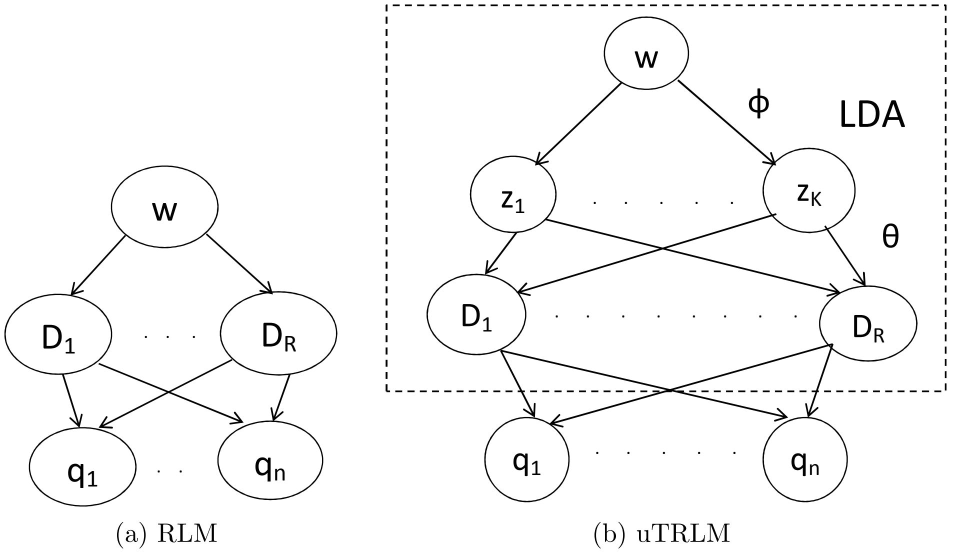
6.2.3 Estimation Details

In this section, we present details on how the topical relevance model (TRLM) is *estimated*. By estimation, we mean inferring the posterior probabilities of generating a term w from the relevance model R itself. Similar to the relevance model (RLM), introduced in Section 2.1.4, these probability estimates are then used to rerank a set of initially retrieved documents by measuring how similar their term generation models are to the estimated relevance models (see the introduction to RLM in Chapter 2.1.4).

Since TRLM uses the same estimation technique as RLM, for the convenience of reading we reproduce the equation for the relevance model estimation in Equation 6.1.

$$P(w|\mathrm{R}) \approx P(w|q_1, \ldots, q_n) = \prod_{i=1}^{n} P(w|q_i) \tag{6.1}$$

A careful look at Equation 6.1 shows that it is impossible to compute $P(w|\mathrm{R})$ exactly because in practice the set of relevant documents for a query is unknown. Assuming its existence would defeat the whole purpose of IR. The estimation of the model, therefore, has to be done by using the observed events of the generation of the query terms. One thus needs to approximate the probability of generating a non-query term w from the relevance model R, by the probability of generating w given that the model has already generated the query terms $q_1, \ldots, q_n$. This probability is $P(w|q_1, \ldots, q_n)$, which is thus used as the approximated probability of generating a term from the relevance model R, as shown in Equation (6.1).



(a) RLM

(b) uTRLM

Figure 6.4: Dependence graphs for the RLM and the unifaceted TRLM.

The dependence graph for the generative model of RLM is reproduced in this chapter as Figure 6.4a for reading convenience. In the TRLM, instead of assuming that a word is directly generated from a document language model we assume that a word w can be generated from a finite universe of topics $z = \{z_1, \ldots, z_K\}$ (see Figure 6.4b, where each topic z_i addresses the relevance criterion expressed in the sub-relevance model $\mathrm{R_i}$, as shown in Figure 6.2. Let us say that $z \in \mathrm{R}^K$ follows a multinomial distribution $\phi \in \mathrm{R}^K$, with the Dirichlet prior β for each ϕ_i. Each document $d \in \{D_j\}_{j=1}^R$ in turn comprises of a number of topics, where it is assumed that a topic $z \in \{z_k\}_{k=1}^K$ is chosen by a multinomial distribution $\theta \in R^K$ with the Dirichlet prior α. With this terminology, we derive the estimation equations for the two variants of TRLM in the next two sections .

Unifaceted TRLM

The dependence graph of a unifaceted TRLM is shown in Figure 6.4b. Let us assume that the query terms $\{q_i\}_{i=1}^n$ are conditionally sampled from multinomial unigram document models $\{D_j\}_{j=1}^R$, where R is the number of top ranked documents obtained after an initial retrieval step. Every query term q_i is generated from a document D_j with $P(q_i|D_j)$, similar to the RLM as shown in Figure 6.4a. Each $P(w|q_i)$ in turn is given by

$$P(w|q_i) = \sum_{j=1}^{R} P(w|D_j)P(D_j|q_i) \tag{6.2}$$

Due to the addition of a layer of latent topic nodes, there is no longer a direct dependency of w on D_j, as in the RLM (see Figure 2.3 and Equation (2.11)). Hence to estimate $P(w|D_j)$, we need to marginalize this probability over the latent topic variables z_k. Thus, we have

$$P(w|D_j) = \sum_{k=1}^{K} P(w|z_k)P(z_k|D_j) \tag{6.3}$$

Substituting Equation (6.3) in Equation (6.2) and applying Bayes rule, we obtain

$$\begin{aligned} P(w|q_i) &= \sum_{j=1}^{R} \frac{P(q_i|D_j)P(D_j)}{P(q_i)} \sum_{k=1}^{K} P(w|z_k)P(z_k|D_j) \\ &\approx \frac{1}{R} \sum_{j=1}^{R} P(q_i|D_j) \sum_{k=1}^{K} P(w|z_k)P(z_k|D_j) \end{aligned} \tag{6.4}$$

The last step in Equation (6.4) is obtained by discarding the uniform prior $P(q_i)$. The inner summation of Equation (6.4) is the LDA document model, which is identical to Equation (2.18). The LDA document models over the set of pseudo-relevant documents can be estimated by the Gibbs sampling. The Gibbs sampling equations for LDA inference, introduced in Chapter 2 are reproduced in this chapter in Equation 6.5.

$$\begin{aligned} P_{LDA}(w|d_i, \hat{\theta}, \hat{\phi}) &= \sum_{j=1}^{K} P(w|z_j, \hat{\phi})P(z_j|d_i, \hat{\theta}) \\ &= \sum_{j=1}^{K} \frac{(n_w^{(z_j)} + \beta)(n_j^{(d_i)} + \alpha)}{(\sum_{w'=1}^{V} n_{w'}^{(z_j)} + V\beta)(\sum_{j'=1}^{K} n_{j'}^{(d_i)} + K\alpha)} \end{aligned} \tag{6.5}$$

The LDA inferencing over the set of pseudo-relevant documents is shown by the box labelled “LDA” in the dependence graph of Figure 6.4b. Substituting Equation (2.18) in (6.4), we obtain

$$P(w|q_i) = \frac{1}{R} \sum_{j=1}^{R} P(q_i|D_j) P_{LDA}(w|D_j, \hat{\theta}, \hat{\phi}) \tag{6.6}$$

$P(q_i|D_j)$ is the standard probability of generating a term q_i from a smoothed unigram multinomial document model D_j, and is defined as

$$\begin{aligned} P(q_i|D_j) &= \lambda P_{MLE}(q_i|D_j) + (1-\lambda)P(q_i) \\ &= \lambda \frac{\mathrm{tf}(q_i|D_j)}{\sum_{q \in D_j} \mathrm{tf}(q, D_j)} + (1-\lambda) \frac{\mathrm{df}(q_i)}{\sum_{q \in V} \mathrm{df}(q)} \end{aligned} \tag{6.7}$$

Equation (6.7) represents the language modelling (LM) similarity of the query term q_i with document D_j, identical to Equation (2.4). λ is a smoothing parameter and $P_{MLE}(t|d)$ is the maximum likelihood estimate of occurrence of a term t in document d. Substituting Equations (6.7) and (2.18) (the expression for) into Equation (6.6) gives

$$\begin{aligned} P(w|q_i) =& \frac{1}{R}\sum_{j=1}^{R}[\{\frac{\lambda \cdot \mathrm{tf}(q_i, D_j)}{\sum_{q\in D_j}\mathrm{tf}(q, D_j)} + \frac{(1-\lambda)\cdot \mathrm{df}(q_i)}{\sum_{q\in V}\mathrm{df}(q)}\}\times \\ &\{\sum_{k=1}^{K}\frac{(n_w^{(z_k)}+\beta)(n_k^{(D_j)}+\alpha)}{(\sum_{w'=1}^{V}n_{w'}^{(z_k)}+V\beta)(\sum_{k'=1}^{K}n_{k'}^{(D_j)}+K\alpha)}\}] \end{aligned} \tag{6.8}$$

Equation (6.6) has a very simple interpretation in the sense that a word w is more likely to belong to the topical relevance model if:

- w co-occurs frequently with a query term q_i in the top ranked documents, and
- w has a consistent topical class across the set of pseudo-relevant documents.

It can also be seen from Equation (6.6) that the uTRLM uses a document model $P_{LDA}(w|D)$, different from the standard unigram LM document probability $P_{LM}(w, D)$ for a document D, as shown in Equation (2.4). This may be interpreted as smoothing of word distributions over topics, similar to that described in (Wei and Croft, 2006). Using marginalized probabilities $P(w|z_k)$ in Equation (6.3) leads to a different maximum likelihood estimate to $P(w|D)$, which is the standard maximum likelihood of a word w computed over the whole document D (see Equation (6.3).

Moreover, the TRLM estimation also ensures that each topic is estimated separately with variable weights as given by the prior for each topic, namely $P(z_k|D_j)$. This is because the product of $P(q_i, D_j)$ and $P_{LDA}(w, D_j, \hat{\theta}, \hat{\phi})$ will be maximized if each of them are maximized individually (i.e. attains values close to 1), which essentially indicates that q_i occurs frequently and w has a consistent topical class across the set of pseudo-relevant documents.

Multifaceted TRLM

The difference between the multifaceted model and the unifaceted one is the way in which query terms are sampled from document models. While in the uTRLM, a query term is directly generated from a document model as shown in Equation (6.7), the query term generation probability is marginalized over the latent topic models in the multifaceted variant, as shown in Figure 6.5. Thus, the mTRLM models the fact that not only the pseudo-relevant documents, but also a query comprises multiple topics. So, in the mTRLM, it is not only the documents which are segmented into topics, but so is the query as well. This is shown by the additional layer of latent topic nodes inserted between the document nodes and the query term nodes in Figure 6.5.

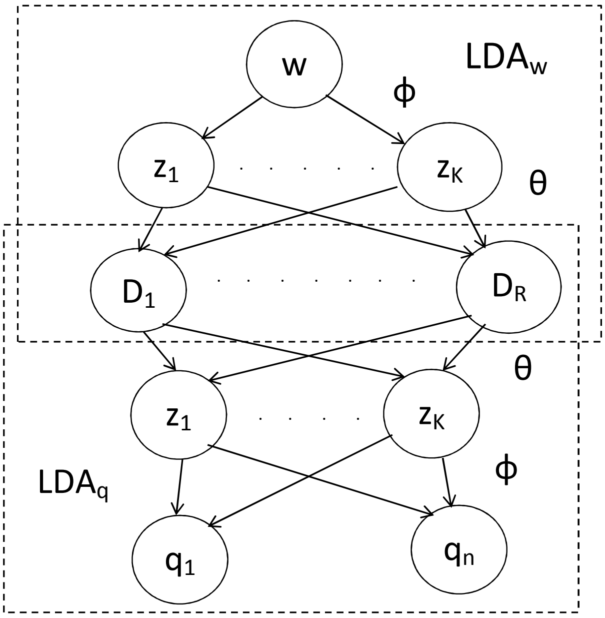


Figure 6.5: Dependence graph for the multifaceted TRLM.

Taking into account the latent topics in a query, $P(q_i|D_j)$, Equation (6.6) has to be marginalized over the topic nodes as shown below. This marginalization ensures that we take into account the topical class of each query term while estimating the model.

$$P(q_i|D_j) = \sum_{k=1}^{K} P(q_i|z_k)P(z_k|D_j) \tag{6.9}$$

Substituting Equation (6.9) in Equation (6.6), and ignoring the denominator $P(D_j)$ by assuming uniform priors, leads to the modified TRLM equation for the multi-faceted model.

$$\begin{aligned}P(w|q_i) =& \frac{1}{R}\sum_{j=1}^{R}\big(\sum_{k=1}^{K}P(q_i|z_k)P(z_k|D_j)\big)P_{LDA}(w|D_j,\hat{\theta},\hat{\phi})\\ =& \frac{1}{R}\sum_{j=1}^{R}P_{LDA}(q_i|D_j,\hat{\theta},\hat{\phi})P_{LDA}(w|D_j,\hat{\theta},\hat{\phi})\end{aligned} \tag{6.10}$$

The LDA probabilities can be substituted from Equation (2.18) as was done in Equation (6.8). Equation (6.10) thus involves two levels of LDA estimated term generation probabilities, one for the words in pseudo-relevant documents and the other for the query terms. This is shown by the two boxes LDA_w and LDA_q respectively in Figure 6.5. Equation (6.10) also has an intuitive appeal in the sense that it assigns higher probability of generating a term from the relevance model, if the term co-occurs with a query term in pseudo-relevant documents and is also likely to belong to the same topic as the query term.

6.2.4 Algorithm for TRLM

Following a formal presentation of the estimation details, we now provide the steps for implementing the two variants of TRLM.

1. Run initial retrieval using standard language model (Hiemstra, 2000) (see Equation 2.4).

2. Let R be the number of top ranked documents assumed to be relevant.

3. Let W be the working set of documents on which LDA is to applied. For uTRLM, $W = \{D_j\}_{j=1}^{R}$ whereas for mTRLM, $W = \{D_j\}_{j=1}^{R} \cup Q$.

4. Perform LDA inference by N iterations of Gibbs sampling on the document set W to estimate the parameters $\hat{\theta}$ and $\hat{\phi}$ (see Equations 2.16 and 2.17).

5. For each word w in the vocabulary V of W, repeat steps 5 (a) and (b).

   (a) Use Equations (6.8) and (6.10) respectively for uTRLM and mTRLM to compute $P(w|q_i)$ for each query term $q_i \in Q$.

   (b) Use Equation (2.10) to compute $P(w|Q) \approx P(w|\mathrm{R})$.

6. Rerank each document by the KL divergence between its unigram document model and the topical relevance model so as to get the final retrieval result (Lavrenko and Croft, 2001). The KL divergence is computed as

$$KL(D||\mathrm{R}) = \sum_{w \in V} P(w|\mathrm{R}) \log(\frac{P(w|\mathrm{R})}{P(w|D)}) \tag{6.11}$$

   The KL divergence or Kullback-Leibler divergence, a metric derived from information theory, is useful in measuring the *distance* between two probability distributions. More formally speaking, the KL divergence between two probability distributions P and Q, i.e. $KL(P, Q)$, denotes the number of extra bits required to code samples from P when using a code based on Q. In our case, while reranking a document, we use this metric to compute the distance between the document model itself and the estimated relevance model. The lower this distance, the more likely it is that the document is relevant. The documents are then reranked by increasing values of their KL divergence values from the reference distribution, which in our case is the estimated relevance model $P(w|\mathrm{R})$. More specifically, we use Equation 6.11 to compute $KL(D||R)$ for each document D in the initially retrieved set of documents, and then rerank this set in ascending order of the $KL(D||R)$ values. After reranking, the top ranked document is the one with least KL divergence value from the R distribution, or in other words, is the closest to the relevance model and hence is the most likely document to be relevant.

The computational complexity of the above algorithm is $O(VRKN)$, where V is the number of terms in the pseudo-relevant documents, R is the number of pseudo-relevant documents, K is the number of topics, and N is the number of iterations used for Gibbs sampling. The computational complexity of RLM on the other hand is $O(VR)$. Since both K and N are small constant numbers independent of R and V, TRLM is a constant times more computationally expensive than RLM. This means that there is a little additional overhead involved in the TRLM in comparison to the RLM.

6.3 TRLM Evaluation

In this section, we evaluate the TRLM. We start this section with a description of the experimental settings and the parameter settings for TRLM and then present the comparative results of TRLM against the baselines.

6.3.1 Experimental Setup

Dataset

The uTRLM models the relevant documents for short queries, where the query itself is not segmented into topics. The uTRLM is thus tested on the TREC collection (TREC 6-8 and TREC Robust) since the titles of the ad-hoc queries comprise of a few keywords. In fact, since the objective of the uTRLM is identical to that of SBQE (cf. Chapter 4), we use the same dataset that was used to test the latter. See Table 3.2 for a more detailed description of the dataset. Choosing an identical dataset enables us to have a direct comparison between these two methods.

In addition, we use a set of longer queries namely the TREC Robust TDN (title, description, narrative) topics to test both uTRLM and mTRLM. The rationale behind using longer queries as compared to TREC title only queries, is to see how the two variants of the model perform for queries which have an intermediate length between the two extremes of either being very short comprising of a few key words or being very long as in associative document search. We expect that the TREC TDN queries may not be ideal candidates for testing mTRLM since it is rather impractical to assume that these queries are truly multi-topical. We thus restrict our choice to only a subset of the TREC queries just to see the transition effect of mTRLM while moving from very short to slightly longer queries, and later on we conduct experimental investigations with mTRLM on the CLEF-IP dataset, the queries in which are truly multi-topical. We chose the TREC Robust subset in particular because these are the queries which are more difficult to improve with PRF (Voorhees, 2004).

To evaluate the TRLM on very long queries we again use the CLEF-IP[1] 2010 dataset, which comprises of a collection of patents from the European patent office. Since mTRLM is an alternative method to achieve the same objective as the segmented query retrieval method described in Section 5.2 does, it is reasonable to conduct the experiments on the same dataset so as to enable a direct comparison between the results.

Baselines

Since the evaluation objective is to examine whether the TRLM outperforms the RLM, we used the RLM as one of our baselines. Additional term-based query expansion with query re-weighting on top of RLM estimation (denoted from now on as RLM_QE) has been found to improve retrieval effectiveness further in comparison to RLM without query expansion (Lang et al., 2010). This approach is thus also used as a stronger baseline for comparison against TRLM.

Note that the TRLM baselines are in fact identical to the SBQE baselines (cf. Section 4.3). The TRLM baseline however does not include the approach where additional expansion terms based on the LM term scores are added to the initial query, namely the LM_QE approach of Chapter 4. The reason is that LM_QE has already been demonstrated to be weaker than the RLM baselines (cf. Figure 4.3a and 4.3b).

[1] `http://www.ir-facility.org/clef-ip/`

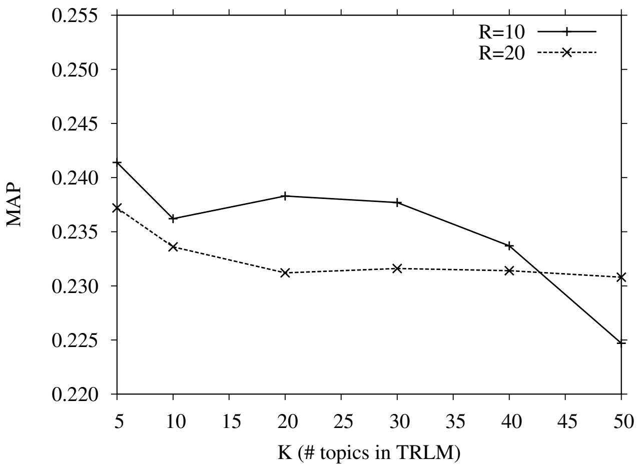


Figure 6.6: Optimizing the TRLM parameters on the TREC-6 dataset.

Implementation of TRLM

The LDA estimation for TRLM was implemented inside the SMART system itself. For this, a part of the C++ code[2] for LDA inference by Gibbs sampling was ported to C. The KL divergence based reranking for the RLM and TRLM was also implemented within SMART.

6.3.2 TRLM Parameters

Common Parameters

The smoothing parameter of the LM in the initial retrieval (see Equation (6.7) was set to 0.4 similar to SBQE (cf. Section 3.3). Similar to our experiments for SBQE (cf. Section 4.3), we used the TREC-6 dataset as the training set to optimize the parameters, namely R (the number of pseudo-relevant documents), T (the number of terms for query expansion in RLM_QE) and K (the number of topics in TRLM). The tuning of these parameters was performed by varying them within the maximum bound of 50.

[2] http://gibbslda.sourceforge.net/

LDA Hyper-Parameters

The hyper-parameters α and β, which control the Dirichlet distributions for TRLM (see Equations 2.16 and 2.17), were set to $\frac{50}{K}$ and 0.1 respectively as suggested in (Griffiths and Steyvers, 2004). This is a reasonable setting since it has been found that a value of $\alpha = \frac{50}{K}$ maximizes the posterior likelihood of $P(w|z)$, whereas it has been reported that values of β considerably higher than 0.1 typically result in formation of coarse-grained topics and values of β much lower than 0.1 usually yield very fine-grained topics. A value of β close to 0.1 is ideal because of the optimality in the granularity of topical representation (Griffiths and Steyvers, 2004). The number of iterations for Gibbs sampling was set to 1000 for all TRLM experiments as suggested in citepgriffiths04.

Number of Topics

An important parameter in the TRLM is the number of topics K. We conducted experiments to investigate the sensitivity of retrieval effectiveness on the number of topics used in TRLM. The results are plotted in Figure 6.6, which shows how the retrieval effectiveness, as measured using the MAP, varies with the number of topics used for uTRLM. The figure shows that optimal results are obtained by using a small value of K, and that the average retrieval effectiveness tends to decrease with increasing values of K. The optimal results are obtained with the setting of $R = 10$ and $K = 5$. We thus use the same settings of R and K for the test datasets.

TRLM Query Expansion Parameter

In principle, similar to RLM_QE, the method of term based query expansion can also be applied to TRLM. Instead of applying the RLM scores for selecting the expansion terms for the subsequent feedback step, we use the TRLM scores for doing so. Similar to RLM_QE, we search for the optimal settings of the parameters R (the number of pseudo-relevant documents for feedback) and T (the number of expansion terms) within the range $[5, 20]$ on the TREC-6 dataset, which is used

Table 6.1: Comparative evaluation of RLM, RLM_QE, uTRLM and uTRLM_QE on TREC topics (TREC-6 topics were used for parameter training).

TREC	MAP				
Dataset	LM	RLM	RLM_QE	uTRLM	uTRLM_QE
6	0.2075	0.2061 (-0.67%)	0.2279 (9.83%)	**0.2484**$^{*+}$(19.71%)	0.2439* (17.54%)
7	0.1614	0.1673 (3.65%)	0.1714 (6.19%)	**0.1816*** (12.51%)	**0.1914**$^{*+}$(18.59%)
8	0.2409	0.2302 (-4.44%)	0.2612 (8.42%)	**0.2631*** (9.21%)	**0.2875**$^{*+}$(19.34%)
Robust	0.2618	0.2796 (6.79%)	0.3236 (23.60%)	**0.3351**$^{*+}$(27.99%)	**0.3410**$^{*+}$(30.25%)

Table 6.2: The recall values for the runs reported in Table 6.1.

TREC	Recall				
Dataset	LM	RLM	RLM_QE	uTRLM	uTRLM_QE
6	0.5167	0.5167 (0.00%)	0.5307 (2.71%)	0.5167(0.00%)	0.5648(9.31%)
7	0.4795	0.4795 (0.00%)	0.5177 (7.98%)	0.4795(0.00%)	0.5348(11.55%)
8	0.5559	0.5559 (0.00%)	0.6055 (8.93%)	0.5559(0.00%)	0.6582(18.40%)
Robust	0.7860	0.7860 (0.00%)	0.8381 (6.64%)	0.7860(0.00%)	0.8715(10.87%)

as the training set for tuning the parameters. The optimal parameter settings on TREC-6 is given by $(R, T) = (10, 10)$, i.e. when 10 documents and 10 terms are used for query expansion. We used this setting of R and T on the test datasets.

6.3.3 uTRLM Evaluation

The unifaceted topical relevance model (uTRLM) without and with query expansion is tested on the TREC 6,7,8 and Robust title-only queries, since due to the short length of these queries it is not reasonable to run the multifaceted version of TRLM which assumes that query terms belong to multiple topical classes.

Table 6.1 shows the mean average precision (MAP) for these TREC title only runs. The run uTRLM_QE involves additional query expansion based on the TRLM scores. Table 6.2 shows the recall values for the corresponding runs reported in Table 6.1. A '*' and a '+' in Table 6.1 indicates statistically significant improvement of TRLM or TRLM_QE over RLM and RLM_QE respectively. It is seen from Table 6.1 that the TRLM significantly outperforms the RLM for all query sets. TRLM also significantly outperforms RLM_QE, i.e. RLM with explicit term-based query

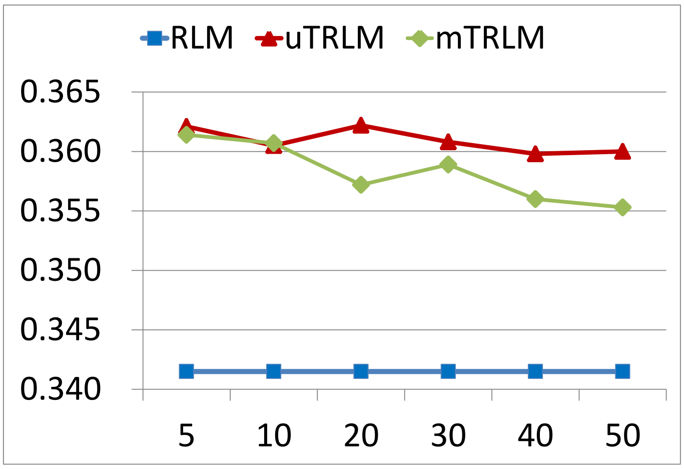


Figure 6.7: Comparative performances of the RLM, uTRLM and the mTRLM on the TREC Robust TDN queries.

expansion, even though the latter performs a second-step retrieval with additional expansion terms yielding a higher recall as seen from Table 6.2.

Query expansion on top of TRLM proves particularly beneficial as can be seen by the uTRLM_QE MAP values. It can be seen that although the uTRLM_QE results are slightly worse than the TRLM results for the TREC-6 dataset, the results are significantly better than uTRLM without query expansion for the other test datasets. This significant increase in MAP can be attributed to the significantly higher recall values achieved by uTRLM_QE. This also demonstrates that the expansion terms selected by uTRLM_QE are more useful than the expansion terms selected by the RLM_QE.

RLM is particularly ineffective on the TREC-8 topic set, where reranking documents using the RLM in fact decreases MAP with respect to the initial retrieval by 4.44%. RLM_QE however is able to increase MAP significantly (8.42%) for this topic set due to an increase in recall caused by the query expansion (see Table 6.2). Even without increasing recall, TRLM is able to outperform RLM_QE. This provides empirical evidence of a more accurate and more robust estimation of the relevance model compared to RLM. TRLM with query expansion further improves performance on the TREC 8 dataset.

Figure 6.7 shows that the uTRLM is more suitable than the mTRLM for short queries. Figure 6.7 plots both uTRLM and mTRLM results on the TREC Robust TDN queries, the number of pseudo-relevant documents being 10 for both the reported sets of results. It can be observed from Figure 6.7 that uTRLM performs slightly better than mTRLM. However both perform significantly better than the RLM.

The reason why uTRLM performs slightly better than mTRLM on TREC Robust TDN queries is that these queries, although almost 10 words in length on average, do not genuinely express a multifaceted information need. The results however indicate that genuinely multi-topical queries are required to evaluate mTRLM, which we report in the next section.

6.3.4 mTRLM Evaluation

Chapter 5 showed that patent prior art search is a challenging problem. Table 5.2 showed that query expansion is not beneficial for patent retrieval, since expansion terms tend to add more ambiguity to the already very long and ambiguous queries Similar observations are reported in (Magdy et al., 2010). A solution to this problem was proposed in Chapter 5 involving decomposition of a patent query into segments of thematically related topics with the help of a text segmentation method using each segment as a separate query, and finally merging the results. The main problem with this method however is that retrieval is slow, as documents need to be retrieved against a number of queries with a subsequent merging of result lists. Furthermore, the notion of topics in a text segmentation algorithm is restricted to contiguous blocks of text.

Multifaceted topical relevance models can solve this problem efficiently because they represent a query as comprised of multiple topics, where each topic can be associated with a particular information need. A further advantage of the mTRLM is that it does not require multiple retrieval steps. It infers topics by analyzing the space of pseudo-relevant documents and the query in contrast to simply segmenting the query text.

Table 6.3: Retrieval by mTRLM on CLEF-IP 2010 data.

Metric	LM	RLM	RLM_QE	SEG	uTRLM	mTRLM
MAP	0.0960	0.1081	0.0947	0.0947	0.1056	**0.1095**
PRES	0.4235	0.4536	0.4260	0.4949	0.4508	**0.4561**

Table 6.3 shows the results for mTRLM. A comparison is provided with RLM, RLM_QE and uTRLM. Along with MAP, we also report the results for the PRES values (Magdy and Jones, 2010b), which focus on recall at early ranks.

We observe that mTRLM outperforms both RLM and uTRLM on these topics. The benefit of mTRLM over RLM is not statistically significant. However mTRLM achieves a significantly higher MAP over the initial retrieval result LM, whereas the RLM improvement over LM is not significant.

We also note that mTRLM yields better retrieval results (in terms of MAP) compared to that of the segmented retrieval method SEG_RR as proposed in Section 5.2 (reproduced in the table with the column name SEG for the sake of direct comparison), without explicitly using separate query segments as sub-queries. This shows that mTRLM can be an effective technique to focus on each query aspect of a long query and retrieve against each query topic (aspect) in only one retrieval step.

The reason, we believe, for the lower PRES values in mTRLM is due to the averaging effect of using the same number of topics for all the patent queries, which in fact are largely different from each other in terms of the granularity of the information need that they express. The SEG method actually takes this into account because the number of topically coherent segments obtained for each query are different. We hypothesize that the performance of mTRLM on the CLEF-IP patent dataset can be further improved by individually choosing the value of K for each query. We discuss more in this regard in Sections 6.4.2 and 6.4.5.

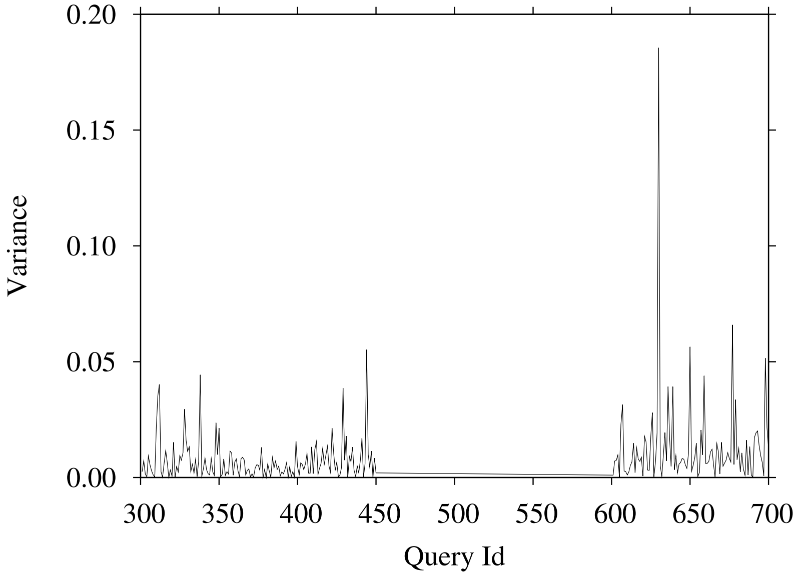


Figure 6.8: Variance in the MAP values for different values of K (number of topics) in the range $[2, 50]$.

6.4 Discussions and Further Analysis of TRLM

In this section, we firstly illustrate with an example that the value of K, i.e. the number of topics, used in the estimation of TRLM, largely depends on the specificity of the information need of the query itself. Next, we show that varying K across the queries does improve the retrieval effectiveness averaged over a query set. Next, we investigate the robustness of TRLM in the presence of noisy feedback documents.

6.4.1 Per-query Sensitivity to Number of Topics

Figure 6.6 shows that TRLM is relatively insensitive to the choice of the number of topics used for LDA estimation, i.e. the value of K. However this is the average effect over a query set, thus we discuss the effect on individual queries in more detail in this section. We manually looked at the MAP values of the title queries of TREC 6, 7, 8 and Robust queries for different K values in the range of $[2, 50]$. We found that only 24 of 250 queries register a MAP standard deviation higher than 0.02, which suggests that MAP is fairly insensitive to the choice of K and performance is stable for the majority of the TREC 6-8 and Robust queries. This is shown in Figure 6.8.

After manually looking at the queries with large variances in the MAP values, we observed three distinct patterns of MAP variations for different values of K: i) a sharp increase, ii) a peak, and iii) a sharp decrease with increasing K. Figure 6.12 highlights the observations for three queries with the highest variances in MAP values.

The first case, i.e. a sharp increase in MAP with increasing K, can be exemplified by query 630, where we note a sharp increase in the MAP values with an increase of K, which suggests that the scope of the information need expressed in this query is broad, as a result of which the pseudo-relevant documents for this query are associated with a high number of diverse topics. The description of this query reads "*Retrieve documents containing information about the symptoms of individuals suffering from 'Gulf War Syndrome' as a result of serving in the Gulf War*", which suggests that the wide range of symptoms occurring in different individuals tend to form separate topics, the and model is thus optimized for a high value of K. The narrative also suggests that there are several facets or aspects of information need in the query such as the illness from service in the Gulf War, contacts with Gulf War veterans, desert storm syndrome etc as shown in Figure 6.9.

<top>
<num> Number: 630
<title> Gulf War Syndrome
<desc> Description:
Retrieve documents containing information about the symptoms of individuals suffering from ’Gulf War Syndrome’ as a result of serving in the Gulf War.
<narr> Narrative:
Documents regarding law suits that claim causes of illness from service in the Gulf War are relevant, as are reports of cases resulting from contact with an ill Gulf War veteran. ‘Dessert Storm Syndrome’ is a synonym for the condition and is considered relevant.
</top>

Figure 6.9: The TREC query 630 which shows an increase in MAP with increasing values of K (number of topics in TRLM).

The case of a distinct peak in MAP is illustrated by query 650. The peak is suggestive of the ideal number of relevant topics for this particular query. The narrative of this query reads "*A relevant document will contain details about large-scale tax evasion. Documents about people who lost in excess of two million dollars as a result of doing business with an organization indicted for tax fraud are relevant*". This query elaborately expresses two broad information needs, firstly about tax evasion, and secondly about the people who lost money. Both these can in turn address individual sub-topics, e.g. there can be many different types of organizations involved in tax evasion.

<top>
<num> Number: 650
<title> tax evasion indicted
<desc> Description:
Identify individuals or corporations that have been indicted on charges of tax evasion of more than two million dollars in the U.S. or U.K.
<narr> Narrative:
A relevant document will contain details about large-scale tax evasion. Documents about people who lost in excess of two million dollars as a result of doing business with an organization indicted for tax fraud are relevant.
</top>

Figure 6.10: The TREC query 650 which shows an optimal value in MAP for a value of K (number of topics in TRLM) in between the two extremes.

The third case is seen for query 444 which suggests that the information need expressed in this query is very specific. The narrative of this query reads "*To be relevant, a document must indicate that the fluid involved is achieved by a process of pressurization producing the supercritical fluid*", which in fact is a very precise information need. TRLM for this query thus yields the optimal result with only 2 topics, and the MAP decreases with increasing number of topics.

The specificity of the information need of a query can somewhat be quantified by the clarity score measure (Cronen-Townsend and Croft, 2002). The clarity scores of the three example queries 630, 650 and 444 are 454.76, 653.53 and 1842.12 respectively, which conforms to the manual analysis of their specificity of information need. We have thus seen that the parameter, namely the number of topics (K), depends

<top>
<num> Number: 444
<title> supercritical fluids
<desc> Description:
What are the potential uses for supercritical fluids as an environmental protection measure?
<narr> Narrative:
To be relevant, a document must indicate that the fluid involved is achieved by a process of pressurization producing the supercritical fluid.
</top>

Figure 6.11: The TREC query 444 which shows a decrease in MAP with increasing values of K (number of topics in TRLM).

largely on the specificity of the information need expressed in a query. The next section explores whether choosing the best settings of K for a given query (assuming the existence of an oracle) can in fact help in improving the retrieval effectiveness over a set of queries significantly.

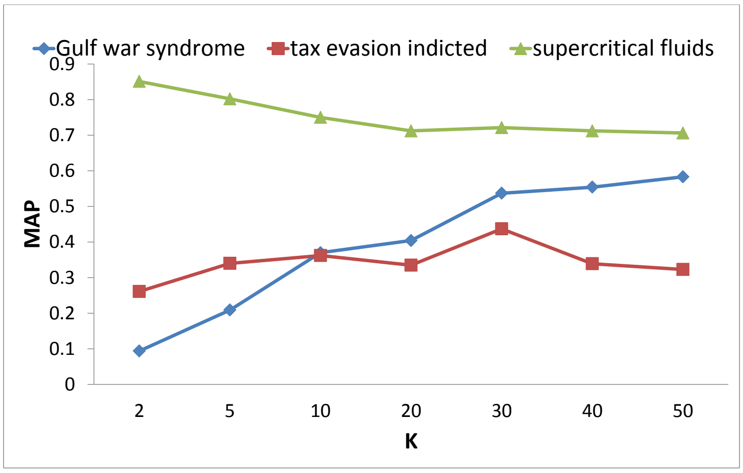


Figure 6.12: Effect of K (number of topics) on MAP for three example queries.

6.4.2 Adapting the Number of Topics

The intention of this part of the study is to see the maximum retrieval effectiveness which can be obtained by choosing the K values individually for each query instead of using a fixed value of K. This is analogous to the targeted improvements in standard query expansion by adapting the number of feedback terms and documents

Table 6.4: MAP values obtained by dynamically choosing the optimal number of topics per query (TRLM*) on the TREC dataset.

	TREC-6	TREC-7	TREC-8	TREC-Robust
TRLM	0.2484	0.1816	0.2631	0.3351
TRLM*	0.2588	0.1855	0.2731	0.3437

Table 6.5: MAP and PRES values obtained by dynamically choosing the optimal number of topics per query (TRLM*) on the CLEF-IP dataset.

	TRLM	TRLM*
PRES	0.4508	0.5028 (11.53%)
MAP	0.1095	0.1261 (15.16%)

per query (Ogilvie et al., 2009), or by selecting only good feedback terms (Cao et al., 2008; Leveling and Jones, 2010).

Let us assume that there is an oracle which tells us the best K value to use for each query by looking at the MAP values obtained for all different values of K, and returns the one for which the MAP is maximum. For example the oracle returns $K = 50$ for query 630 (see Figure 6.12). Table 6.4 shows the best possible results that can be obtained by dynamically choosing the number of topics for each individual query. We see that by using the optimum value of K, additional significant improvements over standard TRLM can be obtained (TRLM* in Table 6.4). This in turn demonstrates the potential of the method to be further optimized by a dynamic choice of the number of topics based on a query feature classification approach, similar to (Cao et al., 2008).

We have already discussed that in the context of patent search, this individual choice of the number of topics can lead to a significant performance gain because the number of topics in a patent query is related to the invention claims which it makes. We now look at the IR effectiveness achieved by the optimal version of mTRLM on the CLEF-IP task. The results are shown in Table 6.5. We see that significant performance gains

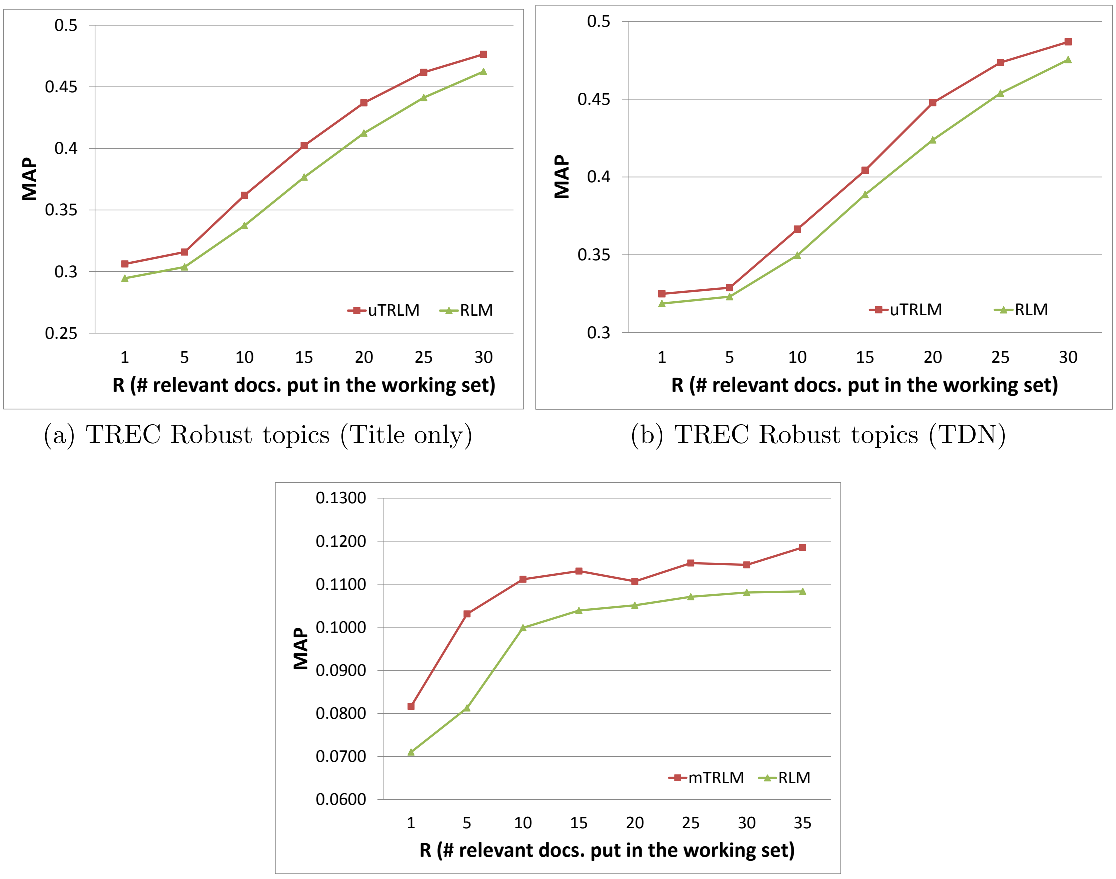
It is generally difficult in practice to implement a predictor approximating such an oracle with satisfactory precision. Topic models which attempt to automatically infer the number of most likely topics from a collection of documents may be used to achieve a varying number of topics for each individual query (Blei et al., 2010). Another idea is to run TRLM with different values of K and use standard fusion techniques such as COMBSUM to merge the resultlists (Fox and Shaw, 1994).

6.4.3 Robustness Analysis

To test the robustness of the model in the presence of supervised training samples, i.e. known relevant documents, we report a series of experiments where we insert a number of true relevant documents in the working set W of top ranked documents, retrieved during the initial retrieval, to see if the model can perform better under the presence of a mixture of pseudo-relevant and true relevant documents. We start with the assumption from PRF that all the $|W|$ top ranked documents are relevant. Then we inject R known relevant documents into this working set, and taking out the first R non-relevant ones from W while doing so. Thus the number of relevant documents in the working set W will be at least R, while the number of non-relevant documents will be at most $|W| - R$. To illustrate with an example, let $W = \{D_1, \ldots, D_5\}$ be the 5 top ranked documents out of which D_4 is relevant. Let us suppose that we want to inject $R = 1$ document into W. We then look for a relevant document down the ranked list. Let us suppose that we find that D_7 is relevant. W is thus modified as $W \leftarrow W \cup \{D_7\} - D_5$. Note that we make use of the available relevance judgements to know if a document in the working set of top ranked documents during the initial retrieval is truly relevant.

Adding more relevant documents into the working set gradually increases the *signal-noise* ratio. To investigate the robustness of the model in the presence of true relevant documents, we first start with inserting 1 relevant document in the working set and then gradually increase this number. The intention is to see whether a topical relevance model can filter out the noise, and utilize the known relevance information better than the standard relevance model.

Results shown in Figure 6.13 indicate that TRLM outperforms RLM for all values



(a) TREC Robust topics (Title only)

(b) TREC Robust topics (TDN)

(c) CLEF-IP topics

Figure 6.13: TRLM and RLM effectiveness in the presence of true relevant documents.

of injected true relevant documents into the working set. In Figure 6.13, MAP is plotted along the Y-axis, while the X-axis shows the number of relevant documents injected into the working set. Note that the length of the queries gradually increases from (a) to (c). For both the T and TDN variants on TREC Robust queries, it can be seen that TRLM outperforms RLM consistently for all ranges of the number of true relevant documents. This suggests that TRLM is more robust to noise, i.e the presence of non-relevant documents in the top ranked documents.

It can also be seen from Figure 6.13c that mTRLM significantly outperforms the RLM for patent queries, again proving that TRLM consistently outperforms RLM even under the presence of non-relevant documents in the feedback step. Particularly interesting is the significant difference in MAP between mTRLM and RLM, i.e. the difference between the left most points on the mTRLM and RLM graphs, which again demonstrates that mTRLM is more effective in utilizing the relevant information from even a small number of documents from the working set of pseudo-relevant documents.

True relevance feedback is often available in patent prior art search domain because patent searches are typically conducted by professional searchers who are willing to meticulously examine a considerable number of retrieved documents and also willing provide feedback to the system (Magdy, 2011). In such scenarios, mTRLM should definitely be the preferred search model to use because of its high retrieval effectiveness in the presence of true relevant information.

6.4.4 Comparison of uTRLM with SBQE

In this section, we compare the results of uTRLM with SBQE, the sentence based query expansion method proposed in Chapter 4. The SBQE method showed us choosing feedback terms from sentences based on the assumption that such terms are topically related due to their proximity helps improve retrieval effectiveness. The likely reason for improvement is due to the fact whole documents are seldom relevant to a query. SBQE makes direct use of this assumption by making use of information from the document segments (sentences) most similar to the query while discarding others.

The same effect is achieved by the TRLM in a bit more subtle way, explained as follows. Recall that TRLM is mainly motivated by the observation that the inherent multi-topical nature of the information need expressed in a query manifests as the clusters of topics in the pseudo-relevant documents. Some of these topics directly relate to the relevant aspects of the information need, as a result of which using information from these topics help improve the retrieval effectiveness. TRLM, in fact achieves document segmentation by imposing a probability distribution $P(w|\mathrm{R})$ over the terms in the pseudo-relevant set of documents, with the effect that terms related to the relevant aspects of the query are assigned more weight (through co-occurrence evidences) than the terms which are not.

Table 6.6: Comparison between SBQE (Chapter 4) and uTRLM without and with query expansion (denoted respectively by the w/o QE and QE columns) on the TREC dataset.

TREC	MAP			P@10			Recall		
Topics	SBQE	uTRLM		SBQE	uTRLM		SBQE	uTRLM	
		w/o QE	QE		w/o QE	QE		w/o QE	QE
6	0.2481	0.2484	0.2439	0.4280	0.4560*	0.4425	0.5668	0.5167	0.5648
7	0.1963	0.1816	0.1914	0.3280	0.3220	0.3220	0.5342	0.4795	0.5348
8	0.2891*	0.2631	0.2875	0.4300	0.4520*	0.4460*	0.6404	0.5559	0.6582
Robust	0.3540*	0.3351	0.3410	0.4182	0.4273*	0.4273*	0.8745	0.7860	0.8715

For a direct comparison of the retrieval effectiveness obtained by SBQE and TRLM, selected information from Tables 4.2 and 6.1 are merged together in Table 6.6. The results show that SBQE performs significantly better on TREC-8 and TREC Robust data sets than uTRLM, as measured by the MAP (shown by the asterisks). For the remaining two data sets, i.e. the TREC 6 and 7, the corresponding MAPs are not significantly different.

TRLM with query expansion shows that the MAP values achieved with uTRLM with QE (uTRLM_QE) is very close to those obtained with SBQE. The retrieval effectiveness of uTRLM_QE is however better due to the higher (significant for two cases, namely TREC-8 and TREC Robust) P@10 values. For both SBQE and uTRLM_QE, the precision at top 10 documents are sacrificed at the cost of increasing the recall. However, the decrease in P@10 is lower in the case of uTRLM_QE.

The decision of whether to use uTRLM or SBQE is a trade-off between average precision quality and the execution time. uTRLM is faster than SBQE because it is a single-step process involving only a re-ranking of the current ranked list as against the two step retrieval process of SBQE where the original query is expanded and results are retrieved in the second step with the expanded query. This also explains why SBQE is better than TRLM because retrieval with the expanded query in the feedback step typically increases recall, thus favouring the MAP metric. However, the precision at high ranks such as P@10 are significantly higher in uTRLM for three topic sets (marked by the asterisks), as shown in Table 6.6. This increase in recall is also achieved by uTRLM_QE. The advantages of uTRLM_QE over SBQE are as follows.

- uTRLM_QE involves a significantly lower number of expansion terms than SQBE yet achieveing statistically indistinguishable performance compared to SBQE in terms of MAP.
- The execution time of uTRLM_QE is lower than SBQE due to the lower number of query terms in the feedback step. The average number of expansion terms for SBQE is 465.88 (cf. Table 4.6), whereas the average number of terms for uTRLM_QE is close to 13 (10 expansion terms plus 3 initial query terms on an average).
- The P@10 values obtained by uTRLM_QE are considerably better (though not significantly) than SBQE.

The second retrieval step in uTRLM can lead to an increase in recall and MAP much faster than SBQE. uTRLM alone without the query expansion could be the preferred choice for high precision oriented searches because the precision at top ranks for uTRLM is significantly better than that of SBQE. A subsequent retrieval step with the expanded query obtained by the use of uTRLM term scores is able to increase the recall while not hurting the precision too much at top ranks. uTRLM_QE could thus be the preferred choice for recall oriented searches because the query execution run-time in the subsequent retrieval step in uTRLM_QE is lower than that of SBQE, yet attaining close (sometimes better) recall values in comparison to SBQE.

6.4.5 Comparison of mTRLM with retrieval by separate query segments

In this section, we compare the results of retrieval by separate query segments (SEG), as proposed in Chapter 5 with that of mTRLM. Recall that one of the motivations of developing the mTRLM was that the method of retrieval with separate query segments is slow, since retrieval has to be done for each segment separately and then the results need to be merged. It is intersting to see how the results of mTRLM, which involves a single retrieval run compares with that of SEG. Ideally, the mTRLM should achieve results close to SEG.

Table 6.7: Comparison between segmented query retrieval with mTRLM.

Method	Metric		
	MAP	PRES	Recall
SEG	0.0947	0.4949	0.5982
SEG_PRF	0.1025	0.5033	0.6166
mTRLM	0.1095	0.4508	0.5792
mTRLM*	0.1261	0.5028	0.6003

In Table 6.7, we revisit the results shown previously in Tables 5.2 and 6.3. Here, we observe that the average precision of the mTRLM approach is higher than that obtained either by the method SEG alone or SEG combined with PRF (cf. Section 5.3).

The segmented query retrieval algorithm however, results in a significantly higher recall and PRES. Previously, we discussed that a likely reason for the low PRES values may be attributed to the fact that we use a single value of K (the number of topics in TRLM) for all queries in the testset. In Section 6.4.2, we discussed how varying K individually for each query may affect retrieval quality. The results from Section 6.4.2 for the CLEF-IP task is reproduced in Table 6.7. We can see that mTRLM*, where we make use of the optimal value of K for each query, is able to achieve better values of PRES and recall compared to SEG.

The decision of whether to use mTRLM or SEG for patent search is a trade-off between the high recall quality and high precision with low execution time. If in an application, the recall quality can be sacrificed to some extent so as to gain much faster execution time and higher average precision, then mTRLM should be used instead of SEG.

6.5 Summary and Outlook

This chapter presented the topical relevance model (TRLM), a theoretical framework of generative relevance, exploiting the topical association of terms in top ranked pseudo-relevant documents. The proposed model is a generalization of the standard relevance model, overcoming its limitations of term independence. The proposed method is conceptually similar to the Markov random field based extensions of RLM (Metzler and Croft, 2007; Lang et al., 2010), in the sense that our model also attempts to model term dependencies, but has a different working principle in the sense that our model applies LDA to group terms into topics in the generative process of RLM.

Two variants of TRLM were proposed: the unifaceted model (uTRLM) and the multifaceted model (mTRLM). While uTRLM assumes that a query expresses slightly different aspects of the same information need in an implicit fashion, mTRLM works on the principle that a query is explicitly structured into a number of diverse topics. The difference in the working principle between uTRLM and mTRLM is that while the former uses LDA smoothed pseudo-relevant document models, the latter additionally uses an LDA smoothed query for relevance model estimation.

Results confirm that the unifaceted model is suitable for short queries such as the TREC style ad-hoc search queries. The multifaceted model on the other hand is suitable for associative document search tasks, where a full document is used to find related information from a collection. In effect, we have integrated the separate document and query segmentation approaches described in Chapters 4 and 5 respectively, into one framework.

The uTRLM produces comparable results with SBQE in the sense that while the former yields better run-time and precision at top ranks, the latter produces better average precision. The mTRLM, on the other hand, results in comparable retrieval effectiveness with respect to SEG (retrieval with separate query segments and fusion of results). In this case, while the former results in a better MAP, the latter produces better recall and PRES at the cost of a much increased run-time.

The key contributions of this chapter are:

- theoretical justification of the use of topic models in local context analysis addressing aspects of relevance;
- investigating the use of LDA smoothed document and query models for relevance model estimation;
- an effective technique for associative document retrieval in a single retrieval step without explicitly fragmenting the query into contiguous segments;
- empirical validation of TRLM, which shows that TRLM outperforms RLM and RLM_QE on queries of diverse types and lengths; and
- investigation of the dynamic choice of the number of topics in TRLM, which further improves retrieval effectiveness.

TRLM relies on modelling the topics in the set of pseudo-relevant documents and the query. The distribution of topics in the documents or the query is however used internally to improve the retrieval quality. The users of a search system, which applies TRLM for feedback, thus may not even know about this processing step which is performed entirely in the backend. In the next chapter, we explore whether this information about the topics can be made available to the users of a search system with an aim to provide more flexibility and interactions in their search behaviour.

Chapter 7

TopicVis: An Interface for Topic based Feedback and Navigation

This chapter demonstrates the application of our topical relevance model (TRLM), introduced in Chapter 6, to help a user locate topically relevant information within a set of retrieved documents. A disadvantage of traditional ranked list retrieval is the difficulty in locating relevant segments of information within individual documents. A standard ranked list returned by a search system in response to a query typically comprises of a list of document titles and snippets with highlighted query words. A user while browsing through this list has to read the snippets and make a decision whether a document is likely to contain the relevant information. Users of standard search interfaces, in general, cannot make this decision quickly without reading the snippet.

Many documents are often expository in nature and contain information on multiple topics. In these cases, the user will often be interested in a single or atmost a few of the topical subsets within a document. Consequently, methods to help the user to locate specific relevant information within the top-ranked retrieved documents, are of considerable potential utility.

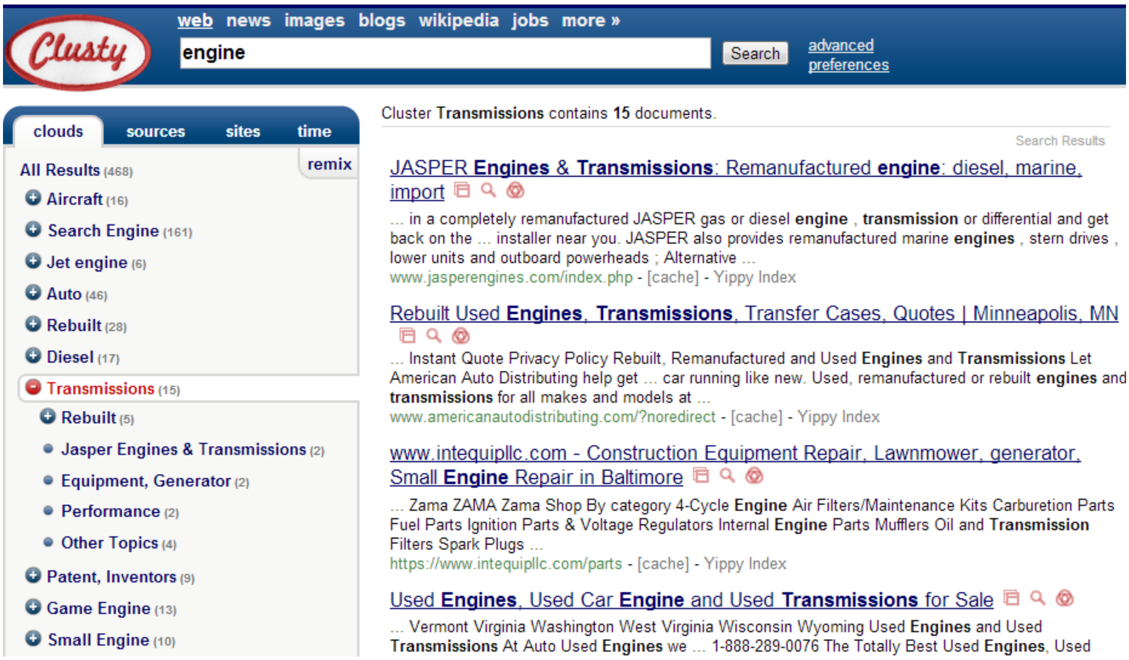
Taking these observations as motivating factors, we propose an information access approach within the list of retrieved documents through topic visualization.

To this end, we describe a web interface supporting topic-based visualization of the retrieved documents and a mechanism to support topic-based navigation through search results coupled with topic-based interactive feedback. This chapter describes the details and evaluation of our developed system, which we name *TopicVis*.

7.1 Background and Motivation

In Chapters 5 and 6, we noted that queries in patent search can be full patent applications and that the information need in such queries can be multi-faceted in nature, in the sense that a document which invalidates any of the individual claims is relevant for such queries. Moreover the retrieved documents which are expository in nature such as patent documents usually contain information on several topics, and hence a document in such a case can be classified into multiple categories. For example, a document titled "engine", may comprise of several broad level topics related to "motor", "transmission system", "gear box", "cooling" etc., and hence may be classified to each of these classes. We have seen that the multi-faceted topical relevance model (mTRLM), proposed in Chapter 6, is able to match the multiple facets of an information need to respective topics prevalent in the retrieved set of documents. The automatic retrieval method proposed in Chapter 6 aims to utilize topic distribution as an internal method in order to improve retrieval effectiveness. The aim of this chapter is to explore whether such topic distribution information can be disclosed to the users so as to help them recognize the latent aspects of a query themselves.

To this end, we develop a search interface where the system provides a visualization of the topic distribution in the retrieved set of documents including the query. This objective of the interface is to facilitate the information seeking task of a user through easier navigation across these topics. Providing visual cues to the relative proportion of topics in each document and the retrieved set as a whole can potentially help a user in discovering the latent aspects of the information need



Cluster **Transmissions** contains **15** documents.

JASPER **Engines** & **Transmissions**: Remanufactured **engine**: diesel, marine, import

... in a completely remanufactured JASPER gas or diesel **engine** , **transmission** or differential and get back on the ... installer near you. JASPER also provides remanufactured marine **engines** , stern drives , lower units and outboard powerheads ; Alternative ...
www.jasperengines.com/index.php - [cache] - Yippy Index

Rebuilt Used **Engines**, **Transmissions**, Transfer Cases, Quotes | Minneapolis, MN

... Instant Quote Privacy Policy Rebuilt, Remanufactured and Used **Engines** and **Transmissions** Let American Auto Distributing help get ... car running like new. Used, remanufactured or rebuilt **engines** and **transmissions** for all makes and models at ...
www.americanautodistributing.com/?noredirect - [cache] - Yippy Index

www.intequipllc.com - Construction Equipment Repair, Lawnmower, generator, Small **Engine** Repair in Baltimore

... Zama ZAMA Zama Shop By category 4-Cycle **Engine** Air Filters/Maintenance Kits Carburetion Parts Fuel Parts Ignition Parts & Voltage Regulators Internal **Engine** Parts Mufflers Oil and **Transmission** Filters Spark Plugs ...
https://www.intequipllc.com/parts - [cache] - Yippy Index

Used **Engines**, Used Car **Engine** and Used **Transmissions** for Sale

... Vermont Virginia Washington West Virginia Wisconsin Wyoming Used **Engines** and Used **Transmissions** At Auto Used **Engines** we ... 1-888-289-0076 The Totally Best Used **Engines**, Used

Figure 7.1: Result-list presentation by the search engine Clusty.

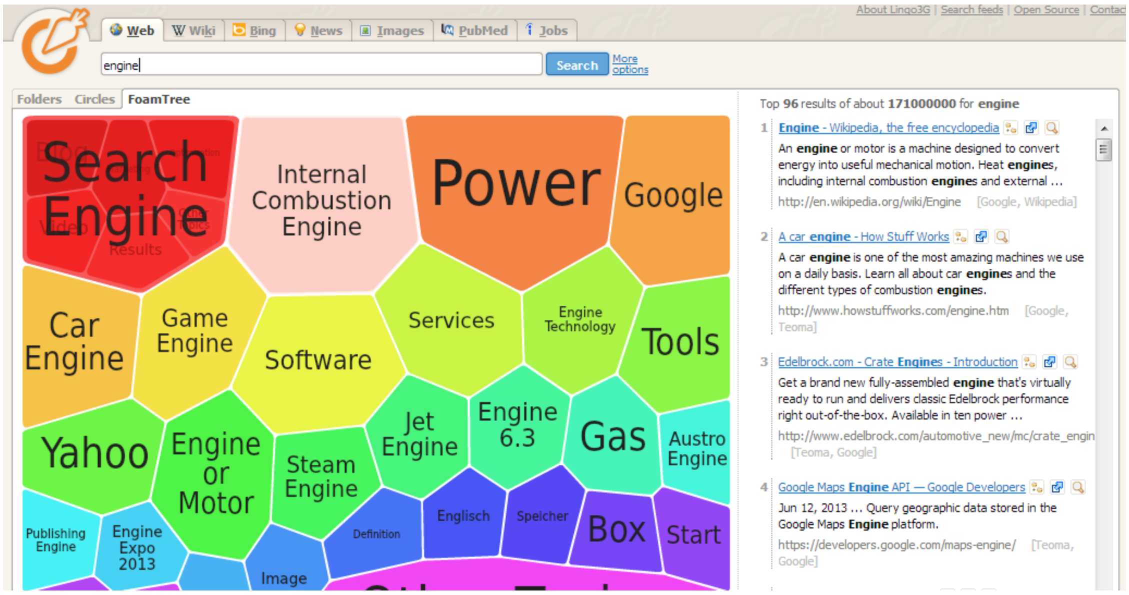
expressed in the query. Visualization of long queries such as the patent claims can also be helpful in matching relevant topics in the queries to those in the documents.

Some existing search interfaces, such as *Clusty*[1] and *Carrot*[2], provide a clustered view of the ranked list of documents. These systems simply categorize each documents in the ranked list into a topic cluster. This categorization of the retrieved set of documents into groups of topics aims to provide a more organized information access to the searcher in comparison to standard web search engines. Figure 7.1 shows a screen-shot for the results retrieved in response to the query "engine" by the search engine Clusty. It can be seen from Figure 7.1 that the retrieved documents are categorized into various clusters labelled as "Air craft", "Search engines" etc. A category can have sub-categories, as evident from the sub-categories "Jasper engines and transmissions", "Equipment generator" within the selected category "Transmissions".

Figure 7.2 demonstrates the presentation of search results by the search engine Carrot on the same query, namely "engine". The category labels in Carrot for this

[1] `http://clusty.com/`

[2] `http://search.carrotsearch.com/carrot2-webapp/search`



Top **96** results of about **171000000** for **engine**

1 **Engine** - Wikipedia, the free encyclopedia
An **engine** or motor is a machine designed to convert energy into useful mechanical motion. Heat **engines**, including internal combustion **engines** and external ...
http://en.wikipedia.org/wiki/Engine [Google, Wikipedia]

2 A car **engine** - How Stuff Works
A car **engine** is one of the most amazing machines we use on a daily basis. Learn all about car **engines** and the different types of combustion **engines**.
http://www.howstuffworks.com/engine.htm [Google, Teoma]

3 Edelbrock.com - Crate **Engines** - Introduction
Get a brand new fully-assembled **engine** that's virtually ready to run and delivers classic Edelbrock performance right out-of-the-box. Available in ten power ...
http://www.edelbrock.com/automotive_new/mc/crate_engin [Teoma, Google]

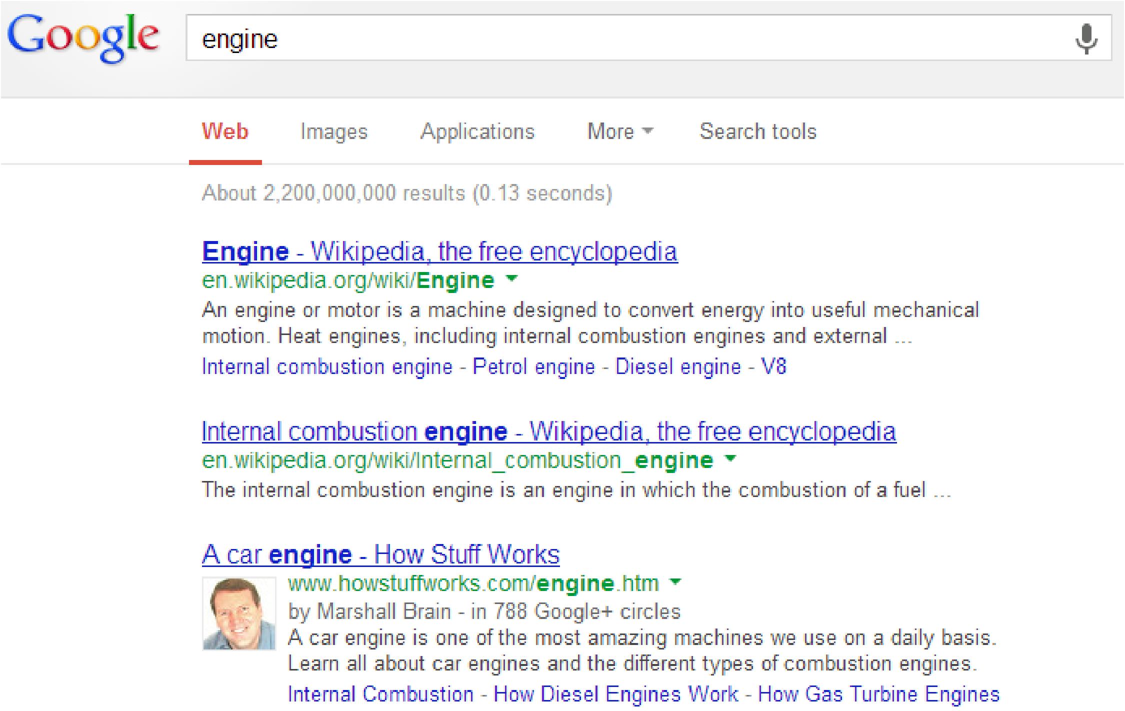
4 Google Maps **Engine** API — Google Developers
Jun 12, 2013 ... Query geographic data stored in the Google Maps **Engine** platform.
https://developers.google.com/maps-engine/ [Teoma, Google]

Figure 7.2: Result-list presentation by the search engine Carrot.

query are “Search Engine”, “Internal Combustion Engine” etc.

Both these systems rely on clustering the set of retrieved documents. However, limitations of the cluster hypothesis are firstly that each document can only belong to a single cluster (in this case a topic class), and secondly that clusters are mutually exclusive. Our developed interface addresses these limitations by modelling each document as a mixture of topics which ensures that a document can belong to multiple topic classes with proportional membership values (since it is a mixture model of topics with relative proportions).

Moreover, these interfaces do not provide any visualization of the retrieved documents themselves. These systems display the standard ranked list of documents along the right half of the screen as shown in Figures 7.1 and 7.2. The visual representation of the documents may provide valuable cues about the document content. In these search systems, it is not possible to know the topical composition of a document without reading it. For example, the top ranked document shown in the retrieved ranked list of Figure 7.2 is the Wikipedia article titled “Engine”. This document comprises of several sub-topics such as “heat engine”, “automobiles”, “electric motor” etc. It is not possible to know this composition simply by looking at the title and the snippet shown on the right half of the retrieved results page (see



engine

About 2,200,000,000 results (0.13 seconds)

Engine - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/Engine
An engine or motor is a machine designed to convert energy into useful mechanical motion. Heat engines, including internal combustion engines and external ...
Internal combustion engine - Petrol engine - Diesel engine - V8

Internal combustion engine - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/Internal_combustion_engine
The internal combustion engine is an engine in which the combustion of a fuel ...

A car engine - How Stuff Works
www.howstuffworks.com/engine.htm
by Marshall Brain - in 788 Google+ circles
A car engine is one of the most amazing machines we use on a daily basis. Learn all about car engines and the different types of combustion engines.
Internal Combustion - How Diesel Engines Work - How Gas Turbine Engines

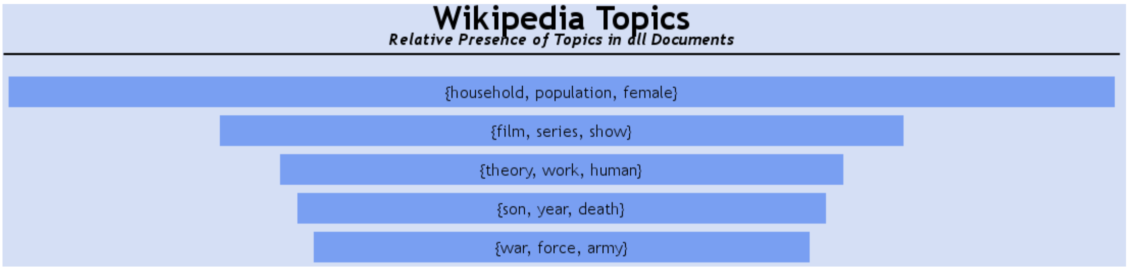
Figure 7.3: Result-list presentation by the Google.

Figure 7.2).

Furthermore, these search systems do not support quick navigation between topically related segments of documents. Navigation between topically related segments of documents is somewhat analogous to the feature of providing hyperlinks to other documents on topics related to the current one, as provided by major commercial search engines such as Google[3]. This is demonstrated in Figure 7.3, where we see the presence of hyperlinks to other related sub-topics of engine such as "internal combustion engine", "petrol engine" and "diesel engine". However, commercial search engines cannot provide hyperlinks to sections within documents where the relevant content is most likely to be found. Our developed interface provides direct hyperlinks from each document in the retrieved results page to sections within the document the contents of which are based on a selected topic. This advantage is that a user can quickly navigate between sections of a document the contents of which are based on his topic of ineterest.

Previous work of visualizing topic models involves application of LDA to cat-

[3] http://google.com



Wikipedia Topics

Relative Presence of Topics in all Documents

{household, population, female}

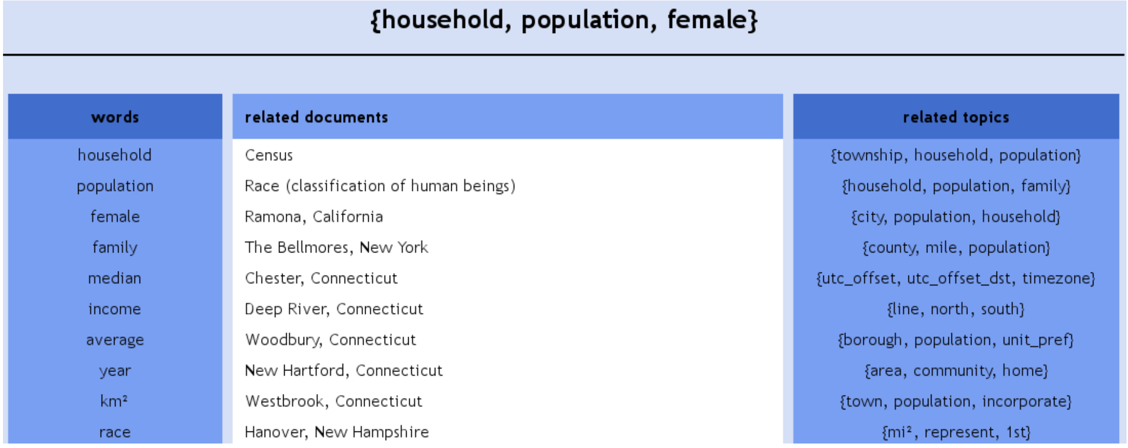
{film, series, show}

{theory, work, human}

{son, year, death}

{war, force, army}

Figure 7.4: Staring page of the LDA-based Wikipedia browser.



{household, population, female}

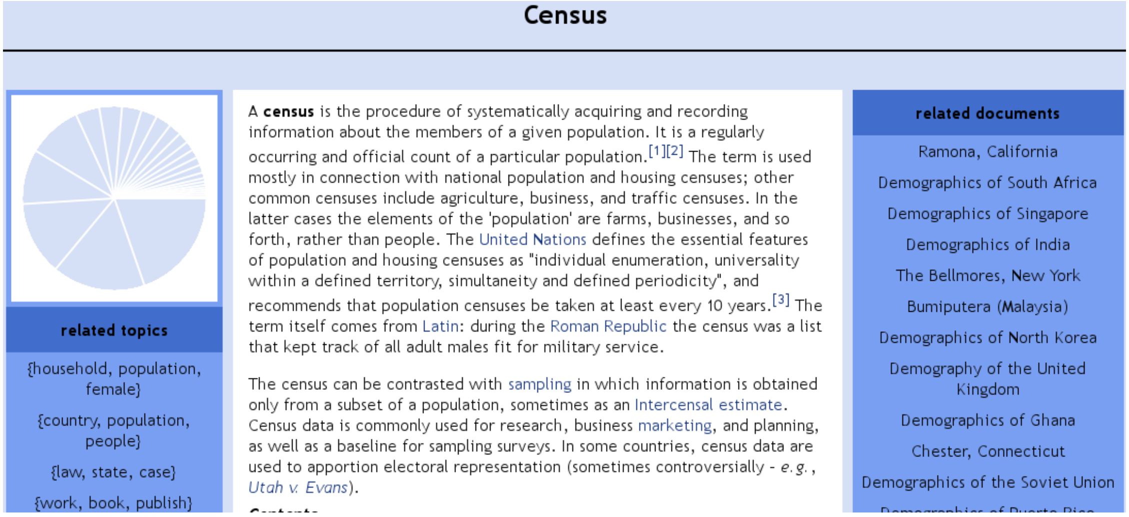
words	related documents	related topics
household	Census	{township, household, population}
population	Race (classification of human beings)	{household, population, family}
female	Ramona, California	{city, population, household}
family	The Bellmores, New York	{county, mile, population}
median	Chester, Connecticut	{utc_offset, utc_offset_dst, timezone}
income	Deep River, Connecticut	{line, north, south}
average	Woodbury, Connecticut	{borough, population, unit_pref}
year	New Hartford, Connecticut	{area, community, home}
km²	Westbrook, Connecticut	{town, population, incorporate}
race	Hanover, New Hampshire	{mi², represent, 1st}

Figure 7.5: The screen-shot after selecting a topic in the LDA-based Wikipedia browser.

egorize each Wikipedia document as a mixture of topics, and allowing navigation through documents related to a chosen topic (Chaney and Blei, 2012). Figure 7.4 shows the starting screen which displays the five most prevalent topics in the Wikipedia document collection. The user can click on any of these topic categories after which the system opens up a screen displaying the list of documents comprising of contents from the selected topic, as shown in Figure 7.5. Additionally, the system also shows a list of words belonging to the selected topic and a list of other topics related to the current one. The user can click on any of the documents listed for viewing them. The view of the first document on the list, namely "Census" is shown in Figure 7.6. Along with the document content, the system shows a list of other related documents and other topics related to the current selected topic.

In summary, the system described in (Chaney and Blei, 2012) allows a topic-based browsing of the Wikipedia corpus. However, some limitations of this system[4]

[4] http://bit.ly/wiki100



Census

related topics

{household, population, female}

{country, population, people}

{law, state, case}

{work, book, publish}

A **census** is the procedure of systematically acquiring and recording information about the members of a given population. It is a regularly occurring and official count of a particular population.[1][2] The term is used mostly in connection with national population and housing censuses; other common censuses include agriculture, business, and traffic censuses. In the latter cases the elements of the 'population' are farms, businesses, and so forth, rather than people. The United Nations defines the essential features of population and housing censuses as "individual enumeration, universality within a defined territory, simultaneity and defined periodicity", and recommends that population censuses be taken at least every 10 years.[3] The term itself comes from Latin: during the Roman Republic the census was a list that kept track of all adult males fit for military service.

The census can be contrasted with sampling in which information is obtained only from a subset of a population, sometimes as an Intercensal estimate. Census data is commonly used for research, business marketing, and planning, as well as a baseline for sampling surveys. In some countries, census data are used to apportion electoral representation (sometimes controversially - *e.g.*, *Utah v. Evans*).

Contents

related documents

Ramona, California

Demographics of South Africa

Demographics of Singapore

Demographics of India

The Bellmores, New York

Bumiputera (Malaysia)

Demographics of North Korea

Demography of the United Kingdom

Demographics of Ghana

Chester, Connecticut

Demographics of the Soviet Union

Demographics of Puerto Rico

Figure 7.6: View of a Wikipedia document in the LDA-based Wikipedia browser.

are that it has no provision for:

- a query-based information search,
- a visualization of the topics within a document, and
- topic-based navigation through sections of documents.

Our developed search interface addresses each of these limitations. The details of our search interface, which we call *TopicVis* are described in the following sections.

7.2 System Overview

In this section, we provide a brief introduction to the features supported by our search interface, named *TopicVis*. The details of each feature is described in the next section. Features supported by TopicVis are listed as follows.

1. The interface shows a visualization of the topics in the retrieved set of documents, which may relate to aspects of the information need expressed in the query. The objective here is to provide a visualization of the fine grained aspects or facets of the information need. The system here makes use of the cluster hypothesis that various aspects in the information need give rise to the different topics in the retrieved set of documents (Xu and Croft, 1996).

Each topic is labelled by a list of 10 words with the top most membership likelihood values in that topical class. This approach of labelling is thus identical to the one adopted in (Chaney and Blei, 2012) as shown in Figure **??**, the only difference being Chaney and Blei (2012) used 3 top most words for each topic, whereas we use 10.

2. For multi-topical long queries, such as in patent search, the interface also shows the topical structure of the query so that relevant sections from documents may potentially be visually matched with those in the query by a user.

3. The interface shows the relative proportion of topics in each retrieved document, providing the user with a visual cue about the topical composition of a document. The intention here is to make the users avoid reading non-relevant documents. Ideally, the users can glance quickly over documents non-relevant to his topic of interest by simply looking at their relative topical compositions.

4. The system allows the user to select a particular topic for feedback by clicking on a particular coloured region. The objective of this feature is to rerank the initially retrieved set of documents in order to ensure that documents consisting of a high proportion of words from the selected topic are reported at top ranks after reranking.

5. The interface facilitates the user firstly to jump to the first section of a document the content of which is predominated by the selected topic and secondly to navigate between the sections on the selected topic. For example, returning back to the example query of “engine”, if the topic selected by the user is related to the concept of “gear box”, on clicking the corresponding topical region of a document thumbnail image, the user is taken directly to the first section of the document which contains information on “gear box”. Each section is accompanied with a next and a previous button, which can be used to jump directly to the next or the previous section of the document containing

information on the same topic. In our example, the user by pressing the next button can view the next section within the current document containing information on “gear box”. This helps the user in quickly locating information on a chosen topic of interest.

The first two features of the system are achieved by utilizing the output of the LDA step in TRLM. Recall from Section 6.2.3 that in TRLM, we compute the probabilities $P(w|D)$s for each of the top R pseudo-relevant documents by marginalizing them over a set of latent topics, and that LDA outputs two distribution vectors θ (from document to topic) and ϕ (from topic to word). These output matrices θ and ϕ from the LDA step of the TRLM are then used to provide a visualization of the topic distribution within the R top-ranked documents. More specifically, the i^{th} row vector of the θ matrix, which represents the relative proportions of topics for the i^{th} document in the retrieved list, is used to visualize the topical composition of the i^{th} document.

The third feature of topic-based feedback is achieved by simply rearranging the retrieved documents, decreasingly sorted on the θ values for the selected topic. More precisely speaking, the document with the highest likelihoood for the chosen topic in its mixture model, is reported at the top rank and so on. The objective of this feature is to rerank the initially retrieved set of documents in order to ensure that documents predominated by the chosen topic of interest by a user are brought at top ranks after reranking. This is somewhat similar to filtering the resultlist of retrieved documents based on the chosen topic of interest, as featured in Clusty and Carrot. For example, selecting the topic “Transmissions”, as shown in Figure 7.1, shows only the documents belonging to that category. A document can thus appear only once in a topic category. The TopicVis, in contrast, takes into account the topical composition of a document while reranking the results, as a result of which a document can appear at different poositions in the ranked list depending on the selected topic of interest. To illusrtate with an example, a document titled “gear box” can appear at the first rank when the selected topic is “gear box”, while the same document may appear at position 10 if the selected topic is "motor transmission". Note that a document on "gear box" can also contain information on "motor transmission". It is thus more reasonable to report this document somewhere down the ranked list when the chosen topic of interest is "motor transmission", rather than not showing it at all for this selected topic as done in Clusty and Carrot, due to the inherent weakness of hard clustering in assuming that a document can only be comprised of a single topic and that these topic classes are mutually exclusive.

The last feature of topic-based navigation involves categorizing sections of documents to topic classes based on the proportion of constituent words in each topic, and then building up navigation links within sections on the same topic. This facilitates to jump from one part of the document to the other without needing to vertically scroll downwards. The ϕ matrix from the LDA output is used to classify each section of a document into one of the topical classes. Segments classified to the identical topic classes are then linked together by the navigtion arrows.

Although the TopicVis interface is quite general in its approach and thus can be applied for any interactive ad-hoc search task, we demonstrate and evaluate the function of the TopicVis search interface on the CLEF-IP 2010 document collection (see Section 3.2.1 for more details on the dataset). The reason we chose this dataset in particular are as follows. Recall that both the patent documents and queries for the prior art search task are multi-topical in nature (see Section 5.1 for more discussion on this), as a result of which, it may be more difficult and time consuming for a user to invalidate certain invention claims of the query by finding relevant prior art contents in segments of documents related to those claims, or in other words, to match sections of documents relevant to those in the query. We expect that the task of a user in this case can considerably be simplified by the use of such an interface which provides the topic visualizations of the documents and the query. Consequently, we demonstrate and evaluate our interface on this particular search task.

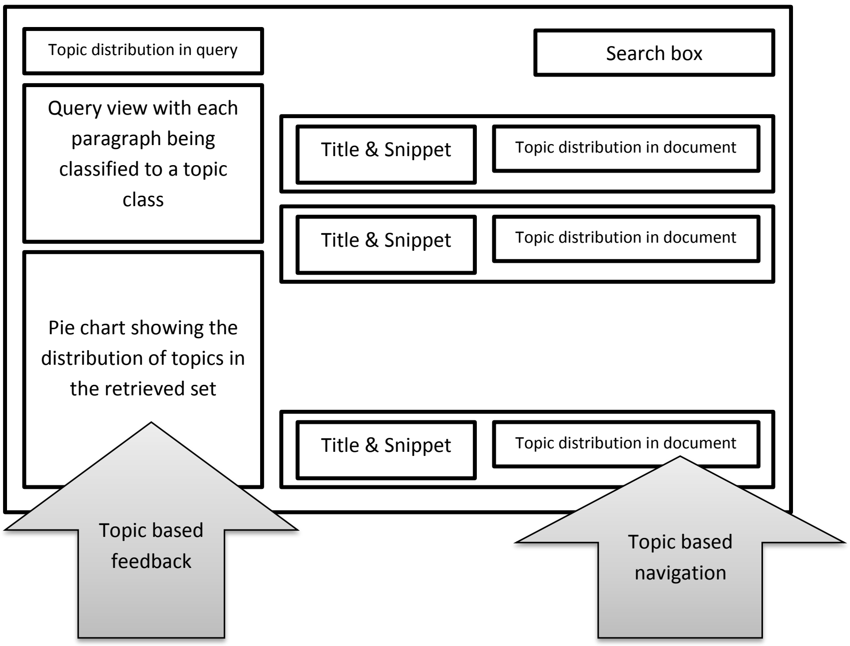
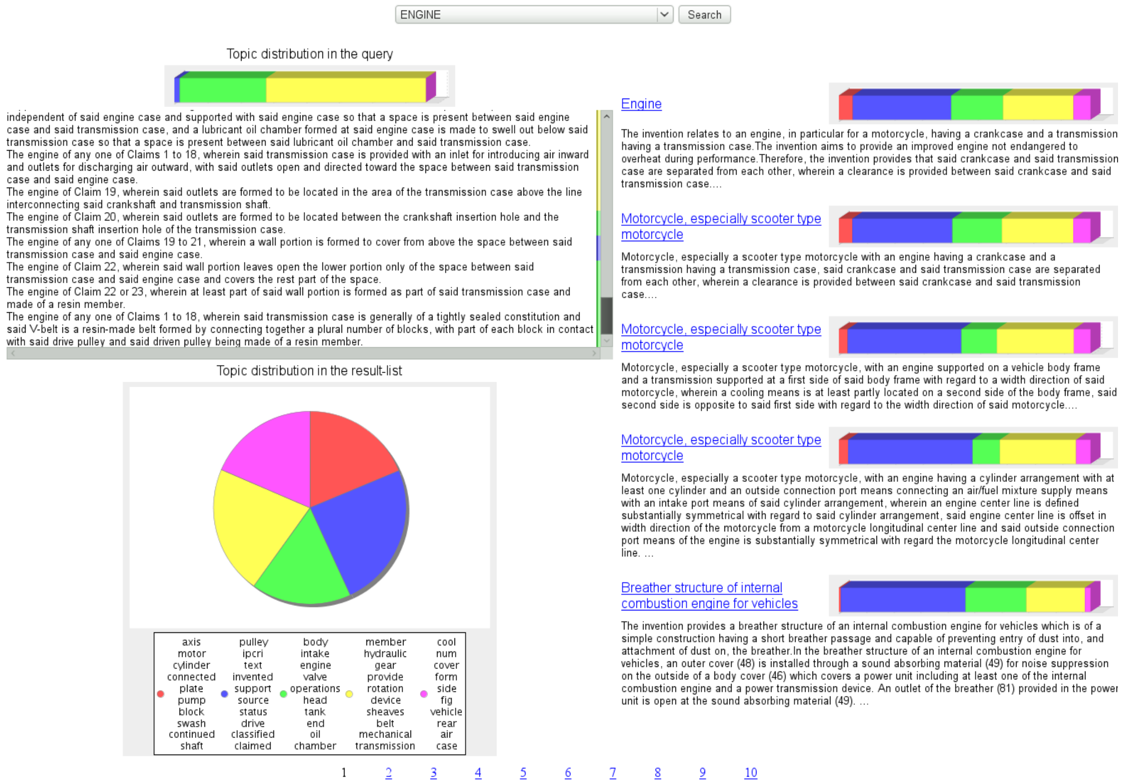

Topic distribution in query
Search box
Query view with each paragraph being classified to a topic class
Title & Snippet
Topic distribution in document
Title & Snippet
Topic distribution in document
Pie chart showing the distribution of topics in the retrieved set
Title & Snippet
Topic distribution in document
Topic based feedback
Topic based navigation

Figure 7.7: Block diagram view of the TopicVis search interface.

A block diagram view of the TopicVis interface is shown in Figure 7.7. On receiving a new patent query from a user, the system executes mTRLM (cf. Section 6.2.3). Recall from Section 6.3.2 that we obtained the best IR effectiveness on an average with 5 topics. The value of the mTRLM parameter K, the number of topics, was thus preconfigured to 5 in the system. The ranked list of results returned by mTRLM is shown on the right part of the screen. The interface shows the title and a snippet for each retrieved document similar to a standard search engine. In addition, the interface renders the topic distribution in the query-document along the left part of the screen. Below it, the interface shows the query text with the paragraphs coloured with appropriate topic classes to which they belong. On the botom-left corner, a pie chart is displayed which shows the distribution of topics in the retrieved set of documents. The pie chart is accompanied by a list of 10 top-most probable words (i.e. top 10 words with the most membership likelihood values) for each topic. Furthermore, in addition to the title and snippet, the system shows a stacked horizontal histogram (also known as a stacked bar chart) in order to render the topic distribution for each retrieved document.



ENGINE Search

Topic distribution in the query

independent of said engine case and supported with said engine case so that a space is present between said engine case and said transmission case, and a lubricant oil chamber formed at said engine case is made to swell out below said transmission case so that a space is present between said lubricant oil chamber and said transmission case.
The engine of any one of Claims 1 to 18, wherein said transmission case is provided with an inlet for introducing air inward and outlets for discharging air outward, with said outlets open and directed toward the space between said transmission case and said engine case.
The engine of Claim 19, wherein said outlets are formed to be located in the area of the transmission case above the line interconnecting said crankshaft and transmission shaft.
The engine of Claim 20, wherein said outlets are formed to be located between the crankshaft insertion hole and the transmission shaft insertion hole of the transmission case.
The engine of any one of Claims 19 to 21, wherein a wall portion is formed to cover from above the space between said transmission case and said engine case.
The engine of Claim 22, wherein said wall portion leaves open the lower portion only of the space between said transmission case and said engine case and covers the rest part of the space.
The engine of Claim 22 or 23, wherein at least part of said wall portion is formed as part of said transmission case and made of a resin member.
The engine of any one of Claims 1 to 18, wherein said transmission case is generally of a tightly sealed constitution and said V-belt is a resin-made belt formed by connecting together a plural number of blocks, with part of each block in contact with said drive pulley and said driven pulley being made of a resin member.

Topic distribution in the result-list

🔴	🔵	🟢	🟡	🟣
axis	pulley	body	member	cool
motor	ipcri	intake	hydraulic	num
cylinder	text	engine	gear	cover
connected	invented	valve	provide	form
plate	support	operations	rotation	side
pump	source	head	device	fig
block	status	tank	sheaves	vehicle
swash	drive	end	belt	rear
continued	classified	oil	mechanical	air
shaft	claimed	chamber	transmission	case

Engine

The invention relates to an engine, in particular for a motorcycle, having a crankcase and a transmission having a transmission case.The invention aims to provide an improved engine not endangered to overheat during performance.Therefore, the invention provides that said crankcase and said transmission case are separated from each other, wherein a clearance is provided between said crankcase and said transmission case....

Motorcycle, especially scooter type motorcycle

Motorcycle, especially a scooter type motorcycle with an engine having a crankcase and a transmission having a transmission case, said crankcase and said transmission case are separated from each other, wherein a clearance is provided between said crankcase and said transmission case....

Motorcycle, especially scooter type motorcycle

Motorcycle, especially a scooter type motorcycle, with an engine supported on a vehicle body frame and a transmission supported at a first side of said body frame with regard to a width direction of said motorcycle, wherein a cooling means is at least partly located on a second side of the body frame, said second side is opposite to said first side with regard to the width direction of said motorcycle....

Motorcycle, especially scooter type motorcycle

Motorcycle, especially a scooter type motorcycle, with an engine having a cylinder arrangement with at least one cylinder and an outside connection port means connecting an air/fuel mixture supply means with an intake port means of said cylinder arrangement, wherein an engine center line is defined substantially symmetrical with regard to said cylinder arrangement, said engine center line is offset in width direction of the motorcycle from a motorcycle longitudinal center line and said outside connection port means of the engine is substantially symmetrical with regard the motorcycle longitudinal center line. ...

Breather structure of internal combustion engine for vehicles

The invention provides a breather structure of an internal combustion engine for vehicles which is of a simple construction having a short breather passage and capable of preventing entry of dust into, and attachment of dust on, the breather.In the breather structure of an internal combustion engine for vehicles, an outer cover (48) is installed through a sound absorbing material (49) for noise suppression on the outside of a body cover (46) which covers a power unit including at least one of the internal combustion engine and a power transmission device. An outlet of the breather (81) provided in the power unit is open at the sound absorbing material (49). ...

1 2 3 4 5 6 7 8 9 10

Figure 7.8: Output of TopicVis for the query "Engine".

A screen-shot for a sample query, titled "Engine", is shown in Figure 7.8. Note that the query is a full patent invention claim, and not just the keyword "engine". The full text of the query is shown at the top left corner of the screen.

With reference to Figure 7.8, it can be seen that the sample query is quite general in nature, with fine grained aspects such as the "motor" (the red region[5] in the pie chart), "valve operations" (the green region), "gear box operations" (the yellow region) etc. By looking at the pie chart and the list of words belonging to each topic, a user can potentially map each topic to an individual aspect of their potential information need. In the list of words displayed below the pie chart, it can be seen that the red coloured region containing words such as "motor", "axis", "cylinder", refer to the topic of combustion in an engine, whereas the yellow coloured topic containing words such as "hydraulic", "gear", "transmission" etc. which roughly

[5] If you are reading this in monochrome, then the colour convention is as follows. For the pie chart, the colours in clock-wise direction are red, blue, green, yellow and magenta respectively. The colours in a stacked bar chart are in the same order from left to right. We however keep on refering to the regions of the pie and the stacked bar charts in colour codes throughout the rest of this thesis.

corresponds to the topic of the gear transmission system of an engine.

The interface shows the topic distribution of the patent query “engine” on top left part of the screen; below it, it shows the text of the query document. Each paragraph of the text is annotated with a coloured bar on its right. The colour represents the topical class assigned to the current paragraph. In Section 7.3.4, we describe how this classification is performed.

We now describe how this interface can be useful to a patent examiner. Since each claim of the query patent is assigned a colour (a topical class), the examiner can look for segments in the retrieved documents which belong to the same topical class, i.e. they are assigned the same colour and the examiner can thus use this information to assist in validating or invaidating this claim. To this end, the system supports topic-based navigation in the following way.

1. The regions in the stacked bar charts for the retrieved documents are clickable and are linked to the first segment of the document marked with the corresponding colour.

2. Each segment in a document, classified to a topic class and hence assigned the corresponding topic colour, contains a next and a previous link respectively to the next and the previous segments of the same colour, i.e. belonging to the same topical class, within the same document. Thus, if a patent examiner wants to invalidate the claim part of the query patent marked with yellow, he can click on the yellow region in the stacked bar chart of a document and continue to view yellow segments by simply pressing on the next links without needing to scroll through the document. This interaction is shown by the arrow labelled “Topic based navigation” in Figure 7.8.

3. On clicking a region of the pie chart, the system reranks the retrieved set of documents ordered by the proportion of the selected topic. The objective of this feature is to rerank the initially retrieved set of documents in order to ensure that documents consisting of a high proportion of words from the

selected topic are reported at top ranks after reranking. Returning back to our example, for invalidating the yellow claim, a patent examiner may find it useful to rerank the documents, so that the document which contains the highest proportion of yellow text is brought at the top rank, followed by the ones with a progressively decreasing proportion of the text marked with yellow.

The following sections describe the details of the features supported by TopicVis.

7.3 TopicVis Features

In this section, we describe the features provided by the TopicVis system in more details with illustrative examples and sample output screen-shots.

7.3.1 Topic Distribution in the Retrieved Set

To illustrate how the pie chart representing the distribution of topics over the retrieved set of documents is rendered, let us consider a simple example of LDA output on 5 documents and 3 topics. Recall from Section 2.2.1, that θ is an $M \times K$ matrix and ϕ is a $K \times V$ matrix, where M represents the number of documents retrieved at top ranks, K the number of topics, and V the vocabulary size. Thus in the case of our example, $M = 5$ and $K = 3$. Let θ be the following matrix.

$$\theta = \begin{array}{c} \\ D_1 \\ D_2 \\ D_3 \\ D_4 \\ D_5 \end{array} \begin{array}{c} \begin{array}{ccc} T_1 & T_2 & T_3 \end{array} \\ \begin{pmatrix} 0.2 & 0.2 & 0.6 \\ 0.6 & 0.2 & 0.2 \\ 0.3 & 0.3 & 0.4 \\ 0.1 & 0.1 & 0.8 \\ 0.97 & 0.01 & 0.02 \end{pmatrix} \end{array} \tag{7.1}$$

The first row of the matrix reveals that the first document (D_1) is composed of 20% terms from topic 1 (T_1), 20% terms from topic 2 (T_2), and 60% from topic 3 (T_3).

The other rows can be similarly interpreted. Note that if the content of a document is not comprised of a certain subset of topics, as may often be the case, then the corresponding column values for those topics in the row vector for that document would be close to zero. For example, the last row of the θ matrix indicates that D_5 is essentially uni-topical with negligible contributions from T_2 and T_3.

Assuming $V = 10$, let the ϕ matrix be

$$\phi = \begin{array}{c|cccccccccc|} & w_1 & w_2 & w_3 & w_4 & w_5 & w_6 & w_7 & w_8 & w_9 & w_{10} \\ T_1 & 0.01 & 0.05 & 0.10 & 0.75 & 0.30 & 0.30 & 0.01 & 0.05 & 0.40 & 0.05 \\ T_2 & 0.10 & 0.05 & 0.80 & 0.20 & 0.30 & 0.40 & 0.98 & 0.05 & 0.10 & 0.90 \\ T_3 & 0.89 & 0.90 & 0.10 & 0.05 & 0.40 & 0.30 & 0.01 & 0.90 & 0.50 & 0.05 \end{array} \tag{7.2}$$

The ϕ matrix is interpreted as follows. Each column vector of the ϕ matrix, pertaining to a particular word in the set of documents, represents the topic class membership likelihoods. More precisely speaking, the first column says that the word w_1 belongs to the first topic T_1 with a probability of 0.01, to the second topic T_2 with a probability of 0.1, and that to the third with the highest likelihood of 0.89.

The pie chart, rendered on the left pane of the TopicVis screen, is computed as follows. From the fuzzy or soft memberships of a word into the topic classes of the ϕ matrix, we compute the hard membership values by taking the max operator, resolviing ties randomly. Thus, with reference to our example, w_1 is assigned to T_3 and so on. The assignments, obtained after taking the max operator are shown below.

$$\begin{array}{c|cccccccccc|} & w_1 & w_2 & w_3 & w_4 & w_5 & w_6 & w_7 & w_8 & w_9 & w_{10} \\ T_1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ T_2 & 0 & 0 & 1 & 0 & 0 & 1 & 1 & 0 & 0 & 1 \\ T_3 & 1 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 1 & 0 \end{array} \tag{7.3}$$

The frequency of a topic T, denoted by $f(T)$, is then computed as the total number of words belonging to that topic, i.e. the number of 1s in its corresponding

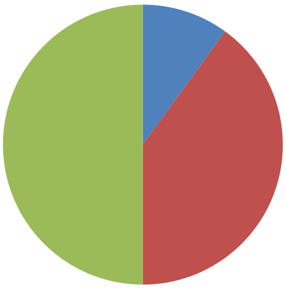


Figure 7.9: Pie-chart, derived from the LDA output matrix ϕ, showing distribution of topics in the retrieved set.

row. For example, $f(T_2) = 4$. The ratio of this frequency, $f(T)$, to the total number of terms, i.e. $\frac{f(T)}{V}$ for each term, is then rendered in the pie chart. For our example, $f(T_1) = \frac{1}{10}$, $f(T_2) = \frac{4}{10}$, and $f(T_3) = \frac{5}{10}$. The pie chart with these values is shown in Figure 7.9.

The pie chart visually represents the contribution of each topic in the retrieved set of documents, and as per the cluster hypothesis (see (Xu and Croft, 1996) and Section 6.2), provides a visual cue about the more specific aspects of a given query.

7.3.2 Topic Distribution in a Single Document

Alongside the title and snippet of a document, TopicVis shows the distribution of topics for this document. The snippet in the TopicVis comprises of the first 500 characters of the text in the page, Note that the snippet is generated by a relatively simple method, since generating complex informative summaries is outside the scope of this work. The purpose of the snippet and the rendering of the word-topic mapping is to convey as much information as possible to the searcher, without him actually needing to open the document by following the hyperlink. The visual representation however is more appealing, in the sense that the user can form an idea of what the document is about without actually needing to read the snippet.

Let us take the example θ matrix of the previous section, and illustrate how this is achieved. The first document in the ranked list pertains to the document D_1, i.e. the first row of the θ matrix of Equation 7.1. For this document, the interface shows a stacked horizontal histogram (stacked bar chart), with adjacent regions of three

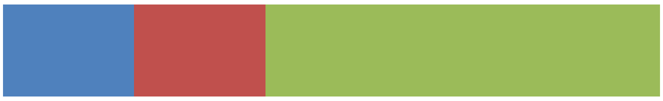


Figure 7.10: Stacked histogram, derived from the LDA output matrix θ, showing distribution of topics in a document.

colours each corresponding to one topic, as shown in Figure 7.10. From Figure 7.10, we see that the first coloured region occupies 20% of the total area, the second again 20%, and the third 60%.

7.3.3 Topic-based Feedback

In topic-based feedback, we rerank documents based on their topical compositions. The objective is to ensure that documents consisting of a high proportion of words from the selected topic are reported at top ranks. The user can use this feature by clicking on the corresponding regions of the pie-chart associated with different topics.

The pie chart is displayed on the left pane of the interface, as shown in Figure 7.7. The regions of the pie-chart are clickable. A click event in a particular region reranks the result-list of retrieved documents based on the selected topic. Since each document is a mixture model of its constituent topics, it can be considered as a vector, with the proportion of each topic being a component of the vector. The document vectors are hence sorted by the component value corresponding to the selected topic.

Returning to our example ranked list shown in matrix form in Equation 7.1, and considering the user clicks on the area pertaining to topic T_1 of the pie chart, the ranked list is rearranged as shown in Equation 7.4. Note that each document vector has been sorted on the first component, which in turn represents the relative contribution from topic T_1 to obtain a row rearranged matrix θ_1 from θ.

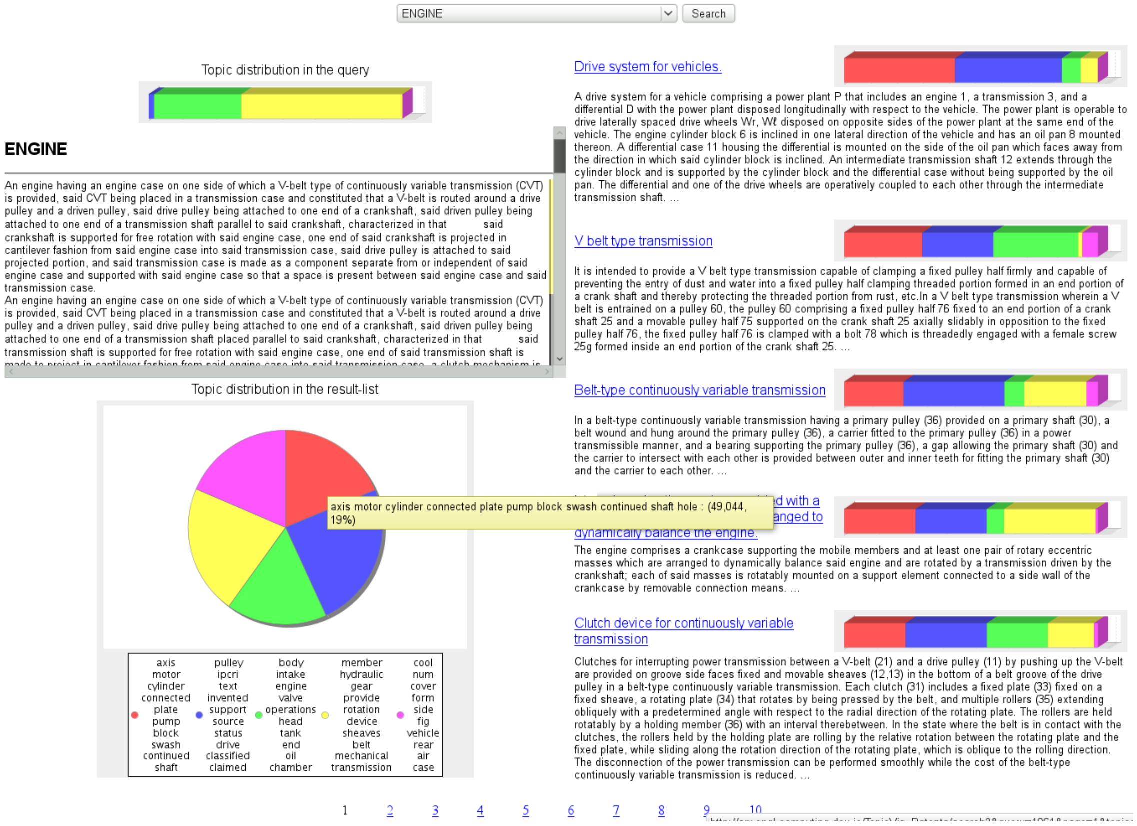
$$\theta_1 = \begin{array}{c|ccc|} & T_1 & T_2 & T_3 \\ D_5 & 0.8 & 0.1 & 0.1 \\ D_2 & 0.6 & 0.2 & 0.2 \\ D_3 & 0.3 & 0.3 & 0.4 \\ D_1 & 0.2 & 0.2 & 0.6 \\ D_4 & 0.1 & 0.1 & 0.8 \end{array} \tag{7.4}$$

The effect of topic-based feedback on the ranked list returned for the query "Engine" (cf. Section 7.2) is shown in Figure 7.11. The selected topic here is T_1, i.e. the topic representing the concept of motor transmission (the red region of the pie chart). It can be seen that documents about motor transmission with titles such as "Drive system for vehicles", "V belt type transmission" etc. are shown at top ranks, as shown in Figure 7.11.

Figure 7.12 shows the case where the selected topic for feedback is T_5 (the magenta coloured region). This topic broadly relates to cooling of the engine. We observe that in this case documents predominantly containing material on this topic with titles such as "starter/generator for motor vehicles", "engine lubricating device" etc. are reported in the top 5 ranks after reranking, as shown in Figure 7.12.

It is worth mentioning that this system has some similarities to faceted search, in which users can explore a collection from mutually exlusive categories of information, such as news, games, movies etc[6]. A major difference between our approach and faceted search is that in the latter, a document exclusively belongs to a single category, whereas in our case, a document is treated as a mixture model of topics. Furthermore, the topics in a faceted search capture the coarse-grained categorical information of a document, but not the fine-grained aspects of the information need

[6] Some commercial faceted search engines are Open Directory Project `http://www.dmoz.org/` and Yahoo Directory `http://dir.yahoo.com/`.



ENGINE
Search
Topic distribution in the query
Topic distribution in the result-list
axis motor cylinder connected plate pump block swash continued shaft
pulley ipcri text invented support source status drive classified claimed
body intake engine valve operations head tank end oil chamber
member hydraulic gear provide rotation device sheaves belt mechanical transmission
cool num cover form side fig vehicle rear air case
1 2 3 4 5 6 7 8 9 10

ENGINE

An engine having an engine case on one side of which a V-belt type of continuously variable transmission (CVT) is provided, said CVT being placed in a transmission case and constituted that a V-belt is routed around a drive pulley and a driven pulley, said drive pulley being attached to one end of a crankshaft, said driven pulley being attached to one end of a transmission shaft parallel to said crankshaft, characterized in that said crankshaft is supported for free rotation with said engine case, one end of said crankshaft is projected in cantilever fashion from said engine case into said transmission case, said drive pulley is attached to said projected portion, and said transmission case is made as a component separate from or independent of said engine case and supported with said engine case so that a space is present between said engine case and said transmission case.

An engine having an engine case on one side of which a V-belt type of continuously variable transmission (CVT) is provided, said CVT being placed in a transmission case and constituted that a V-belt is routed around a drive pulley and a driven pulley, said drive pulley being attached to one end of a crankshaft, said driven pulley being attached to one end of a transmission shaft placed parallel to said crankshaft, characterized in that said transmission shaft is supported for free rotation with said engine case, one end of said transmission shaft is made to project in cantilever fashion from said engine case into said transmission case, a clutch mechanism is

axis motor cylinder connected plate pump block swash continued shaft hole : (49,044, 19%)

Drive system for vehicles.

A drive system for a vehicle comprising a power plant P that includes an engine 1, a transmission 3, and a differential D with the power plant disposed longitudinally with respect to the vehicle. The power plant is operable to drive laterally spaced drive wheels Wr, Wℓ disposed on opposite sides of the power plant at the same end of the vehicle. The engine cylinder block 6 is inclined in one lateral direction of the vehicle and has an oil pan 8 mounted thereon. A differential case 11 housing the differential is mounted on the side of the oil pan which faces away from the direction in which said cylinder block is inclined. An intermediate transmission shaft 12 extends through the cylinder block and is supported by the cylinder block and the differential case without being supported by the oil pan. The differential and one of the drive wheels are operatively coupled to each other through the intermediate transmission shaft. ...

V belt type transmission

It is intended to provide a V belt type transmission capable of clamping a fixed pulley half firmly and capable of preventing the entry of dust and water into a fixed pulley half clamping threaded portion formed in an end portion of a crank shaft and thereby protecting the threaded portion from rust, etc.In a V belt type transmission wherein a V belt is entrained on a pulley 60, the pulley 60 comprising a fixed pulley half 76 fixed to an end portion of a crank shaft 25 and a movable pulley half 75 supported on the crank shaft 25 axially slidably in opposition to the fixed pulley half 76, the fixed pulley half 76 is clamped with a bolt 78 which is threadedly engaged with a female screw 25g formed inside an end portion of the crank shaft 25. ...

Belt-type continuously variable transmission

In a belt-type continuously variable transmission having a primary pulley (36) provided on a primary shaft (30), a belt wound and hung around the primary pulley (36), a carrier fitted to the primary pulley (36) in a power transmissible manner, and a bearing supporting the primary pulley (36), a gap allowing the primary shaft (30) and the carrier to intersect with each other is provided between outer and inner teeth for fitting the primary shaft (30) and the carrier to each other. ...

d with a anged to dynamically balance the engine.

The engine comprises a crankcase supporting the mobile members and at least one pair of rotary eccentric masses which are arranged to dynamically balance said engine and are rotated by a transmission driven by the crankshaft; each of said masses is rotatably mounted on a support element connected to a side wall of the crankcase by removable connection means. ...

Clutch device for continuously variable transmission

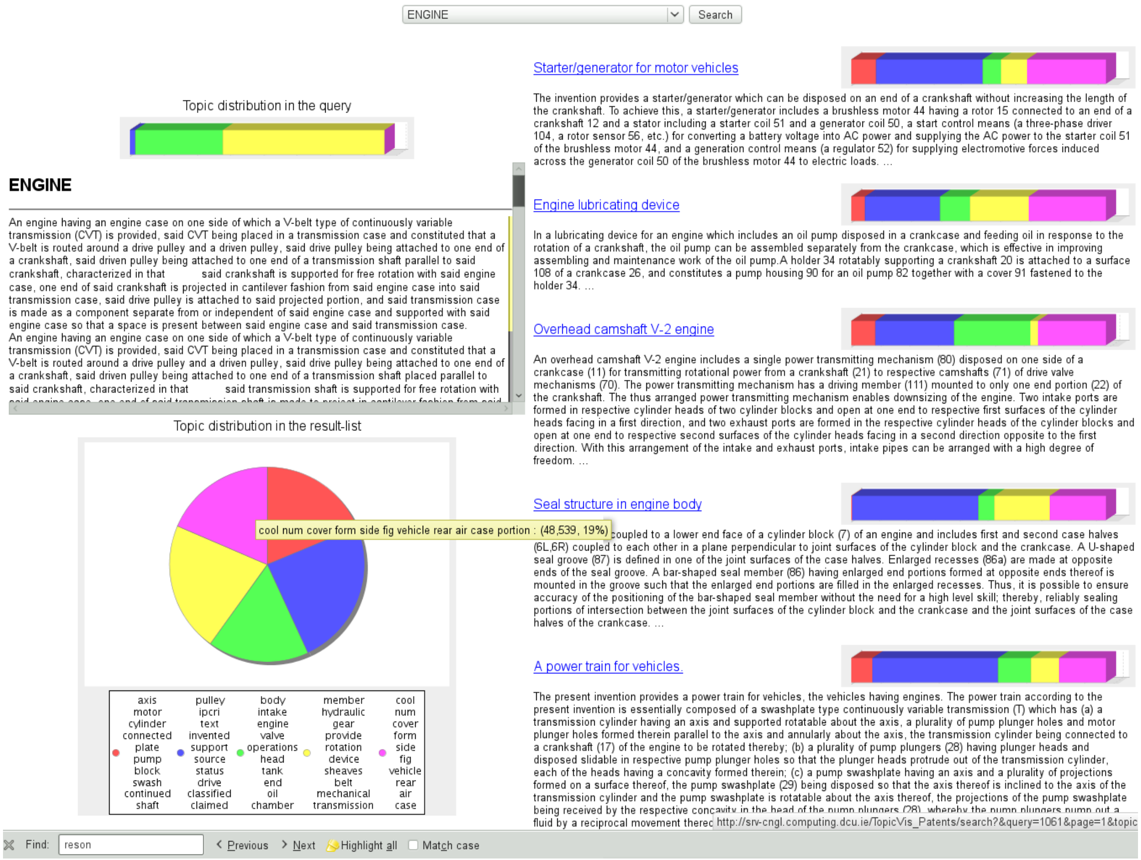
Clutches for interrupting power transmission between a V-belt (21) and a drive pulley (11) by pushing up the V-belt are provided on groove side faces fixed and movable sheaves (12,13) in the bottom of a belt groove of the drive pulley in a belt-type continuously variable transmission. Each clutch (31) includes a fixed plate (33) fixed on a fixed sheave, a rotating plate (34) that rotates by being pressed by the belt, and multiple rollers (35) extending obliquely with a predetermined angle with respect to the radial direction of the rotating plate. The rollers are held rotatably by a holding member (36) with an interval therebetween. In the state where the belt is in contact with the clutches, the rollers held by the holding plate are rolling by the relative rotation between the rotating plate and the fixed plate, while sliding along the rotation direction of the rotating plate, which is oblique to the rolling direction. The disconnection of the power transmission can be performed smoothly while the cost of the belt-type continuously variable transmission is reduced. ...

Figure 7.11: Topic-based Feedback on Topic 1, i.e. the topic related to the concept of "motor transmission".

sought. However, using our interface the user can visualize the more subtle aspects of a query, and hence refine his search accordingly. For example, a user may not have known the fine grained aspects such as "motor", "transmission system", "gear box", "cooling" etc. associated with the query "engine". A visualization of these concepts through the search interface is likely to help him in choosing a particular aspect for further exploration.

7.3.4 Topic-based Navigation

The title of a retrieved document is hyperlinked with the help of HTML anchor link to the standard text-based view of the corresponding patent article, as done in standard search engines. However, it is difficult for a user to locate the sought information from such long expository articles. A guided walk through the sections of a document related to a given topic should be beneficial for a user. For example let


ENGINE
Search
Topic distribution in the query
Topic distribution in the result-list
cool num cover form side fig vehicle rear air case portion : (48,539, 19%)
● axis motor cylinder connected plate pump block swash continued shaft
● pulley ipcri text invented support source status drive classified claimed
● body intake engine valve operations head tank end oil chamber
● member hydraulic gear provide rotation device sheaves belt mechanical transmission
● cool num cover form side fig vehicle rear air case
Find: reson
Previous
Next
Highlight all
□ Match case
http://srv-cngl.computing.dcu.ie/TopicVis_Patents/search?&query=1061&page=1&topic

ENGINE

An engine having an engine case on one side of which a V-belt type of continuously variable transmission (CVT) is provided, said CVT being placed in a transmission case and constituted that a V-belt is routed around a drive pulley and a driven pulley, said drive pulley being attached to one end of a crankshaft, said driven pulley being attached to one end of a transmission shaft parallel to said crankshaft, characterized in that said crankshaft is supported for free rotation with said engine case, one end of said crankshaft is projected in cantilever fashion from said engine case into said transmission case, said drive pulley is attached to said projected portion, and said transmission case is made as a component separate from or independent of said engine case and supported with said engine case so that a space is present between said engine case and said transmission case.
An engine having an engine case on one side of which a V-belt type of continuously variable transmission (CVT) is provided, said CVT being placed in a transmission case and constituted that a V-belt is routed around a drive pulley and a driven pulley, said drive pulley being attached to one end of a crankshaft, said driven pulley being attached to one end of a transmission shaft placed parallel to said crankshaft, characterized in that said transmission shaft is supported for free rotation with said engine case, one end of said transmission shaft is made to project in cantilever fashion from said

Starter/generator for motor vehicles

The invention provides a starter/generator which can be disposed on an end of a crankshaft without increasing the length of the crankshaft. To achieve this, a starter/generator includes a brushless motor 44 having a rotor 15 connected to an end of a crankshaft 12 and a stator including a starter coil 51 and a generator coil 50, a start control means (a three-phase driver 104, a rotor sensor 56, etc.) for converting a battery voltage into AC power and supplying the AC power to the starter coil 51 of the brushless motor 44, and a generation control means (a regulator 52) for supplying electromotive forces induced across the generator coil 50 of the brushless motor 44 to electric loads. ...

Engine lubricating device

In a lubricating device for an engine which includes an oil pump disposed in a crankcase and feeding oil in response to the rotation of a crankshaft, the oil pump can be assembled separately from the crankcase, which is effective in improving assembling and maintenance work of the oil pump.A holder 34 rotatably supporting a crankshaft 20 is attached to a surface 108 of a crankcase 26, and constitutes a pump housing 90 for an oil pump 82 together with a cover 91 fastened to the holder 34. ...

Overhead camshaft V-2 engine

An overhead camshaft V-2 engine includes a single power transmitting mechanism (80) disposed on one side of a crankcase (11) for transmitting rotational power from a crankshaft (21) to respective camshafts (71) of drive valve mechanisms (70). The power transmitting mechanism has a driving member (111) mounted to only one end portion (22) of the crankshaft. The thus arranged power transmitting mechanism enables downsizing of the engine. Two intake ports are formed in respective cylinder heads of two cylinder blocks and open at one end to respective first surfaces of the cylinder heads facing in a first direction, and two exhaust ports are formed in the respective cylinder heads of the cylinder blocks and open at one end to respective second surfaces of the cylinder heads facing in a second direction opposite to the first direction. With this arrangement of the intake and exhaust ports, intake pipes can be arranged with a high degree of freedom. ...

Seal structure in engine body

coupled to a lower end face of a cylinder block (7) of an engine and includes first and second case halves (6L,6R) coupled to each other in a plane perpendicular to joint surfaces of the cylinder block and the crankcase. A U-shaped seal groove (87) is defined in one of the joint surfaces of the case halves. Enlarged recesses (86a) are made at opposite ends of the seal groove. A bar-shaped seal member (86) having enlarged end portions formed at opposite ends thereof is mounted in the groove such that the enlarged end portions are filled in the enlarged recesses. Thus, it is possible to ensure accuracy of the positioning of the bar-shaped seal member without the need for a high level skill; thereby, reliably sealing portions of intersection between the joint surfaces of the cylinder block and the crankcase and the joint surfaces of the case halves of the crankcase. ...

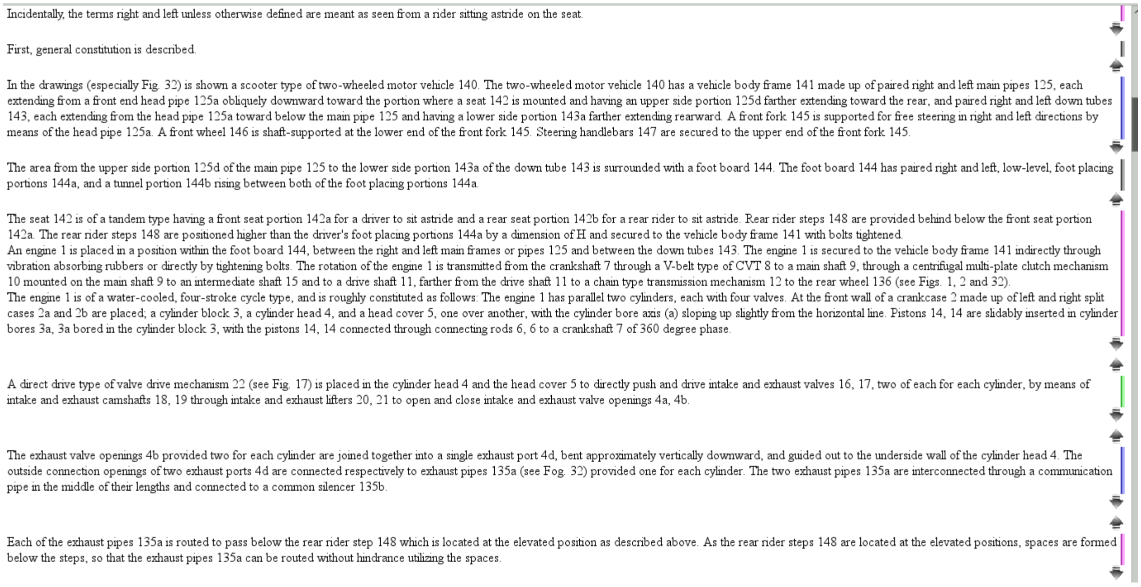
A power train for vehicles.

The present invention provides a power train for vehicles, the vehicles having engines. The power train according to the present invention is essentially composed of a swashplate type continuously variable transmission (T) which has (a) a transmission cylinder having an axis and supported rotatable about the axis, a plurality of pump plunger holes and motor plunger holes formed therein parallel to the axis and annularly about the axis, the transmission cylinder being connected to a crankshaft (17) of the engine to be rotated thereby; (b) a plurality of pump plungers (28) having plunger heads and disposed slidable in respective pump plunger holes so that the plunger heads protrude out of the transmission cylinder, each of the heads having a concavity formed therein; (c) a pump swashplate having an axis and a plurality of projections formed on a surface thereof, the pump swashplate (29) being disposed so that the axis thereof is inclined to the axis of the transmission cylinder and the pump swashplate is rotatable about the axis thereof, the projections of the pump swashplate being received by the respective concavity in the head of the pump plungers (28), whereby the pump plungers pump out a fluid by a reciprocal movement therec

Figure 7.12: Topic-based Feedback on Topic 5, i.e. the topic related to the concept of engine cooling.

us consider that in the case of our example query "Engine", the user is interested in topic 5, i.e. the topic related to the concept of engine cooling. This topic contains words such as "rear", "seat", "front" etc. It can be seen from Figure 7.8 that the first document in the retrieved list of documents, namely the document titled "Engine", has some segments classified to topic 5 (see the rightmost region coloured in magenta of the stacked bar chart on top of the result-list). In order to read this piece of information quickly, it is convenient for a user to *jump-in* to the first segment of the document on this particular topic. The interface enables the user to do this by providing hyperlinks to the sections of documents on respective topics. In this case, the user can directly jump in to the first magenta coloured segment without needing to scroll through the document himself; this is demonstrated in Figure 7.13.

To provide the navigation functionality, the system classifies the sections of a



Incidentally, the terms right and left unless otherwise defined are meant as seen from a rider sitting astride on the seat.

First, general constitution is described.

In the drawings (especially Fig. 32) is shown a scooter type of two-wheeled motor vehicle 140. The two-wheeled motor vehicle 140 has a vehicle body frame 141 made up of paired right and left main pipes 125, each extending from a front end head pipe 125a obliquely downward toward the portion where a seat 142 is mounted and having an upper side portion 125d farther extending toward the rear, and paired right and left down tubes 143, each extending from the head pipe 125a toward below the main pipe 125 and having a lower side portion 143a farther extending rearward. A front fork 145 is supported for free steering in right and left directions by means of the head pipe 125a. A front wheel 146 is shaft-supported at the lower end of the front fork 145. Steering handlebars 147 are secured to the upper end of the front fork 145.

The area from the upper side portion 125d of the main pipe 125 to the lower side portion 143a of the down tube 143 is surrounded with a foot board 144. The foot board 144 has paired right and left, low-level, foot placing portions 144a, and a tunnel portion 144b rising between both of the foot placing portions 144a.

The seat 142 is of a tandem type having a front seat portion 142a for a driver to sit astride and a rear seat portion 142b for a rear rider to sit astride. Rear rider steps 148 are provided behind below the front seat portion 142a. The rear rider steps 148 are positioned higher than the driver's foot placing portions 144a by a dimension of H and secured to the vehicle body frame 141 with bolts tightened.
An engine 1 is placed in a position within the foot board 144, between the right and left main frames or pipes 125 and between the down tubes 143. The engine 1 is secured to the vehicle body frame 141 indirectly through vibration absorbing rubbers or directly by tightening bolts. The rotation of the engine 1 is transmitted from the crankshaft 7 through a V-belt type of CVT 8 to a main shaft 9, through a centrifugal multi-plate clutch mechanism 10 mounted on the main shaft 9 to an intermediate shaft 15 and to a drive shaft 11, farther from the drive shaft 11 to a chain type transmission mechanism 12 to the rear wheel 136 (see Figs. 1, 2 and 32).
The engine 1 is of a water-cooled, four-stroke cycle type, and is roughly constituted as follows: The engine 1 has parallel two cylinders, each with four valves. At the front wall of a crankcase 2 made up of left and right split cases 2a and 2b are placed; a cylinder block 3, a cylinder head 4, and a head cover 5, one over another, with the cylinder bore axis (a) sloping up slightly from the horizontal line. Pistons 14, 14 are slidably inserted in cylinder bores 3a, 3a bored in the cylinder block 3, with the pistons 14, 14 connected through connecting rods 6, 6 to a crankshaft 7 of 360 degree phase.

A direct drive type of valve drive mechanism 22 (see Fig. 17) is placed in the cylinder head 4 and the head cover 5 to directly push and drive intake and exhaust valves 16, 17, two of each for each cylinder, by means of intake and exhaust camshafts 18, 19 through intake and exhaust lifters 20, 21 to open and close intake and exhaust valve openings 4a, 4b.

The exhaust valve openings 4b provided two for each cylinder are joined together into a single exhaust port 4d, bent approximately vertically downward, and guided out to the underside wall of the cylinder head 4. The outside connection openings of two exhaust ports 4d are connected respectively to exhaust pipes 135a (see Fog. 32) provided one for each cylinder. The two exhaust pipes 135a are interconnected through a communication pipe in the middle of their lengths and connected to a common silencer 135b.

Each of the exhaust pipes 135a is routed to pass below the rear rider step 148 which is located at the elevated position as described above. As the rear rider steps 148 are located at the elevated positions, spaces are formed below the steps, so that the exhaust pipes 135a can be routed without hindrance utilizing the spaces.

Figure 7.13: A sample screen-shot of TopicVis showing the topic classified sections bordered on the right with different colours.

document into one of the topical classes. This classification is achieved as follows. The sections are first identified by the explicitly marked XML tag pairs “<p>” and “</p>”. For each section, we compute its relative topical composition using the word-topic mapping given by the ϕ matrix.

A section is then classified to the topic having the maximum proportion only if the relative difference between the topic with the largest proportion and that with the second largest proportion is higher than a pre-configured threshold, which in our case is set to 0.5. More specifically, assuming that the relative proportions of the two most frequent topics in a section are η_1 and η_2, the section is assigned to the topic class corresponding to η_1 if and only if or in other words, η_1 is at least 50% higher than η_2.

This thresholding is important, because some sections of a document contain more or less a uniform contribution from all topic classes. Such sections are treated as unclassified in our approach. The specific value of 0.5 was chosen after manual inspection of the classified segments for one sample document in the collection.

$$\tau = \frac{\eta_1 - \eta_2}{\eta_1} > 0.5 \tag{7.5}$$

Section Name	Topical Composition	Assigned Topic
S_1	(0.8, 0.2, 0.1)	1
S_2	(0.2, 0.1, 0.8)	3
S_3	(0.05, 0.05, 0.9)	3
S_4	(0.98, 0.01, 0.01)	1
S_5	(0.1, 0.05, 0.85)	3
S_6	(0.25, 0.35, 0.40)	N/A

Table 7.1: An example of a query with 5 sections.

Let us illustrate this topic classification of the document segments with a simple example. Consider a patent document (or a query) comprised of 6 sections as shown in Table 7.1. The table also shows the topical composition of each section, i.e. the relative proportion of constituent words in each topic. It can be seen that the first and the fourth sections are assigned to topic 1. For the first section, $\eta_1 = 0.8$ and $\eta_2 = 0.2$. By Equation 7.5, $\tau = (0.8 - 0.2)/0.8 = 0.75$, which implies that the contribution from topic 1 is 75% higher than the contribution from topic 2. Since the value of τ is higher than the threshold 0.5 in this case, this particular segment of the document is assigned to topic 1. Similarily, the other remaining segments are assigned to the respective topic classes, as shown in Table 7.1. An exception is the last segment where we see that the relative difference between the most frequent and the second most frequent topic is less than the theshold ($\tau = (0.8 - 0.2)/0.8 = 0.14 < 0.5$). This segment is thus not classified to any topic class.

While the process of choosing the thresholding parameter may seem ad-hoc, it is worth mentioning here that a change in the value of the thresholding parameter is responsible for only changing the topical class membership of sections in a document, and hence is less likely to have any significant impact on the user experience.

After the classification of each section into a topic, the text of a section, belonging to a partcular topic, is bordered with the corresponding colour. The unclassified sections, i.e the sections which could not be assigned to any topical class, are bordered in black (see the second and the fourth sections in Figure 7.13). Each section of text has two links - the *next* and the *previous*, with links respectively to the next and to the previous section within a document belonging to the current topic being viewed. The *next* (*previous*) link of the last (first) section of a document enables the user to view the first (last) section of the next (previous) document on the current topic, thus supporting inter-document navigation as well. No links are generated from and to unclassified sections.

Figure 7.13 shows a screen-shot of the interface obtained after the magenta area of the top stacked bar chart of Figure 7.8 is clicked. The figure demonstrates that the system has automatically jumped to the first magenta section of the document "Engine". In fact the figure shows two more such magenta sections, one at the middle part of the screen and the other one right at the bottom. These (and the others not shown in the figure) can be accessed by sequentially clicking the next links provided at the bottom of each section. Additionally, the system also supports the traditional way of accessing each section of a document by vertical scrolling.

7.4 Experimental Investigation of TopicVis

In this section, we describe an experimental investigation of the effectiveness of the TopicVis interface. We break up the evaluation task into two independent phases. Firstly, we perform an automatic quantitative evaluation of the topic-based feedback in order to investigate the effectiveness of the proposed method of reranking the retrieved list of documents. Next, in order to demonstrate the usability of the interface, we perform a task oriented user study to investigate whether users on average are able to seek information faster using our interface in comparison to a standard IR interface.

To evaluate TopicVis, we used the CLEF-IP 2010 dataset. Since for qualitative evaluation of the interface we needed to peform a task driven user study, we restricted the test query set to a subset of 25 queries. These are the first 25 queries taken from the full list of CLEF-IP 2010 queries, ordered lexicographically.

The TopicVis web interface is developed using Java servlets[7] and Java server pages (JSPs)[8]. The client-side HTML pages uses Javascript[9] for validation checking. The retrieval system used in the back-end of TopicVis is the extended SMART used for our earlier experiments (see Section 3.3 for more details). The communication between the web server application and SMART is achieved by the standard interprocess communication API of the Java virtual machine (JVM). The retrieval results obtained from SMART are then loaded in memory by the web application, and then rendered with the help of the JFreeChart API[10].

We set the number of topics, K, to a value of 5, as used in the TRLM experiments described in Section 6.3.2. For each query, the TopicVis interface reports 50 documents ranked by the mTRLM, with 5 documents on each page.

The following two sections describe the automatic evaluation of our topic-based feedback method and the user study experiment.

7.4.1 Quantitative Evaluation of Topic-based Feedback

In this section, we evaluate the effectiveness of the topic-based feedback. Our objective is to see how effective is the reranking step in topic-based feedback to report documents relevant to the chosen aspect of the query at top ranks. Consequently, the relevance judgements of the CLEF-IP 2010 dataset were used for the automatic evaluation.

The relevance judgements in the CLEF-IP 2010 dataset were obtained by real life patent examiners (Piroi et al., 2011). A document is marked as relevant if its contents invalidates any of the claims expressed in the patent query. However, the relevance judgement file has no information as to which particular claim(s) in the query patent is (are) invalidated by a relevant document. This information is however required for computing the retrieval effectiveness of the topic-based feedback

[7] http://www.oracle.com/technetwork/java/index-jsp-135475.html

[8] http://www.oracle.com/technetwork/java/javaee/jsp/index.html

[9] http://en.wikipedia.org/wiki/JavaScript

[10] http://www.jfree.org/jfreechart/

where the primary interest is to measure how effectively a retrieval system can report documents relevant to a particular claim.

In the absence of such information in the relevance judgements, for the purpose of automatic evaluation of topic-based feedback, we generate this information automatically. The way this information is generated, is as follows. First, we classified sections in the query patent to topic classes by following the methodology presented in Section 7.3.4. Recall that this way of classifying a section of the query to a topic involves selecting the topic with the maximum relative proportion in that section. Each section in the patent query is thus labelled into one of the topic classes. For each topic class, i.e. for a total of 5 in our case, we formed a sub-query by concatenating the text belonging to that class. At the end of this step, we thus had at most K non-empty subqueries, the content of each being solely constituted of a single topic.

To illustrate with an example, let us revisit the example document shown in Table 7.1. From this particular query we would obtain 2 subqueries one comprising of the concatenated segments S_1 and S_4 corresponding to topic 1, while the other constituting the concatenated segments S_2, S_3 and S_5 associated with topic 3. Note that in this case, there is no subquery formed for the second topic (see Figure 7.14). Figure 7.14 thus demonstrates that if K topics are used at most K non empty subqueries are formed. Note that the subquery corresponding to topic 2 is empty in the particular example of Figure 7.14.

S_1	1
S_2	3
S_3	3
S_4	1
S_5	3
S_6	-

S_1	1
S_4	1

S_2	3
S_3	3
S_5	3

Figure 7.14: Illustrative example of constituting topic focused subqueries from the CLEF-IP 2010 patents.

In the next step, we used each such topic focussed sub-query to retrieve results from the patent document collection. The set of documents, assumed to be relevant to this particular topic of the current patent query, was then the set of relevant documents (as obtained from the overall relevance judgements file) occuring in the top (say R) ranks of the retrieved result. The assumption here is that an artificially constructed query where the text pertains to a particular topic, would primarily retrieve documents relevant to that topic at top ranks. For example, the query formed from the segments S_1 and S_4 (see Figure 7.14) are likely to retrieve documents predominant in topic 1.

We therefore compute the intersection of the R top ranked documents retrieved for the sub-query with the set of full relevance judgements, so as to compute a new set of relevance judgements pertaining to each individual topic. We then use these *per-claim* relevance judgements for computation of the effectiveness of topic-based feedback, i.e. whether it retrieves topic focused relevant documents at early ranks, i.e. documents belonging to the set of *per-claim* relevant documents.

To illustrate the process of generating per-claim relevance assessments, which from now on we simply refer to as *sub-qrels*, let us consider a simple example. Let the full set of relevance assessments for a query comprise of the documents $D1, D_2 \ldots, D_5$. Let the set of documents retrieved for each sub-query be S_k, where $k = \{1, 2\}$. We take a subset of the top R ranked documents from each of these lists and call it S_k^R. In our example, let $R = 20$. The figure below shows a sample scenario where the documents D_1, D_2 and D_3 are marked as relevant for topic 1, since these documents occur within the top 20 documents of the ranked list retrieved with sub-query 1. Similarly, D_2, D_3 and D_4 are marked relevant for the second topic.

$$\overbrace{D_1 \quad \rlap{$\underbrace{}_{S_2^{20}}$}D_2 \quad D_3}^{S_1^{20}} \quad D_4 \quad D_5$$

The computation of the per-claim relevance assessment files for each topic this way, enables us to report the effectiveness of topic-based feedback. The objective of the experiment is to find whether per-topic based feedback is able to retrieve more documents relevant to a particular chosen topic at top ranks, as compared to standard retrieval where all topics are given equal weighting. As a baseline for these per-topic feedback experiments, we take the ranked list of documents as obtained by mTRLM, i.e. the one obtained through TopicVis prior to clicking an area in the pie chart.

As an evaluation metric, we use the PRES, which is a standard evaluation metric for patent search (Magdy and Jones, 2010b). The baseline evaluation metric, which we name *overall relevance*, is the PRES computed for the mTRLM retrieval using the full relevance assessments, averaged over the set of 25 topics. To evaluate the per-topic feedback we make use of the per-topic relevance assessments. We compute the total PRES as obtained by evaluating against each per-topic qrel file and then divide it by the number of such per-topic qrel files. Thus, we obtain the PRES for the per-topic based feedback averaged over those topics for which there exists at least one relevant document within the top R ranks. We then take the arithmetic mean of this average PRES over the set of 25 queries as the evaluation metric for per-topic feedback. We call this evaluation score the *per-claim relevance*, since these relevance assessments are derived from the artificially generated per-topic relevance assessments. The aim of this evaluation score is to determine how much topic focussed relevant content is reported at top ranks after the initial resultlist is reranked by topic focused feedback.

The parameter R, which is the cut-off rank considered for computing the per-claim qrels, is varied in the range of $[5, 50]$. Figure 7.15 shows a comparison of the two evaluation metrics, the first labelled as the *overall relevance*, and the second labelled as the *average per-claim relevance*. It can be seen from the figure that the average per-claim relevance is always higher than the overall relevance which shows that reranking documents by their topical compositions, i.e. topic focussed feedback, can report topic-focussed documents within top ranks.

As expected, with an increase in the value of the cut-off, the number of relevant documents included in each sub-qrel files increases as well. This is because a relevant document can be falsely assumed to be relevant to a certain claim only because it is

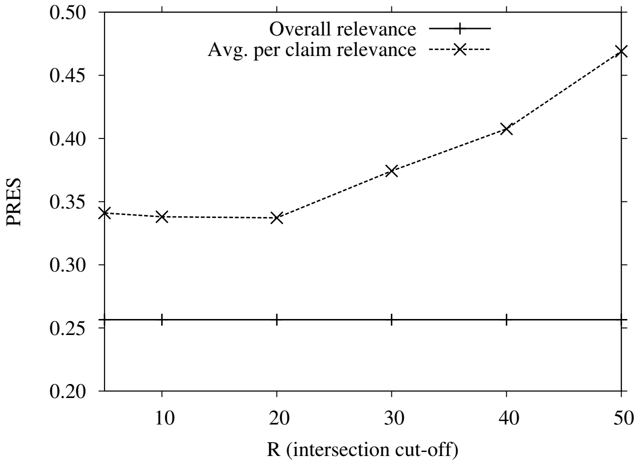


Figure 7.15: Per-topic feedback PRES averaged over topics.

within the top R retrieved documents, and the more we increase the value of R the less realistic our assumption becomes. The assumption however, is realistic enough for the value of $R = 5$, which leads to the conclusion that that a significant number of the topic focussed relevant documents is reported within top 5 ranks by the use of topic-based feedback.

In summary, we have shown in this section that the topic-based feedback feature of TopicVis can effectively report topic-focussed relevant content within top ranks. In the next section, we present a user based evaluation of the TopicVis interface.

7.4.2 Qualitative Evaulation of TopicVis

For a qualitaitive evaluation of the TopicVis interface, we conducted a task-based user study. For the study, a number of volunteer participants were given a task to complete using the TopicVis search interface. The interaction of the users with the interface while performing the tasks, was then analyzed to understand the usefulness of the system features as prescribed in (Vakkari, 2003). In this study, since the task of invalidating patent claims requires professional expertise as of a patent examiner (Piroi et al., 2011), we used a simpler user task of finding a yes/no answer to three questions for each patent query. The questions were formulated so that rather than being related to particular invention claims, they pertained to general information. For example, a sample question for the patent query titled "Sleep apnea therapy device using dynamic overdrive pacing" is

Q: *Is it true that an increasing metabolic demand causes an elevation in stroke volume, but an increasing heart rate from pacing causes a decrease in stroke volume?*

Participants were instructed to find answers to the three yes/no questions from documents retrieved in response to a patent query. The questions distributed to the participants are listed in Appendix **??**.

We distributed 25 queries among 8 participants, out of which 7 participants were assigned 3 queries each and one was asked to accomplish the task on 4 queries. Subjects were recruited by distributing a "call for participation" email among the researchers in the Centre for Next Genneration Localisation (CNGL). The email contained detailed instructions for the experiment and the URL of the TopicVis interface, so that interested researchers could visit the system themselves. The email asked recipients to sign up for the experiment by filling in an online registration form.

The set of registered users willing to participate in the study mainly comprised of PhD. students and post doctoral research fellows conducting research on mainly on IR and machine translation. The average age of the participants was approximately 28 years. None of the participants were familiar with patent prior art search and did not previously use any patent search tool. Most of them however, were familiar with topic modelling concepts.

A three minute tutorial video of the interface was shown to the registered participants. This was followed by a three minute practise session in which they were allowed to freely interact with the system. After this, the participants were asked to use the system to find answers to the three yes/no questions. Note that the users were allowed to freely interact with the system while understanding these tasks, i.e. they were free to use either the new topic-based features, such as the topic-based feedback and navigation, or use the standard search interface features only, such as the standard snippet and document views.

The interactions of the users with the interface during the task sessions were logged for subsequent analysis. The information stored in the logs comprised of the click information in certain action regions of the interface such as the pie-chart, the stacked bar charts, the title of retrieved documents etc.

Log Analysis

The collected user logs were analyzed to investigate how users interacted with the TopicVis interface while executing their tasks. The intention was to see whether the subjects were using the new features such, i.e. the topic-based feedback and navigation, more frequently than the standard baseline features of a search interface such as switching search result pages and clicking on the document titles. As a part of the log analysis, we calculated frequencies of four different events outlined as follows:

Page change: Denotes the number of times the user clicks on the next, previous or the direct pagination links to change the search result page.

Title: Represents the number of times the user clicks on the title of a document to open it in the standard view.

Pie chart: Represents how many times the user clicks on the pie chart areas for topic-based feedback.

Stacked bar chart: Represents the number of times the user clicks on the areas in the stacked bar charts and the navigational arrows for topic-based navigation.

Note that the first two of these events correspond to any standard search interface. The last two however are unique to TopicVis and correspond respectively to the fetures topic-based feedback and navigation.

In Figure 7.16, we show the frequency of user clicks on each for these four events. It can be seen that subjects were using the new features of TopicVis more than the standard features. The fact that there are more clicks on the pie chart as compared

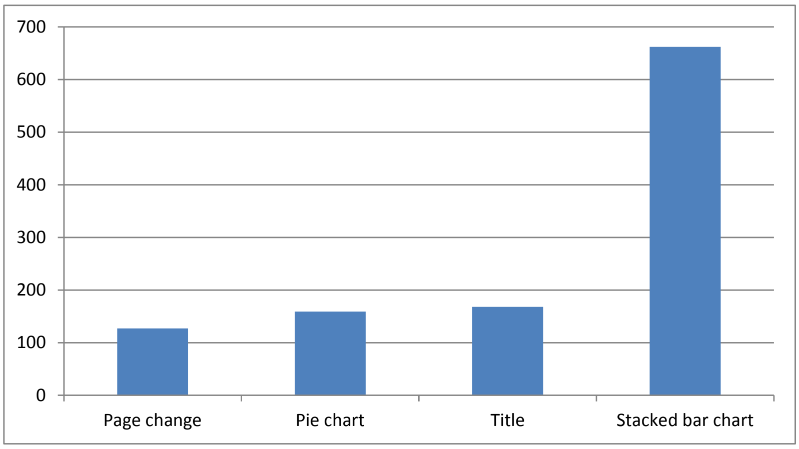


Figure 7.16: Frequency of user clicks on TopicVis features.

to the number of clicks on the pagination links, shows that the subjects were using topic-based feedback during their search tasks to locate the answers to the questions. Moreover, the number of clicks in the stacked bar charts is almost 5 times the number of times the standard view link was clicked, which shows that the subjects preferred the feature of topic-based navigation over that of standard browsing.

Questionnaire-based survey

After the termination of a user task session, the participants were presented with a questionnaire in order to get feedback about the usefulness and usability of the system features. The survey questions were intended to get user feedback on the novel user-interface features provided by TopicVis. Each question was formulated to qualitatively judge the usefulness of a particular feature of TopicVis.

For each feature, we presented questions providing five-point Likeart choice items, namely “strongly disagree”, “disagree”, “neither agree nor disagree”, “agree” and “strongly agree” respectively. The Likeart values assigned to these items were 1 to 5 respectively.

The inter-annotator agreement for the answers provided by the partcipants was fair with a measured value of 0.3043 for the Fleiss kappa (κ) showing that the subjects agreed fairly between themselves (Landis and Koch, 1977).

Appendix B lists the survey questions in details. Table **??** shows the Likeart scale answer values averaged over the set of participants. Table ?? does not show the full question text given to the participants but rather highlights the feature associated with each question, the usefulness of which is to be judged (see the question identification numbers in Table ?? to read the corresponding questions from Appendix B). Since, there was a fair amount of inter-annotator agreement, the average values are suggestive enough of the usefulness of each feature.

Feature	Question Id	Survey Item	LikeArt
Pie chart	1(a)	Accurate topic visualization	4.625
	1(b)	Accurate word labels	4.125
	1(c)	Effective topic-based feedback	4.000
	1(d)	Useful to find topic-focussed relevant documents	4.375
Stacked bar charts	2(a)	Useful query view	4.625
	2(b)	Useful document view	4.250
	2(c)	Accurate topical composition	4.000
	2(d)	Usefulness of snippets	1.250
Topic-based navigation	3(a)	Usefulness of next and previous links	4.500
	3(b)	Usefulness of vertical scroll	2.25
	3(c)	Efficient reading	4.375
	3(d)	Usefulness of standard document view	1.625

From the scores, we can see that the participants provided positive feedback on the new TopicVis features, e.g. the average score 4.625 for item 1a) shows that users on average strongly agreed that the pie chart accurately visualizes the topics in the retrieved set of documents. For the standard search interface features, the users on average tended to disagree, e.g. 2d), which shows that the subjects were not using the snippet view for to decide whether to view a document. Rather, they were using the stacked bar chart for the decision as indicated by the average score of 4 for 2c). In some cases, the subjects were using both the new and the standard features, as can be inferred from the average score of 2.25, which is somewhere in between the two extremes of agreeing and disagreeing, in 3b) suggesting that the subjects were using both the topic-based navigational arrows as well as the vertical scrolling. We suspect that in this particular case, the familiarity of vertically scrolling through a document might have affected their choice.

7.5 Summary and Outlook

This chapter presented a novel search interface, which in addition to the standard search engine features of retrieving a ranked list of documents and presenting these with associated titles and snippets, also provides the following features.

Firstly, it visualizes the query and the retrieved documents as a mixture of topics. The sections of the query are shown with associated topic classes. The visualization of the query is designed to assist the searcher in matching relevant information against the parts of the query. The visualization of the documents, in turn, is designed to provide visual cues relating to the content of a document and to save time in deciding whether to open a document for reading.

Secondly, the system lets a user select a particular topic of his choice for feedback. As a result of this topic-based feedback, the documents are reranked acccording to the decreasing contributions from the selected topic. This, in turn, is intended to lead to finding a relevant piece of information for a particular topic.

Thirdly, the interface provides easy access to document content by letting a user follow hyperlinks from one part of a document to another on the same topic. Using our system, the user can thus read through segments of documents on the same topic by following the links, without requiring him to look for related pieces of information in a document through manual scrolling.

In summary, this chapter showed that topic modelling can be applied for developing a user friendly search interface which not only allows a user to view the topical composition of retrieved documents but also allows him to browse through sections of documents on his chosen topic of interest. In the next chapter, we conclude the thesis by revisiting the research questions explored so far and also provide directions for future work.

Chapter 8

Conclusions

In this thesis, we have shown that topic modelling in the retrieved set of documents and also in the query proves beneficial not only for improving retrieval quality, but is also useful in developing a novel interactive search interface for presentation of retrieval results and providing effective means of navigation through the retrieved documents. In this chapter, we summarize the overall and individual contributions of this study and outline potential directions for future work.

8.1 Research Questions Revisited

In this section, we revisit the research questions, introduced in Chapter 1, and summarize how each one of them has been addressed in the previous chapters.

8.1.1 Sentence-based Query Expansion

The work in this thesis was motivated by the hypothesis that the pseudo-relevance feedback (PRF) in IR can potentially be improved by using information from topics that are relevant to one or more aspects of the given information need, rather than using information from whole documents. The first research question examined whether a simple measure such as term proximity is able to capture topical association between terms, i.e. whether two terms in close proximity in a document belongs to the same topical class. The unit of proximity explored in our experimental investigation was sentences, the reason being sentences can be considered as natural semantic units. The first research question, RQ-1, introduced in Chapter 1, is reproduced below.

RQ-1: *Can additional terms in close proximity to query terms from retrieved documents enrich the information need of the query and improve the retrieval effectiveness of ad-hoc IR?*

The objective of RQ-1 is to see whether terms extracted from close proximity of the query terms and thus by our hypothesis topically associated with the query terms, can help improve PRF. Our proposed method of sentence-based query expansion (SBQE) thus involves decomposition of the pseudo-relevant documents into smaller units, which in our case are sentences. We then expand the given query by adding sentences which are most similar to the query.

SBQE adds full sentences to the query in contrast to the standard term based approaches of adding top scoring terms to the query. The sentence-based approach is thus able to utilize the context information of a sentence, and is empirically shown to outperform the standard term based approaches to query expansion, which ignore the context information altogether. SBQE also discriminates between the relevant documents retrieved at rank 1 as against those retrieved at higher ranks (say 10), in the sense that it adds more sentences from the former and less from the latter.

An advantage of SBQE is that it is simple and straight-forward to implement; yet it produces significantly better results than the more involved techniques of generative models of relevance (Lavrenko and Croft, 2001).

The disadvantage of SBQE is that it is a two-step retrieval in comparison to the relevance model (RLM) (Lavrenko and Croft, 2001), which is a one-step method comprising of reranking the initially retrieved result-list. A second disadvantage is that the expanded queries are very long thus contributing to an increase in run-time of the feedback step.

The first research question RQ-1 is thus answered in positive. We conclude that using topically related terms relevant to the information need expressed in the query improves PRF.

8.1.2 Query Segmentation

Successful exploration of the first research question RQ-1 motivated us to study the complementary problem of multi-topical nature of long queries. In some retrieval domains, such as the patent search, the queries instead of being short and comprised of a few keywords, are full length documents. In patent search a query is a new patent claim and the objective is to retrieve prior articles (in)validating the new claims. Segmenting patent queries into topically coherent segments can be beneficial in these cases, because individual query segments are more focused on particular sub-information needs, and hence are able to retrieve more relevant documents pertaining to it, in contrast to the approach of using the whole document as a query which may fail to retrieve relevant documents for each sub-information need. The research question on query segmentation, introduced in Chapter 1, is reproduced below.

RQ-2: *Can segmentation of very long queries into topically coherent segments be utilized to improve IR effectiveness?*

Our work in Chapter 5 demonstrated that segmenting the queries into topically coherent blocks of text, treating each such segment as a separate query and merging the documents retrieved from each such segment improves retrieval quality in comparison to using full patent claims as queries. We also showed that the approach of using such segmented queries is also able to improve the PRF quality over that of using full queries. Chapter 5 thus answered research question RQ-2 in affirmative and the conclusion is that each topically focused query segment is able to focus on one particular aspect of the information need and hence leads to more effective retrieval than when full queries are used.

Chapters 4 and 5 thus show that both **RQ-1** and **RQ-2** have been answered affirmatively, the implication of which is that using topically coherent text units can improve the PRF quality both for short and long queries. Addition of terms topically related to the query terms helps to enrich each aspect of the initial information need whereas topic focussed query segments of long queries serves to focus on each fine-grained aspect of the queries during retrieval.

A disadvantage of the methods proposed in Chapters 4 and **??** is that both these involve multi-step retrieval, that is to say, SBQE involves a two-step retrieval with the expanded query, whereas the method of segmented query retrieval involves as many different retrieval steps as the number of query segments obtained. Moreover, these methods are distinctly different from each other, one is applicable for keyword type queries and the other for the very long queries. It would be ideal to combine the working principle of these two approaches into a single integrated framework. The work in Chapter 6 presented a way to achieve this.

8.1.3 Topical Relevance Model: Topical Segmentation of Pseudo-Relevant Documents and Queries

Instead of having two separate complementary methods, one applicable for short queries and the other for long queries, the next research question investigated whether we can combine the working principles of the methods explored in RQ-1 and RQ-2 under a single framework. We also explored techniques of modelling the topic distribution of terms in pseudo-relevant documents and queries instead of relying on the proximity hypothesis of term relatedness. Keeping these objectives in mind, we thus formulated the third research question, RQ-3, which is reproduced as follows.

RQ-3: *Can topic modelling prove beneficial in improving the retrieval effectiveness for both short and long queries thus unifying the solutions of RQ-1 and RQ-2?*

Note that In Chapter 6, we develop the model, namely topical relevance model (TRLM), and show that it works well both for short keyword type queries as well as for the very long patent queries. We thus provided an affirmative answer to research question RQ-3.

Towards the end of Chapter 6 in Section 6.4.4, the TRLM is compared with the SBQE. We see that TRLM (without query expansion) is able to achieve a higher precision at top 10 ranks compared to SBQE. The recall and MAP values are higher for SBQE, the reason for which is that SBQE being a two step retrieval process is able to retrieve more relevant documents (thus leading to an increase in recall) during the feedback step with the expanded query, whereas TRLM being a single step retrieval process relies only on reranking the retrieved set of documents. The advantage of TRLM is that it is faster than SBQE. TRLM with query expansion (TRLM_QE) is further able to increase the MAP by retrieving additional relevant documents in the second retrieval step with the expanded query. TRLM_QE achieves MAP values very close to SBQE, while achieving significantly higher P@10 values. Moreover, TRLM_QE, owing to a smaller number of additional expansion terms, is computationally more efficient that SBQE.

In Section 6.4.5, TRLM is compared against the segmented query retrieval method, namely SEG, proposed in Chapter 5. Again, TRLM is much faster than SEG because of the obvious disadvantage of executing as many retrieval steps as the number of query segments. In spite of being a single step retrieval process, TRLM is able to outperform SEG in retrieval effectiveness measured in terms of MAP. SEG however scores higher for recall oriented metrics, such as the percentage recall and PRES, due to the fact that retrieving with different query segments enables the method to find more relevant documents form the collection. The likely reason for this is due to the averaging effect of using the same number of topics in the topic modelling step of TRLM. We also showed that a version of TRLM (denoted as TRLM*, which uses the optimal number of topics for each individual query, is able to outperform SEG in terms of PRES and MAP.

8.1.4 Topic Visualization

The last research question, explored in this thesis is about exploring the potential benefits of topic modelling for providing a more convenient access to relevant information.

RQ-4: *Can topical segmentation of the documents and the queries be helpful in providing a topic-based access to relevant information?*

Towards answering this question, we developed a user interface, which we named TopicVis, designed to facilitate topic-based navigation through search results and topic-based feedback to rerank retrieved documents on the basis of a user selected topic.

We evaluated TopicVis both quantitatively and qualitatively on the CLEF-IP 2010 patent prior art search task. The quantitative evaluation showed that TopicVis is able to effectively retrieve documents relevant to a particular claim of the patent query. The qualitative evaluation showed that the visualization of the query helps in matching relevant information against the parts of the query compared to standard ranked retrieval interfaces. Moreover, the visualization of the documents helps in providing a visual cue about the content of a document and saves time in deciding whether to open a document for reading. thus leading to quickly finding a relevant piece of information for a particular topic. Moreover, TopicVis provides an easy access to document content by letting a user follow hyperlinks from one part of a document to another on the same topic. A user of TopicVis can thus read through segments of documents on the same topic by following the links, without requiring him to look for related pieces of information in a document through manual scrolling.

The work in Chapter 7 thus demonstrated that topical segmentation of documents and queries can provide convenient access to relevant pieces of information, thereby providing an affirmative answer to research question RQ-4.

8.2 Future Work

While this thesis has applied techniques of topic modelling in retrieved documents and queries to improve the quality of retrieval and user satisfaction in relevant information access, there remain a number of avenues for future work, which we believe deserve further exploration.

Chapter 4: For term expansion, it is observed that a variable number of expansion terms chosen dynamically for the individual topics provides best effective results (Ogilvie et al., 2009). Future work in this direction can involve exploring whether employing a variable number of sentences, i.e. using different values for the parameter m in SBQE for different topics, yields further improvement in the retrieval effectiveness.

Moreover, the SBQE method can also be extended to handle fixed length word windows (pseudo-sentences) instead of natural sentences.

Furthermore, exploring whether applying any of the sentence scoring mechanisms outlined in (Murdock, 2006; Losada, 2010) instead of the cosine similarity for selecting the candidate sentences proves more beneficial for SBQE is also worth investigating.

Chapter 5: For the segmented query retrieval algorithm - SEG, we observed that the round-robin merging technique outperforms the standard merging technique of COMBSUM. The merging technique applied for our experiments described in Chapter 5 was unweighted. Future work in this direction may investigate the whether applying a round-robin technique weighted by the similarities between documents and query segments can improve the results further.

Chapter 6: The underlying topic model applied in the TRLM method, proposed in Chapter 6, is the latent Dirichlet allocation (LDA). A limitation of LDA is that it is a parametric method, that is to say the number of topics has to be pre-configured before inferring the posterior probabilities of the model. This limitation of LDA is also applicable to the TRLM. In the context of TRLM, we have shown in Section 6.3.2 that the performance of TRLM averaged over a set of queries is relatively insensitive to the choice of the number of topics. However, a per query analysis presented in Section 6.4.1 showed that a judicious choice of K can lead to a significant difference in retrieval effectiveness. The reason for this is mainly due to the specificity in the information need expressed in the query. For a more general query, we expect a higher number of topics manifested in the retrieved set of documents in comparison to a query which is more specific.

This limitation of using a fixed value of the parameter K can be overcome by employing a non-parametric generalization of LDA. One such generalization is the hierarchic LDA (hLDA) (Blei et al., 2010). The output of hLDA is a rooted tree, where the most general topic represents the root, and more specific topics are encountered as one traverses down the tree. The single layer of hidden nodes in the TRLM may thus be replaced by this rooted tree hierarchy of topic nodes. As a result of this extension, the tree of topic nodes would be deeper for a query with broad information need, whereas for a query with more specific information need the tree would be shallower. It will be of particular interest to see the effect of this extension of TRLM on the retrieval effectiveness.

Chapter 7: The proposed extension to TRLM with hLDA can also be applied to extend the TopicVis interface. It is expected that due to such an extension, the mapping from the topics to the information need aspects of the query could potentially be more accurate. This in turn should lead to a user in more accurately discovering latent aspects of the information need, as a result of which he could experience more accurate topic-based information access through topic-based feedback and navigation.

Moreover, the topic-based navigation can be extended to organize the segments of a document into a hierarchy of topics by using a hierarchic topic modelling approach such as the hLDA, in contrast to organizing the document into a flat list of segments classified into one of the topic classes. This way of organizing the information would enable users to view sections of documents covering a broad topic following which they can progressively view sections of documents on more specific topics.

8.3 Closing Remarks

We believe that the work presented in this thesis has opened potential new research directions for exploiting sub-document or sub-query level information not only in improving retrieval effectiveness, but also in providing a more convenient topic-based access to relevant pieces of information to the users of a search system. We hope that this work will act as a starting point for other researchers to continue investigations on the problems that we addressed in an endeavour to further improve the techniques presented in this thesis and find further applications for them.

Appendices

Appendix A

Publications

The research presented in this dissertation was published in several peer-reviewed conference proceedings. The work on document segmentation, presented in Chapter 4 is presented in (Ganguly et al., 2011a). The work in Chapter 5 appears in two papers, namely (Ganguly et al., 2011c) and (Ganguly et al., 2011b). The work in Chapter 6 is presented in (Ganguly et al., 2012b). The methodology, developed in Chapter 6 has also been applied successfully on cross-language information retrieval (CLIR). The details appear in (Ganguly et al., 2012a). The search interface *TopicVis*, presented in Chapter 7, will appear as a demonstration paper in SIGIR 2013 (Ganguly et al., 2013).

A.1 Publications from this work

1. **D.Ganguly**, M.Ganguly, J.Leveling and Gareth J.F. Jones, TopicVis: A GUI for Topic-based Feedback and Navigation, In *Proceedings of SIGIR 2013*, Dublin, Ireland, July 2013.

2. **D.Ganguly**, J.Leveling and Gareth J.F. Jones, Cross-Lingual Topical Relevance Models, In *Proceedings of 24th International Conference on Computational Linguistics (COLING 2012)*, Mumbai, India, December 2012.

3. **D.Ganguly**, J.Leveling, and G. J. F. Jones, Topical Relevance Models, In *Pro-*

ceedings of the Eighth Aisa Information Retrieval Societies Conference (AIRS 2012), Tianjin, China, December 2012. *(Best Poster Paper Award).*

4. **D.Ganguly**, J.Leveling, and G. J. F. Jones, Utilizing sub-topical structure of documents for Information Retrieval, In *Proceedings of the Workshop for Ph.D. Students in Information and Knowledge Management (PIKM 2011)*, CIKM 2011, Glasgow, UK, October 2011.

5. **D.Ganguly**, J.Leveling, and G. J. F. Jones, United we fall, Divided we stand: A study of Query Segmentation and PRF for Patent Prior Art Search, In *Proceedings of the Patent Information Retrieval Workshop, CIKM 2011*, Glasgow, UK, October 2011.

6. **D.Ganguly**, J.Leveling, and G. J. F. Jones, Patent Query Reduction using Pseudo Relevance Feedback, In *Proceedings of the 20th Conference on Information and Knowledge Management(CIKM 2011)*, Glasgow, UK, October 2011.

7. **D.Ganguly**, J.Leveling and Gareth J.F. Jones, Query Expansion for Language Modeling using Sentence Similarities. In *Proceedings of the 2nd Information Retrieval Facility Conference, IRFC 2011*, Vienna, Austria. pp. 62-77.

8. **D.Ganguly**, J.Leveling and Gareth J.F. Jones, Exploring Accumulative Query Expansion for Relevance Feedback. In *Proceedings of INEX-2010.*

9. **D.Ganguly**, J.Leveling and Gareth J.F. Jones, Exploring Sentence level Query Expansion in Language Model based IR. In *Proceedings of the 8th International Conference on Natural Language Processing, ICON 2010*, Kharagpur, India. pp. 18-27

A.2 Other publications during my PhD. study

1. **D.Ganguly**, J.Leveling and Gareth J.F. Jones, A Case Study in Decompounding for Bengali Information Retrieval, In *Proceedings of Conference and Labs of the Evaluation Forum (CLEF 2013)*, Valencia, Spain (To Appear).

2. **D.Ganguly**, J. Leveling and Gareth J.F. Jones, LDA-smoothed Relevance Model for Document Expansion: A Case Study for Spoken Document Retrieval, In *Proceedings of SIGIR 2013*, Dublin, Ireland, July 2013.

3. J.Leveling, **D.Ganguly** and Gareth J.F. Jones, Approximate Sentence Retrieval for Scalable and Efficient Example-based Machine Translation, In *Proceedings of 24th International Conference on Computational Linguistics (COLING 2012)*, Mumbai, India, December 2012.

4. **D.Ganguly**, J.Leveling and Gareth J.F. Jones, DCU@FIRE-2012: Rule-based Stemmers for Bengali and Hindi, In *Working Notes of Forum of Information Retrieval and Evaluation (FIRE 2012)*, Calcutta, India, 17-19 December 2012.

5. **D.Ganguly**, J.Leveling, and Gareth J.F. Jones, DCU@INEX 2012: Exploring Sentence Retrieval For Tweet Contextualization In *Proceedings of CLEF 2012.*

6. **D.Ganguly**, J.Leveling, and Gareth J.F. Jones, Overview of the Personalized and Collaborative Information Retrieval (PIR) Track at FIRE-2011, In *Proceedings of Forum of Information Retrieval and Evaluation (FIRE) 2011*, Mumbai, India, 2011 (To appear).

7. **D.Ganguly**, J.Leveling, and Gareth J.F. Jones, Simulation of Within-Session Query Variations using a Text Segmentation Approach, In *Proceedings of the Conference on Multilingual and Multimodal Information Access Evaluation (CLEF 2011)*, Amsterdam, The Netherlands, September 2011.

8. **D.Ganguly**, J.Leveling, W.Li and Gareth J.F. Jones, Towards Evaluation of Personalized and Collaborative Information Retrieval, In *Proceedings of the First Workshop on Personalised Multilingual Hypertext Retrieval (PMHR 2011) at ACM Hypertext 2011*, Eindhoven, The Netherlands, June 2011.

9. **D.Ganguly**, J.Leveling and Gareth J.F. Jones, Automatic Generation of Query Sessions using Text Segmentation, In *Proceedings of the Information Retrieval Over Query Sessions Workshop at ECIR 2011*, Dublin, Ireland, April 2011.

10. W.Li, **D.Ganguly** and Gareth J.F. Jones, Using Recommenders for Enhanced Domain-Specific Information Retrieval, In *Proceedings of the 3rd International Conference on the Theory of Information Retrieval (ICTIR'11)*, Bertinoro, Italy, September 2011.

11. **D.Ganguly**, J.Leveling and Gareth J.F. Jones, DCU and ISI@INEX 2010: Ad-hoc and Data-Centric tracks. In *Proceedings of INEX-2010.*

Appendix B

Qualitative Evaulation of TopicVis

The survey questions for the qualitative evaulation of TopicVis features are listed as follows along with the average answer scores obtained from the participants.

1. The pie chart

   a) The pie chart accurately visualizes different topics in the retrieved documents. (4.625)

   b) The word labels in the pie chart are helpful in distinguishing between the topics. (4.125)

   c) Clicks on different regions in the pie chart (topic-based feedback) show documents focusing on the selected topic at top ranks. (4)

   d) The pie chart was beneficial in completing the assigned task, i.e. finding answers to the questions. (4.375)

2. The stacked bar chart

   a) The stacked bar chart for the query (top left) accurately visualized the different topics in the query. (4.625)

   b) The stacked bar chart for each retrieved document helped in completing my task. (4.25)

c) My decision of opening a document for viewing was primarily based on the stacked bar chart. (4)

d) My decision of opening a document for viewing was primarily based on the snippet content. (1.25)

3. The topic-based navigation links within the document view

   a) I used the links to navigate within documents. (4.5)

   b) I used the scroll-bar to navigate within documents. (2.25)

   c) Using the links for navigation saved a lot of my reading effort. (4.375)

   d) The standard document view is more convenient to use than the topical navigation view. (1.625)

Table B.1 shows the individual Likeart answers provided by the participants. Clearly it can be seen that the participants gave positive response for the new TopicVis features almost unanimously. They also evenly agreed that the standard search engine features such as the snippet view (2(d)) and the standard document view (3(d)) were not useful enough in the patent search task.

	Question#											
User#	1(a)	1(b)	1(c)	1(d)	2(a)	2(b)	2(c)	2(d)	3(a)	3(b)	3(c)	3(d)
1	4	3	3	4	5	4	4	1	5	1	4	1
2	4	4	3	3	4	3	4	2	4	2	4	2
3	5	5	4	4	5	4	4	1	5	2	4	1
4	5	5	5	4	5	5	4	2	4	3	5	1
5	4	4	5	5	5	5	4	1	4	3	5	1
6	5	4	4	5	4	4	4	1	5	2	4	2
7	5	4	5	5	5	5	4	1	4	3	5	3
8	5	4	3	5	4	4	4	1	5	2	4	2
Avg.	4.63	4.13	4.00	4.38	4.63	4.25	4.00	1.25	4.50	2.25	4.38	1.63

Table B.1: Likeart values assigned by individual partcipants for each question.

Appendix C

Task-based User Study of TopicVis

Three yes/no questions were formulated for each of the 25 patent queries used in the evaluation test set of TopicVis. The total number of questions is therefore 75. The yes/no questions, which the participants were asked to answer for each of the 25 queries, are listed below. The answers are also provided alongside each question.

1. Longitudinal coupled multiple mode surface acoustic wave filter

   (a) A SAW device typically comprises of input/output interdigital transducer (IDT) electrodes on piezoelectric substrates. (TRUE)

   (b) A filter with good characteristics is one in which the attenuation near the pass band is superior, with ripples being produced in the band. (FALSE)

   (c) It has been a common trend to reduce the number of parts and combine several parts into a composite form in the circuit configuration of recent elastic surface wave filter devices. (TRUE)

2. Hybrid film, antireflection film comprising it, optical product, and method for restoring the defogging property of hybrid film.

   (a) The use an epoxy compound with the epoxy group at only one end of the molecule, is advantageous for the antireflection effects. (FALSE)

   (b) In direct bonding, hydrogen bonds and covalent bonds are not formed together. (FALSE)

   (c) A low reflection plastic with a low reflectance and stain resistance comprising a plastic substrate is a patented technology. (TRUE)

3. Organic electroluminescent device having stacked electroluminescent units

   (a) Organic compounds makes use of spontaneous light, has a high response speed and has no dependence on an angle of field. (TRUE)
   (b) An organic light-emitting diode (OLED) device includes a substrate, and a cathode disposed over the substrate. (FALSE)
   (c) A patterned donor transfer substrate and a laser light absorbing layer is placed over the donor transfer substrate in an OLED device. (FALSE)

4. Fragranced compositions comprising encapsulated material:

   (a) "Microencapsulation" is a process by which one or more ingredients become encased in a hardened polymer. (TRUE)
   (b) Microcapsules having colorant on their exterior surfaces can transfer the colorant when the capsules contain liquids which wet the colorant. (TRUE)
   (c) In a machine dishwashing tablet the coating layer comprises of materials selected from the group consisting of fatty acids, alcohols, diols, esters and ethers and mixtures thereof. (TRUE)

5. Cyclosporin-based pharmaceutical compositions

   (a) The transesterified and polyethoxylated vegetable oil may also comprise esters of saturated or unsaturated C12-20 fatty acids with glycerol or propylene glycol, for example glycerol monooleate. (TRUE)
   (b) The term "C2-C12 alkynyl" refers to a straight or branched alkynyl chain having from two to twelve carbon atoms. (TRUE)
   (c) Many anti-cancer agents drugs are readily absorbed in the digestive tract. (FALSE)

6. Sleep apnea therapy device using dynamic overdrive pacing

   (a) If a person restfully rides in car on a very bumpy road, a pacemaker can erroneously increase his heart rate dramatically at a time when such an increase is not wanted. (TRUE)
   (b) In general, an increasing metabolic demand causes an elevation in stroke volume, but an increasing heart rate from pacing causes a decrease in stroke volume. (TRUE)
   (c) The step of administering a polarization calibration pulse within the refractory period does not need to wait after the depolarization event. (FALSE)

7. 5-AMINOLEVULINIC ACID SALT, PROCESS FOR PRODUCING THE SAME AND USE THEREOF

   (a) The laser peeling is a therapeutic method wherein the skin surface is burnt by irradiating with laser beams instead of the application of chemicals. (TRUE)
   (b) Reacting an oil-soluble organic compound with molecular oxygen in the presence of a water-insoluble sensitizer in an organic solvent phase under irradiation of light produces a water-insoluble organic oxide. (FALSE)
   (c) Nitrate nitrogen present in food is partly reduced into nitrous acid by enteric bacteria in the living body. (TRUE)

8. Oil compositions for improved fuel economy

   (a) Diesel internal combustion engines mounted on motor-driven vehicles, constructions machines and power generators are generally driven using gas oil or heavy oil. (TRUE)
   (b) A low sulfate ash lubricating oil composition comprises of an oil of lubricating viscosity, 0.1 to 3.0% of a calcium overbased acidic material. (TRUE)

(c) Borated dispersants may be prepared by boration of succinimide, succinic ester, benzylamine and their derivatives. (TRUE)

9. Laser thermal transfer donor including a separate dopant layer

   (a) The spectral distribution, of emitted light cannot be modified by introducing a "dopant" into the electron-transporting layer. (FALSE)
   (b) Many CMOS structures often employ borderless diffusion regions adjacent to isolation regions. (TRUE)
   (c) 2-methyl-8-hydroquinoline aluminum is a light-emissive organic fluorescent dye which is useful as a donor layer for selective transfer onto an organic EL display device to form red, green, or blue light emitting subpixels. (TRUE)

10. Image forming apparatus, method of controlling the same, computer product, and process cartridge.

    (a) An image forming apparatus may refer to an electrophotographic system such as a copying machine, a printer, a facsimile machine etc. (TRUE)
    (b) The overlap deviation of images can be increased by the multicolor development with a single image forming unit. (FALSE)
    (c) A scorotron charger, which is a charging means, is used for image forming processes of each color of RED (R), GREEN (G), and BLUE (B). (FALSE)

11. Integral belt for an extended nip press

    (a) A cylindrical endless elastic body layer can be formed by impregnating a liquid elastic body precursor into a fibrous material and curing the liquid elastic body precursor. (TRUE)
    (b) A press fabric for the press section of a paper machine has a base fabric which includes a nonwoven mesh fabric. (TRUE)

(c) A long nip press is used in a papermaking machine to dewater a fibrous web. (TRUE)

12. Surgical stapling instruments including a cartridge having multiple staple sizes

    (a) A ”Plunger” is a rod having threaded screw mounting portions at only the proximal end. (FALSE)
    (b) A ”Thumbwheel” is disk shaped piece rotatably mounted in a circumferential mounting notch. (TRUE)
    (c) A pusher travels longitudinally through the cartridge carrying member and acts upon the staples to sequentially eject them from the cartridge. (TRUE)

13. Antireflective coating compositions

    (a) A coating layer of a photoresist is formed on a substrate and the photoresist layer is then exposed through a photomask to a source of activating radiation. (TRUE)
    (b) Higher absorbance values for a particular resin can be obtained by decreasing the percentage of chromophore units on the resin. (FALSE)
    (c) Relatively low etch selectivity can be achieved between the organic hard mask layer and the overcoated patterned organic-based resist. (FALSE)

14. Silicon nitride sintered material and production prodess thereof

    (a) Silicon nitride bodies exhibit low strength at high temperature. (FALSE)
    (b) Silicon nitride bodies comprise at least one of the rare earth elements Y, Er, Tm, Yb and Lu. (FALSE)
    (c) Conventionally, silicon nitride - tungsten carbide composite sintered material is used as a wear-resistant member such as a bearing ball or as a material for a heater of a glow plug. (TRUE)

15. Repositionable memory element in a single reel tape cartridge

(a) An ink cartridge for an ink jet printing apparatus has a printhead which ejects ink droplets onto a recording medium and an ink supply needle introduces ink to the printhead. (TRUE)

(b) It is necessary that the memory element is disposed in a position on the surface of the magnetic tape cartridge or inside the magnetic tape cartridge where the memory element does not interfere with the reel. (TRUE)

(c) The cartridge is loaded such that three reference points defining a fixed plane within the housing engages the cartridge in a variable orientation with respect to a head within the drive. (FALSE)

16. Damping arrangements for Y25 bogies

(a) Running velocity may be increased in curves by decreasing the load applied to an inner wheel of the bogie during body tilt operation. (FALSE)

(b) A bogie for wagons of high-speed freight trains includes a frame formed by two side members interconnected centrally for allowing relative angular movements only in horizontal planes. (FALSE)

(c) During a braking operation the torque transmitted it is necessary that the torque tube does not change the annular configuration of that tube. (FALSE)

17. Electrodeless lighting system

(a) A conventional microwave electrodeless lamp is so arranged that the electrodeless lamp is provided in a microwave cavity resonator having an opening with the appendant mesh impenetrable to microwave and a microwave oscillator is linked therewith. (TRUE)

(b) An excimer laser gas in a laser tube is excited by a infra-red wave introduced from a waveguide. (FALSE)

(c) A conventional microwave oven generates microwave energy which is absorbed by water and other molecules in food to make them move at high speeds to create frictional heat which cooks the product evenly in a short space of time. (TRUE)

18. Wear resistant, flame-retardant composition and electric cable covered with said composition.

    (a) Polyphenylene ethers are a class of polymers which are widely used in industry, especially as engineering plastics in applications which require such properties as toughness and heat resistance. (TRUE)

    (b) Dicarboxylic acids which are suitable for use in the preparation of the resins are aliphatic, cycloaliphatic, and/or aromatic dicarboxylic acids. (TRUE)

    (c) Polyphenylene sulfide resins (PPS resins) are engineering plastics with bad heat resistance and flame resistance while having good electric characteristics. (FALSE)

19. On-press exposure and on-press processing of a lithographic material

    (a) Direct-to-plate method bypasses the creation of film because the digital data are transferred directly to a plate precursor by means of a platesetter. (TRUE)

    (b) The average molecular weight of polymers may range from 5,000 to 1,000 g/mol. (FALSE)

    (c) A jet of pressurised water cannot always be used for erasing a lithographic printing master. (FALSE)

20. Nematic liquid crystal device

    (a) A liquid crystal display device comprises a liquid crystal display panel and a refractor disposed on the side of the liquid crystal display panel, opposite from the visible side thereof. (FALSE)

(b) There has been available a conventional liquid crystal display device for displaying three-dimensional information, wherein a plurality of liquid crystal display panels are deposited on one after another. (TRUE)

(c) The interface of the liquid crystal layer and the substrate without the positive alignment process has an alignment regulating force (surface energy) stronger than that of the substrate with the positive alignment process. (FALSE)

21. STABILIZED ALBUMIN PREPARATIONS

(a) Specific examples of the pH controllers include acetic acid-sodium acetate. (TRUE)

(b) The intensity of scattered radiation does not depend on the size of the scattering centers. (FALSE)

(c) A water-soluble, cationic, quaternary ammonium compound can be prepared with a lipophilic end group. (TRUE)

22. Metal coordination compound, luminescence device and display apparatus

(a) Various compounds such as oxadiazole derivatives are used as hole transporting materials. (TRUE)

(b) Aluminum quinolinol complexes are used in the luminescence layer. (TRUE)

(c) The polymeric fluorescent substance is insoluble in organic solvents. (FALSE)

23. Method for producing group III nitride compound semiconductor

(a) Group III nitride compound semiconductor are direct-transition semiconductors exhibiting a wide range of emission spectra from UV to red light. (TRUE)

(b) In a group III nitride compound semiconductor light-emitting device, a light-emitting layer having a portion where an InGaN layer is interposed between AlGaN layers on both sides. (TRUE)

(c) Generally, light-emitting devices using the nitride system III - V compound semiconductor are manufactured by sequentially growing layers made of the nitride system. (TRUE)

24. Control device and method for an electrically driven fan of a vehicle

    (a) Various different techniques have been used in an attempt to flow air through a contained space of a system including air distribution systems for conditioning the temperature of the air with the rate of such air flow being related to the static pressure in the system. (TRUE)
    (b) In a hybrid vehicle wherein the rotation torque of a motor and engine are input to a continuously variable transmission, a target speed ratio is determined from a target engine rotation speed set based on a target drive torque of said vehicle and a vehicle speed. (TRUE)
    (c) Pulse Width Modulation (PWM) is carried out by detecting a refrigerant pressure of the air conditioner and a coolant temperature and calculating a duty ratio of the cooling fan in accordance with the coolant temperature and the refrigerant pressure. (TRUE)

25. Engine

    (a) The cylinder block and the differential are positioned on two different sides of the crankshaft. (FALSE)
    (b) A starter/generator apparatus used with a conventional internal combustion engine has a starter coil and a generator coil which are mounted on the stator of a motor. (TRUE)
    (c) A continuously variable transmission has an endless V belt running across a driving pulley. (TRUE)

Bibliography

Allan, J. (1995). Relevance feedback with too much data. In *SIGIR 1995*, pages 337–343. ACM Press.

Amati, G., Carpineto, C., and Romano, G. (2004). Fondazione Ugo Bordoni at TREC 2004. In *Proceedings of TREC '04*.

Azzopardi, L., Joho, H., and Vanderbauwhede, W. (2010). A survey on patent users search behaviour, search functionality and system requirements. In *IRF Report*.

Baeza-Yates, R. A., Hurtado, C. A., Mendoza, M., and Dupret, G. (2005). Modeling user search behavior. In *LA-WEB*, pages 242–251.

Berger, A. and Lafferty, J. (1999). Information retrieval as statistical translation. In *Proceedings of the SIGIR*, SIGIR '99, pages 222–229, New York, NY, USA. ACM.

Billerbeck, B. and Zobel, J. (2004). Questioning query expansion: An examination of behaviour and parameters. In *ADC 2004*, volume 27, pages 69–76. Australian Computer Society, Inc.

Bishop, C. (2006). *Pattern Recognition and Machine Learning.* Springer.

Blei, D. M., Griffiths, T. L., and Jordan, M. I. (2010). The nested chinese restaurant process and bayesian nonparametric inference of topic hierarchies. *J. ACM*, 57(2).

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022.

Buckley, C., Allan, J., and Salton, G. (1993). Automatic routing and ad-hoc retrieval using SMART: TREC 2. In *TREC*, pages 45–56.

Buckley, C., Salton, G., Allan, J., and Singhal, A. (1994). Automatic query expansion using SMART: TREC 3. In *Overview of the Third Text REtrieval Conference (TREC-3)*, pages 69–80. NIST.

Buckley, C. and Voorhees, E. M. (2000). Evaluating evaluation measure stability. In *Proceedings of the SIGIR*, pages 33–40.

Cao, G., Nie, J.-Y., Gao, J., and Robertson, S. (2008). Selecting good expansion terms for pseudo-relevance feedback. In *SIGIR 2008*, pages 243–250. ACM.

Chaney, A. J.-B. and Blei, D. M. (2012). Visualizing topic models. In *Proceedings of ICWSM*.

Choi, F. Y. Y. (2000). Advances in domain independent linear text segmentation. In *Proceedings of the NAACL 2000*, pages 26–33.

Cleverdon, C. W. (1960). Aslib cranfield research project on the comparative efficiency of indexing systems. In *ASLIB Proceedings*, XII, pages 421–431.

Cleverdon, C. W. (1991). The significance of the cranfield tests on index languages. In *Proceedings of the 14th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '91, pages 3–12, New York, NY, USA. ACM.

Croft, W. B. (1993). Knowledge-based and statistical approaches to text retrieval. *IEEE Expert*, 8(2):8–12.

Cronen-Townsend, S. and Croft, W. B. (2002). Quantifying query ambiguity. In *Proceedings of HLT '02*, pages 104–109.

Deerwester, S. C., Dumais, S. T., Landauer, T. K., Furnas, G. W., and Harshman, R. A. (1990). Indexing by latent semantic analysis. *JASIS*, 41(6):391–407.

Fox, E. A. and Shaw, J. A. (1994). Combination of multiple searches. In Harman, D. K., editor, *The Second Text REtrieval Conference (TREC-2)*, pages 243–252. National Institute for Standards and Technology. NIST Special Publication 500-215.

Ganguly, D., Leveling, J., and Jones, G. J. F. (2011a). Query expansion for language modeling using sentence similarities. In *Proceedings of the IRFC 2011*, pages 62–77.

Ganguly, D., Leveling, J., and Jones, G. J. F. (2011b). United we fall, Divided we stand: A study of Query Segmentation and PRF for Patent Prior Art Search. In *Proceedings of the Patent Information Retrieval Workshop, CIKM 2011*. ACM.

Ganguly, D., Leveling, J., and Jones, G. J. F. (2012a). Cross-Lingual Topical Relevance Models. In *Proceedings of the COLING 2012*.

Ganguly, D., Leveling, J., and Jones, G. J. F. (2012b). Topical Relevance Models. In *Proceedings of the AIRS 2012*.

Ganguly, D., Leveling, J., and Jones, G. J. F. (2013). Topicvis: A gui for topic-based feedback and navigation. In *Proceedings of the SIGIR '13*.

Ganguly, D., Leveling, J., Magdy, W., and Jones, G. J. F. (2011c). Patent Query Reduction using Pseudo Relevance Feedback. In *Proceedings of CIKM 2011*. ACM.

Geman, S. and Geman, D. (1987). *Readings in computer vision: issues, problems, principles, and paradigms*, chapter Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images, pages 564–584. Morgan Kaufmann Publishers Inc.

Griffiths, T. L. and Steyvers, M. (2004). Finding scientific topics. *PNAS*, 101(suppl. 1):5228–5235.

Harman, D. (1993). Overview of the second text retrieval conference (trec-2). In *TREC*, pages 1–20.

Harman, D. and Buckley, C. (2004). The NRRC Reliable Information Access (ria) workshop. In *SIGIR 2004*, pages 528–529, New York, NY, USA. ACM.

Hearst, M. (1997). TextTiling: Segmenting text into multi-paragraph subtopic passages. *Computational Linguistics*, 23(1):33–64.

Hiemstra, D. (2000). *Using Language Models for Information Retrieval.* PhD thesis, Center of Telematics and Information Technology, AE Enschede.

Hiemstra, D. and Kraaij, W. (2005). A language modeling approach for TREC. In Voorhees, E. M. and Harman, D., editors, *TREC: Experiment and Evaluation in Information Retrieval.* MIT press.

Hofmann, T. (1999). Probabilistic latent semantic indexing. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '99, pages 50–57, New York, NY, USA. ACM.

Hull, D. (1993). Using statistical testing in the evaluation of retrieval experiments. In *Proceedings of the 16th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '93, pages 329–338, New York, NY, USA. ACM.

Itoh, H., Mano, H., and Ogawa, Y. (2003). Term distillation in patent retrieval. In *Proceedings of the ACL-2003 workshop on Patent corpus processing - Volume 20*, pages 41–45, Stroudsburg, PA, USA.

Järvelin, K. (2009). Interactive relevance feedback with graded relevance and sentence extraction: simulated user experiments. In *CIKM 2009*, pages 2053–2056. ACM.

Kwok, K., Grunfeld, L., Sun, H. L., and Deng, P. (2004). TREC 2004 robust track experiments using PIRCS. In *TREC*.

Lam-Adesina, A. M. and Jones, G. J. F. (2001). Applying summarization techniques for term selection in relevance feedback. In *SIGIR 2001*, pages 1–9. ACM.

Landis, J. R. and Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33:159–174.

Lang, H., Metzler, D., Wang, B., and Li, J.-T. (2010). Improved latent concept expansion using hierarchical markov random fields. In *Proceedings of CIKM '10*, pages 249–258.

Lavrenko, V. and Croft, B. W. (2001). Relevance based language models. In *SIGIR 2001*, pages 120–127. ACM.

Leveling, J. and Jones, G. J. F. (2010). Classifying and filtering blind feedback terms to improve information retrieval effectiveness. In *RIAO 2010*. CID.

Liu, X. and Croft, W. B. (2004). Cluster-based retrieval using language models. In *Proceedings of the SIGIR '04*, pages 186–193, New York, USA. ACM.

Losada, D. E. (2010). Statistical query expansion for sentence retrieval and its effects on weak and strong queries. *Inf. Retr.*, 13:485–506.

Lovins, J. B. (1968). Development of a stemming algorithm. *Mechanical Translation and Computational Linguistics*, 11:22–31.

Luhn, H. P. (1958). The automatic creation of literature abstracts. *IBM J. Res. Dev.*, 2(2):159–165.

Lv, Y. and Zhai, C. (2010). Positional relevance model for pseudo-relevance feedback. In *SIGIR 2010*, pages 579–586. ACM.

Magdy, W. (2011). *Toward Higher Effectiveness for Recall-Oriented Information Retrieval: A Patent Retrieval Case Study.* PhD thesis, School of Computing, Dublin City University.

Magdy, W. and Jones, G. J. F. (2010a). Applying the KISS principle for the CLEF-IP 2010 prior art candidate patent search task. In *CLEF-2010 (Notebook Papers/LABs/Workshops).*

Magdy, W. and Jones, G. J. F. (2010b). PRES: A score metric for evaluating recall-oriented information retrieval applications. In *SIGIR 2010*, pages 611–618.

Magdy, W., Leveling, J., and Jones, G. J. F. (2010). Exploring structured documents and query formulation techniques for patent retrieval. In *10th Workshop of the Cross-Language Evaluation Forum, CLEF 2009*, pages 410–417.

Magdy, W., Lopez, P., and Jones, G. J. F. (2011). Simple vs. sophisticated approaches for patent prior-art search. In *ECIR*, pages 725–728.

Metzler, D. and Croft, W. B. (2007). Latent concept expansion using markov random fields. In *Proceedings of SIGIR '07*, pages 311–318.

Minka, T. and Lafferty, J. (2002). Expectation-propagation for the generative aspect model. In *Proceedings of the Eighteenth conference on Uncertainty in artificial intelligence*, UAI'02, pages 352–359, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Mitra, M., Singhal, A., and Buckley, C. (1998). Improving automatic query expansion. In *SIGIR 1998*, pages 206–214. ACM.

Murdock, V. (2006). *Aspects of Sentence Retrieval*. PhD thesis, University of Massachusetts - Amherst.

Ogilvie, P., Vorhees, E., and Callan, J. (2009). On the number of terms used in automatic query expansion. *Information Retrieval*, 12(6):666–679.

Piroi, F., Lupu, M., Hanbury, A., and Zenz, V. (2011). CLEF-IP 2011: Retrieval in the intellectual property domain. In *CLEF (Notebook Papers/Labs/Workshop)*.

Ponte, J. M. (1998). *A language modeling approach to information retrieval*. PhD thesis, University of Massachusetts.

Robertson, S. E. (1977). The Probability Ranking Principle in IR. *Journal of Documentation*, 33(4):294–304.

Robertson, S. E. (1990). On Term Selection for Query Expansion. *Journal of Documentation*, 46:359–364.

Robertson, S. E., Walker, S., Jones, S., and Hancock-Beaulieu, M. (1994). Okapi at TREC-3. In *Proceedings of the Third Text REtrieval Conference (TREC 1994)*. NIST.

Rocchio, J. J. (1971). Relevance feedback in information retrieval. In *The SMART retrieval system – Experiments in automatic document processing*. Prentice Hall.

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

Salton, G. and Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Inf. Process. Manage.*, 24(5):513–523.

Salton, G. and McGill, M. (1984). *Introduction to Modern Information Retrieval*. McGraw-Hill Book Company.

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

Singhal, A. (1997). *Term Weighting Revisited*. PhD thesis, Cornell University, Department of computer science.

Singhal, A., Salton, G., Mitra, M., and Buckley, C. (1996). Document length normalization. *Inf. Process. Manage.*, 32(5):619–633.

Sparck-Jones, K. (1973). Index term weighting. *Journal of Documentation*, 9(11):619–633.

Sparck-Jones, K., Walker, S., and Robertson, S. E. (2000). A probabilistic model of information retrieval: development and comparative experiments. *Inf. Process. Manage.*, 36(6):779–840.

Takaki, T. (2005). Query terms extraction from patent document for invalidity search. In *NTCIR-5*.

Takaki, T., Fujii, A., and Ishikawa, T. (2004). Associative document retrieval by query subtopic analysis and its application to invalidity patent search. In *Proceedings of CIKM*, CIKM '04, pages 399–405.

Terra, E. L. and Warren, R. (2005). Poison pills: harmful relevant documents in feedback. In *CIKM 2005*, pages 319–320. ACM.

Tombros, A. and Sanderson, M. (1998). Advantages of query biased summaries in information retrieval. In *SIGIR 1998*, pages 2–10. ACM.

Vakkari, P. (2003). Task-based information searching. *Annual Review of Information Science and Technology*, 37(1):413–464.

Voorhees, E. M. (2004). Overview of the TREC 2004 robust track. In *TREC*.

Wanagiri, M. Z. and Adriani, M. (2010). Prior art retrieval using various patent document fields contents. In *CLEF-2010 (Notebook Papers/LABs/Workshops)*.

Wei, X. and Croft, W. B. (2006). LDA-based document models for ad-hoc retrieval. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '06, pages 178–185, New York, NY, USA. ACM.

Wilkinson, R., Zobel, J., and Sacks-Davis, R. (1995). Similarity measures for short queries. In *In Fourth Text REtrieval Conference (TREC-4)*, pages 277–285.

Xu, J. and Croft, W. B. (1996). Query expansion using local and global document analysis. In *SIGIR 1996*, pages 4–11. ACM.

Xue, X. and Croft, W. B. (2009). Transforming patents into prior-art queries. In *SIGIR*, pages 808–809.

Zhou, D. and Wade, V. (2009). Latent document re-ranking. In *EMNLP '09*, pages 1571–1580.