SEARCH BEYOND TRADITIONAL PROBABILISTIC INFORMATION RETRIEVAL

QINMIN HU

A DISSERTATION SUBMITTED TO THE FACULTY OF GRADUATE STUDIES
IN PARTIAL FULFILMENT OF THE REQUIREMENTS
FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

GRADUATE PROGRAM IN COMPUTER SCIENCE & ENGINEERING
YORK UNIVERSITY
TORONTO, ONTARIO
MARCH 2013

SEARCH BEYOND TRADITIONAL PROBABILISTIC INFORMATION RETRIEVAL

by **Qinmin Hu**

a dissertation submitted to the Faculty of Graduate Studies of York University in partial fulfilment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

SEARCH BEYOND TRADITIONAL PROBABILISTIC INFORMATION RETRIEVAL

by **Qinmin Hu**

By virtue of submitting this document electronically, the author certifies that this is a true electronic equivalent of the copy of the dissertation approved by York University for the award of the degree. No alteration of the content has occurred and if there are any minor variations in formatting, they are as a result of the coversion to Adobe Acrobat format (or similar software application).

Examination Committee Members:

1. Jimmy Huang
2. Nick Cercone
3. Xiaohui Yu
4. Xuwen Chen
5. Ali Asgary
6. Huaiping Zhu

Abstract

This thesis focuses on search beyond probabilistic information retrieval. Three approached are proposed beyond the traditional probabilistic modelling. First, term association is deeply examined. Term association considers the term dependency using a factor analysis based model, instead of treating each term independently. Latent factors, considered the same as the hidden variables of "eliteness" introduced by Robertson et al. to gain understanding of the relation among term occurrences and relevance, are measured by the dependencies and occurrences of term sequences and subsequences. Second, an entity-based ranking approach is proposed in an entity system named "EntityCube" which has been released by Microsoft for public use. A summarization page is given to summarize the entity information over multiple documents such that the truly relevant entities can be highly possibly searched from multiple documents through integrating the local relevance contributed by proximity and the global enhancer by topic model. Third, multi-source fusion sets up a meta-search engine to combine the "knowledge" from different sources. Meta-features, distilled as high-level categories, are deployed

to diversify the baselines. Three modified fusion methods are employed, which are reciprocal, CombMNZ and CombSUM with three expanded versions. Through extensive experiments on the standard large-scale TREC Genomics data sets, the TREC HARD data sets and the Microsoft EntityCube Web collections, the proposed extended models beyond probabilistic information retrieval show their effectiveness and superiority.

Acknowledgements

Foremost, I would like to express my sincere gratitude to my supervisor Prof. Jimmy Huang for the continuous support of my doctoral study and research, for his patience, motivation, enthusiasm, and immense knowledge. His guidance helped me in all the time of research and writing of this thesis. I could not have imagined having a better supervisor and mentor for my graduate study.

Besides my supervisor, I would like to thank the rest of my thesis committee: Prof. Nick Cercone, Prof. Xiaohui Yu, Prof. Xuehua Chen, Prof. Ali Asgary and Prof. Huaiping Zhu, for their encouragement, insightful comments, and hard questions.

I also would like to thank my lab mates in the Information Retrieval and Knowledge Management Lab: Jiashu Zhao, Zheng Ye and Miao Jun for their help when we were working together, and for all the fun we have had in the past six years.

Last but not the least, I would like to thank my husband Laurence Zhu, for his love, understanding and supporting in our life. Furthermore, I would like to welcome my daughter Pearl to this world and thank her to give me more courage to finish this thesis.

Table of Contents

List of Tables

List of Figures

1 Introduction

An information retrieval (IR) system starts with a user's query and aims to find information relevant to the given query, where a query is a statement of an information need and the retrieved information is from the documents of the data collection. In most traditional retrieval systems, queries are translated into query representations. Similarly, documents are converted into document representations. Here in Figure 1.1, a basic IR system is described, in which the IR model proposes to match the query representation against the document representation. The model first computes a numeric score on how well each document representation satisfies the query, and then ranks the document representations according to their scores. Finally, the document representations are recovered to be the original documents for user's reading.

1.1 Research Problems

Ideally, an IR system is supposed to be able to estimate the relevance of a document with absolute certainty to a given query. That is to say that the system is intelligent enough to

predict or know the exact relevance of each document in a collection. Then the system picks up the documents who are truly relevant and shows them to the user. However, during the real retrieval process, the relevance is hidden and difficult to estimate. There are three obstacles in an IR system: (1) most users' information needs are not clear, such as too simple or too many query terms; (2) the nature relevance of a document is actually uncertain, because the users have different opinions about the relevancy, even if the information needs are identified exactly; (3) the answers have to be highlighted and concluded from multiple documents, instead of simply putting documents together.

Probabilistic modelling is the most popular during past fifty years, and has been rather successful to make contributions to the above obstacles. However, there are still some open research problems. First, most probabilistic models are based on the assumption that query terms are independent of each other, such as the famous BM25 functions

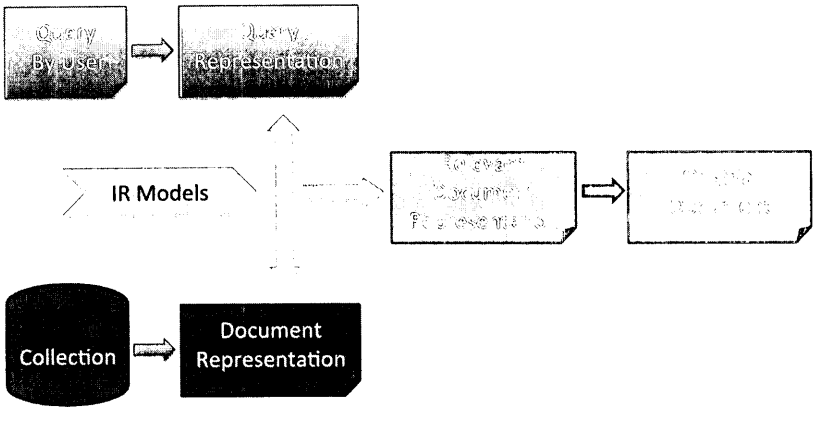

Figure 1.1: A Basic IR System

(Robertson and Walker 1994). Second, the current probabilistic models treat the documents in a collection independently, by calculating the probability for each document. It would be more desirable to include information among and over the documents. Third, probabilistic model deals with the collection based on its own single function, which can be better to consider multiple sources of functions to draw upon each other's strength.

Therefore, three approaches beyond traditional probabilistic models are motived to be proposed, which are term association (Hu et al. 2011a, 2012a), entity ranking (Hu et al. 2012b) and multi-source fusion (Hu et al. 2010, 2011b) respectively. The specific research problems proposed to solve in this research are: (1) term association considers the term dependency using a factor analysis based model, instead of treating each term independently. Latent factors, considered the same as the hidden variables of "eliteness" introduced by Robertson et al. (Robertson and Walker 1994) to gain understanding of the relation among term occurrences and relevance, are measured by the dependencies and occurrences of term sequences and subsequences; (2) entity ranking focuses on the entity level and presents entities over multiple documents. A summarization page is given to summarize the entity information over multiple documents such that the truly relevant entities can be highly possibly searched from multiple documents by integrating local relevance contributed by proximity and global enhancer by topic model; (3) multi-source fusion sets up a meta-search engine to combine the "knowledge" from different sources. Meta-features, distilled as high-level categories, are deployed to diversify the baselines.

Three modified fusion methods are employed, which are reciprocal, CombMNZ and CombSUM with three expanded versions.

1.2 Background Knowledge of Probabilistic Modelling

The literature on the probabilistic approach is by now extensive and densely technical in the past three decades. The first attempts to develop a probabilistic theory of retrieval are Maron and Kuhns (Maron and Kuhns 1960), Miller (Miller 1971). After that, probabilistic models step into a steady development stage, which results in several well known operational systems, such as Okapi (Beaulieu et al. 1997, Huang et al. 2003, Robertson and Jones 1976, Robertson and Walker 1994, Robertson et al. 1982), Indri (Callan et al. 1995, Croft et al. 1983, 1993, Metzler et al. 2004), Lemur (Avrahami et al. 2006, Metzler and Croft 2004, Ogilvie and Callan 2001, Ponte and Croft 1998) and Terrier (Lioma et al. 2004, McCreadie et al. 2012, Ounis et al. 2005, Santos et al. 2010). Some details about probabilistic modelling history can be found in the survey papers of Crestani et. al. (Crestani et al. 1998) and Fuhr (Fuhr 1992).

Probabilistic modelling finds methods for estimating the probabilities to evaluate the probability of relevance that are both theoretically sound and computationally efficient. Compared to vector space model (Salton 1968) in which documents are ranked according to a measure of similarity, probabilistic models are more interpretable and theoretically sound. Compared to the Boolean IR model (Wartik 1992) in which the documents are

searched based on Boolean logic and classical Set theory, probabilistic models provide a solution to computing relevance certainty, instead of shifting the uncertainty problem to the user. In summary, probabilistic models contribute a better solution to relevance certainty. That is, a user assigns relevance judgements to documents with respect to her/his query, and the task of the IR system is to yield an approximation of the set of relevant documents (Fuhr 1992). Furthermore, the approaches formulated by van Rijsbergen (van Rijsbergen 1986, 1992) overcomes the subjective definition of an answer in an IR system by generalizing the proof-theoretic model towards uncertain inference.

The indexing model is first proposed by Maron and Kuhns (Maron and Kuhns 1960), in which they report a technique for literature indexing and searching in a mechanized library system. "Relevance" is taken as the key concept in the theory of IR and is explicated in terms of the theory of probability. This model computes a probability for each document as a measure of how relevant a document will satisfy the given query.

The binary independence retrieval (BIR) model is first proposed by Robertson and Sparck Jones (Robertson and Jones 1976). The basic assumption is that terms are distributed differently within relevant and non-relevant documents, known as the "cluster hypothesis" which has been verified experimentally in (van Rijsbergen and Jones 1973). However, as pointed out by (Cooper 1995), this assumption that actually underlays the BIR model is not that of binary independence but is a weaker assumption of linked dependence (Crestani et al. 1998, Fuhr 1992).

The binary independence indexing (BII) model (Fuhr and Buckley 1991) is a variant of very first probabilistic IR model (Fuhr 1992) after the indexing model (Maron and Kuhns 1960). The major advantage of this model is that the document representations are not specified, but are observed in relation to a number of query terms. The relationship between the query and a document is represented by a binary vector z, where $z = 1$ if a query term q is in that document d, otherwise $z = 0$. As a consequence, the BII model estimates the probability $P(R|z, d)$ instead of $P(R|q, d)$. However, the BII model is hard to be applied in the real practice.

The framework of the Darmstadt Indexing Approach (DIA), as a description-oriented indexing approach, has been developed by (Biebricher et al. 1988, Fuhr 1989). Within the DIA, the indexing task is subdivided in a description step and a decision step. The authors adapt the definition of relevance descriptions to the representations of documents, since the DIA makes no additional assumptions about the choice of the attributes and the structure of relevance. The major advantage of this indexing approach is its flexibility with respect to the representations of documents.

The 2-Poisson model is a more explicit model relating to the elements of the representation, which has been proposed first by (Bookstein and Swanson 1974). This model aims to decide if an index term should be assigned to a document or not, in which two document classes are produced. However, experimental evaluations of this model are only partially successful.

The unified model proposed by Robertson et al. (Robertson et al. 1982), integrates the probabilistic indexing with the binary indexing relevance model. In particular, this model contains four specific models, including the indexing model by Maron and Kuhn (Maron and Kuhns 1960), the BIR model (Robertson and Jones 1976) and another two modified models (Robertson et al. 1982). However, this unified model makes an independence assumption that terms in queries and documents are independent of each other.

Language modelling is used by the speech recognition community to refer to a probability distribution that captures the statistical regularities of the generation of language (Yamron 1997). Later Ponte and Croft (Ponte and Croft 1998, Song and Croft 1999) propose a language approach to IR. The idea is that a document is a good match to a query if the document model is likely to generate the query, which will in turn happen if the document contains the query words. This model thus provides a different realization for document ranking, instead of overtly modelling the probability $P(R = 1|q, d)$ of relevance R of a document d to a query q.

1.3 Term Association

The use of large-scale experimental techniques and domain-dependent tools has increased the pace at which biologists produce useful information. This also promotes the growth of the scientific literature, which contains information on those experimental results in the form of free text that is structured in a way which makes it straightforward for humans

to read but more difficult for computers to interpret automatically and search efficiently. As a consequence, there is increasing interest in methods that can handle collections of free texts. Such methods include systems that efficiently retrieve and classify information in response to complex user queries, and beyond this, systems carry out a deeper analysis of the literature to extract specific associations.

IR deals with text analysis, text storage, and the retrieval of stored records having similarity between them (Salton et al. 1983). In context of biomedical domain, IR systems are to retrieve documents/passages that a user gets relevant to satisfy his or her information need. Many information seekers, really desire to be provided short, specific answers to questions and put them in context by providing supporting information and linking to original sources (Hersh et al. 2005). There are situations when the terms retrieved by IR systems, are not the only desirably independent but associations among the terms within different contexts or a single text, which provide an insight into the text as answers, might be of interest in some specific domains like biomedical domain, text summarization, question answering systems and so on.

Here I focus on discovering term associations among the keywords from a query. Taken all the keywords as a sequence, some subsequences are treated as terms and a factor analysis based model is proposed to provide knowledge for finding the importance of term associations statistically. In the scientific fields, variables such as "intelligence" or "leadership quality" of terms can not be measured directly. Such variables, called latent

variables, can be measured by other "quantifiable" variables, which reflect the underlying variables of interest. Factor analysis attempts to explain the correlations between the observed term associations in terms of the underlying factors, which are not directly observable. These latent factors can be considered the same as the hidden variables of "eliteness" introduced by Robertson et al (Robertson and Walker 1994) in order to gain some understanding of the relation among multiple term occurrences and relevance. The observations for the proposed approach can be obtained from the keywords that are extracted from the queries, and from the passages retrieved by an IR system. In order to find the latent factors for term associations, the factor loadings (Subbaraoand et al. 1995) are computed by MATLAB (Reyment and Joreskog 1996). Then the communalities (Subbaraoand et al. 1995) are calculated based on the factor loadings to indicate the importance and reliance of latent factors. After that, the baseline is re-ranked recursively according to the communalities for improving retrieval performance. In addition, in order to evaluate the superiority of the proposed approach, the generalized sequential pattern (GSP) algorithm is adopted as a comparison.

1.4 Entity Ranking

With tremendous development of the World Wide Web (WWW) as a huge information repository, various semantic information is available for entities, i.e., people, organizations and locations embedded in the static Web pages or Web databases. However, most

existing Web search engines generally treat a whole Web page as the unit for retrieval and consuming. Therefore, entity-oriented search tasks are necessary on the WWW to make users' information needs be better answered by object-level entities instead of any type of documents.

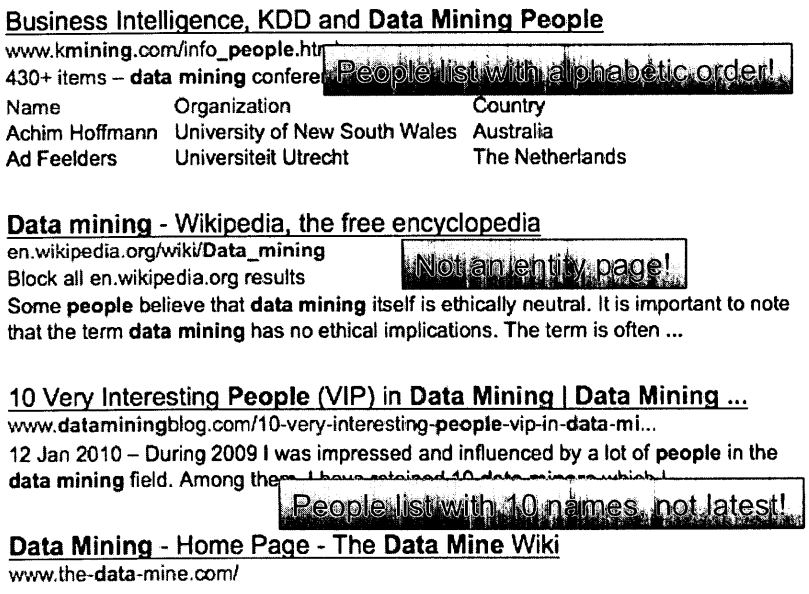

Figure 1.2: Google results, given a query "*data mining people*".

As an example, Figure 1.2 presents the document-level results by Google [1] given a query "*Data Mining People*", where the first and third results are people related. The query gives a task of whom have the best reputations/importances in the data mining field. The first result gives a table where the researchers' information is manually collected and presented with an alphabetic order. The second one is a wikipedia page, instead of an

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

entity page. The third one is to recommend ten interesting persons in 2009.

There are three major barriers in Figure 1.2. First, the entity information (i.e., the persons in the data mining field) potentially spreads out across many Web pages (i.e., both the first and third Web pages) such that the truly relevant entities can not be possibly searched from just a document. Second, the retrieved entities are not ranked properly (i.e., the people are sorted by the alphabetic order in the first Web page). Third, the specific entity information is not targeted (i.e., the general introduction about entity search in the wikipedia page as the second one).

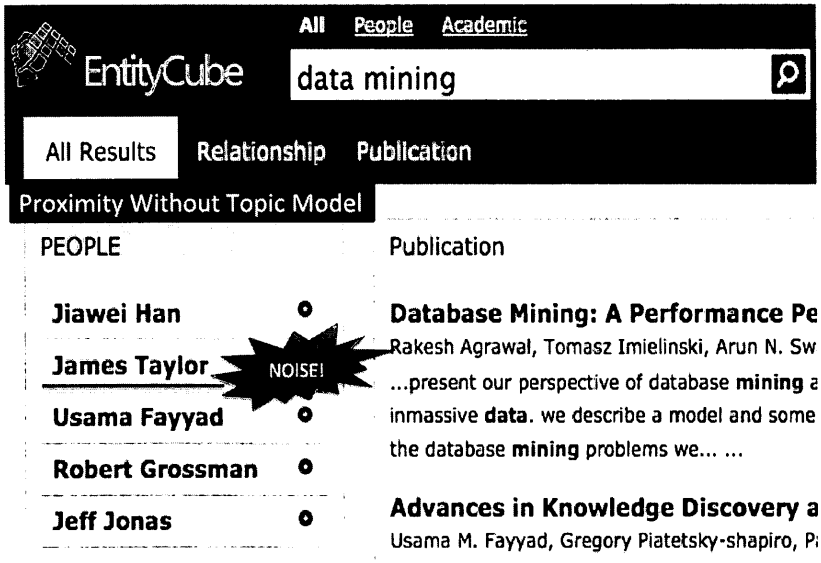

Figure 1.3: EntityCube results, given a query "data mining people".

EntityCube is a solution to solving the above problems, which is an entity search

engine and has been released for public use[2]. It provides summarizations for real-world entities like people, locations and organizations on the Web. In EntityCube, users can ask queries about the entities and explore their relationships. Entity extraction technologies have been proposed in (Zhu et al. 2009) and (Liu et al. 2010a). Here I focus on searching and ranking entities properly given a query.

Relevance is a very important factor on Web search(Blanco and Zaragoza 2010, Na and Ng 2009). This motivates me to consider entity dependency at least. As one of the state-of-the-art dependency methodologies, proximity has been well studied in the past decades (Lv and Zhai 2009, Petkova and Croft 2007, Zhao et al. 2011). Entity proximity considers the relevant impact of an entity on its neighbour entities in a window within a document. Hence, I design an entity-based proximity model with an embedded N-gram model in EntityCube as Figure 1.3, where EntityCube presents the results upon the same query "*Data Mining People*". A list of people are given in the left column and the supportive documents are presented in the right column. For each people, there is a summarization page which summarizes this entity information from multiple Web pages.

However, the second barrier of entity ranking is not well solved. In Figure 1.3, "*James Taylor*" is a noise but is ranked at the second position. This case is deeply analysed as: (1) "*James Taylor*" is a very popular name which is related to over thirteen fields; (2) proximity correctly promotes "*James Taylor*" in the data mining field, but still

[2]http://entitycube.research.microsoft.com

mistakenly promotes "*James Taylor*"s in the physics field and the chemistry field, when "*James Taylor*" is connected in the same windows with "*data*" or "*mining*".

The major disadvantage of proximity is that it works well on a pre-defined window σ (Lv and Zhai 2009, Petkova and Croft 2007, Zhao et al. 2011), but can not perceive the whole picture over a collection of documents. Entity search should be determined by the entity distribution in a collection of documents. Then the overall distribution is motivated to be put into the approach.

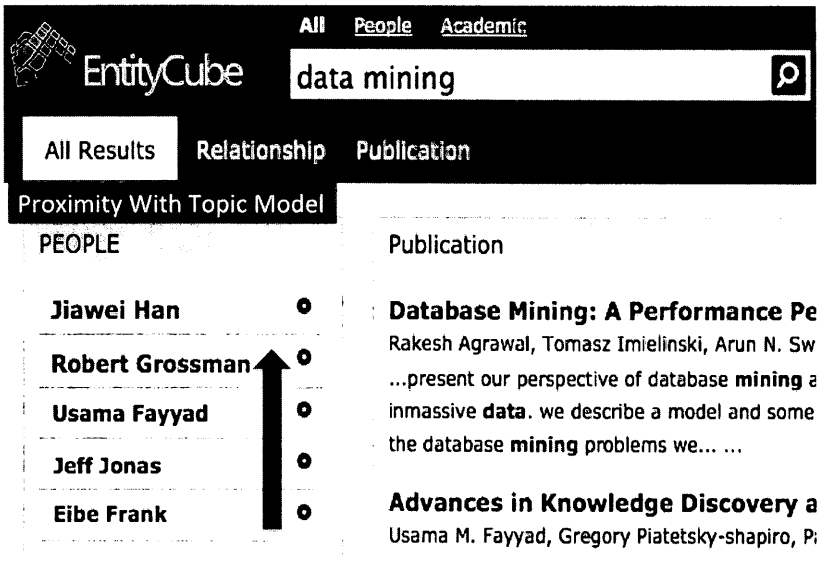

Figure 1.4: EntityCube results, given a query "*data mining people*".

Topic modelling is proposed as a global enhancer to proximity in Figure 1.4. A topic model is a type of statistical model for discovering the abstract "topics" that occur in a collection of documents (Blei et al. 2003). These abstract "topics", known as the hid-

den variables, satisfy the requirement to distinguish the entity's overall distribution in a collection of documents. Therefore, in an entity-based topic model, a document is a bag of entities and modelled as a mixture of hidden topics, the collection of documents is a multivariate distribution. The entities are searched and ranked according to the distribution over the "topics". Given a corpus, the latent "topics" can be obtained by a parameter estimation procedure. In Figure 1.3, topic model first labels "*James Taylor*"s with the hidden "topics", by calculating the probabilities of every "*James Taylor*" given each "topic". Then, same "*James Taylor*"s are connected in a group and different "*James Taylor*"s are disambiguated over the groups. This enhancer discriminates "*James Taylor*"s from different fields so that only those "*James Taylor*"s related to "*data mining*" are ranked. Finally, "*James Taylor*" has been removed from the top five list.

In summary, an entity search approach is proposed to model entities by their relevance and popularity. The relevance is represented by a local relevance and a global enhancer. The local relevance is modelled by proximity with an embedded N-gram model to show how entities related to a query within a pre-defined window. The global enhancer is defined by topic model to simulate the entity distribution probabilistically over the documents. After that, entity-based proximity and topic model are smoothly combined. In addition, popularity is carefully described how to be integrated in the proposed approach.

1.5 Multi-source Fusion

Through the use of IR technologies, information systems retrieve information to index data based on all kinds of pre-defined searching techniques/functions. Each information system has its own models to rank the output. A metasearch system will get access to multiple IR systems and combine their ranking results into a single ranking output generated by the metasearch system. Metasearch systems do not crawl the raw data or maintain a database as most IR systems do, but instead they search several IR systems simultaneously, which act as an agent to pass the query to the search systems and then return the results. Since there are different results retrieved by IR systems/models, metasearch systems provide a quick way to determine which systems are retrieving the best match for information needs.

The major goal of the TREC Genomics Tracks is to create test collections for evaluation of IR and its related tasks in the genomics domain. The users desire to be provided short, specific answers to questions and put them in context by providing linking to original sources from the biomedical literature. This motivates the TREC Genomics Track to implement a new task in 2006 that focuses on passage retrieval using full-text documents from the biomedical literature (Hersh et al. 2006). For the TREC 2006 and 2007 Genomics Track, systems are tasked with extracting out relevant passages of text that answer topic questions and focus on retrieval of short passages (from phrase to sentence

Table 1.1: Meta-Features of Runs

Meta-Feature	Description
FusionIR	fusion - combining results from 2 or more systems regardless of fusion operator used
OkapiIR	passage retrieval using an Okapi BM25 model
TfIdfIR	passage retrieval using a vector space model with any variant of TF-IDF
LmIR	passage retrieval using any language model
DfrIR	passage retrieval using a vector space model with any variant of divergence from randomness (DFR)
...	...

to paragraph in length) that specifically address an information need, along with linkage to the location in the original source document (Hersh et al. 2006, 2007). Here a passage is defined to be a string of characters within a natural paragraph (Hersh et al. 2006). Systems are not only tasked to return passages of text, but also measured on how well they retrieve relevant information at the document-level, aspect-level and passage2-level, which will be presented in the results and discussion section.

In the TREC 2007 Genomics Track, there are a total of 66 runs submitted, in which 49 are classified as automatic. Among the 49 submitted runs, submissions have employed multiple approaches for retrieval processes, such as query expansion, various levels of passage retrieval granularity, and varying IR models with many different scoring schemes. Therefore, meta-features are distilled from the submissions as high-level categories, which are shown in Table 1 (Hersh et al. 2007). For example, "TfidfIR" uses passage retrieval by a vector space model with any variant of TF-IDF (Salton et al. 1983), "OkapiIR" indicates passage retrieval using an Okapi BM25 model (Robertson and Jones 1976, Robertson and Walker 1994), "LmIR" means passage retrieval using a language model, and "FusionIR" combines results from two or more systems regardless of fusion operator usage. This motives me to consider a multi-source fusion approach in a metasearch system to utilize these meta-features. In addition, The performance of NLMFusion, the top scoring automatic run for all three measures (the document-level, the passage2 level and the aspect-level) in 2007 (Hersh et al. 2007), suggests that com-

bining results from different IR models may improve the final results (Hersh et al. 2007).

A robust approach is proposed to combine multiple IR baselines from multiple sources in the genomics domain. First, the proposed approach employs three modified fusion methods, reciprocal, CombMNZ and CombSUM, where CombMNZ is generated into three versions to deeply evaluate this popular combination method. Second, considered the diversity of baselines, it is assumed that the proposed approach in the metasearch system has access to the baselines from three kind of individual models, DFR, BM25 and language model. Therefore, five baselines are selected from the official submissions of the TREC 2007 Genomics Track for combination as the main part of the experiments. Third, in order to evaluate the superiority of the proposed approach, the experiments have been conducted not only on the base runs from different sources, but also on the baselines from a single source of Okapi BM25 with different indices, using the 2007 and 2006 genomics data sets. Fourth, the experimental results demonstrate the viability and superiority of the propose approach with reciprocal to better performance fusion. In addition, as an extension of my preliminary work (Hu et al. 2010), CombSUM is adopted as the third combination method and CombMNZ is further evaluated by considering its normalization, assigned weights and multiple times application.

1.6 Main Contributions

The major contribution of this research work is that diversified approaches have been proposed beyond classical probabilistic modelling, which enriches the developments of traditional probabilistic models and deploys probabilistic ranking principle into multiple domains. All the experimental results show the proposed approaches achieve good performance.

In the term association appraoch, the "eliteness" theory of Robertson at el. is revisited. The assumption of term independence has been relaxed and the informative content has been well measured by term association which evaluates the latent variables of "leadership quality" among the keyword sequences and subsequences as "eliteness". Term association considering co-occurrence and dependency among the keywords produces better results than the baselines which treat the keywords independently. In the other hand, the unigrams, bigrams and trigrams are terms independently computed by the factor analysis based model, which means that the trigrams are not dependent on the bigrams' importance, and the bigrams are not dependent on the unigrams' importance. Their importance is decided by the model and the appearances in the passages. This is also confirmed by the GSP algorithm.

In the entity ranking approach, EntityCube is a well-built entity search engine and has been released for public use, not only with its sophisticated entity extraction tech-

nologies, but also with its innovative searching and ranking approach. The effectiveness of the position information in topic model has been evaluated, since topic model treats a document as a bag of words without orders. The experimental results can be duplicated by using EntityCube. Fourth, named disambiguation is enhanced by the proposed approach.

In the multi-fusion appraoch, compared to the CombMNZ and CombSUM methods, the reciprocal method provides notable improvements using the baselines from a DFR model, a BM25 model and a language model respectively. While CombMNZ does not achieve good performance, three versions as CombMNZ-with-normalization, CombMNZ-with-assigned-weight and CombMNZ-with-multiple, have been conducted to further improve and evaluate the CombMNZ method. Although the CombSUM method does not work as well as reciprocal, CombSUM makes progress on the passage2-level, also works better than CombMNZ on all the three versions. In addition, the proposed robust approach makes improvements not only for combining the baselines from different sources, but also for combining the baselines from the single source such as Okapi BM25.

1.7 Outline

The thesis consists of six chapters, appendices and a list of references. The chapters are organized as follows. Chapter 1 presents the introduction and the research problems I

propose to solve, where term association, entity ranking and multi-source fusion are proposed respectively, motivation and contributions are highlighted. Then in Chapter 2, the related work of traditional probabilistic models, factor analysis, topic model, proximity and meta-search engine is described. In the following three chapters, the term association approach is proposed in Chapter 3, a proximity-based entity retrieval with topic model in Chapter 4 and a multi-source fusion approach in Chapter 5. Particularly, the experimental settings, experiment results and discussions are shown respectively in each chapter as well. Finally, the conclusions are drawn in Chapter 6. The appendices are presented as sample data, sample topics, evaluation measures and scripts etc.

2 Literature Review

The related work is briefly reviewed on probabilistic modelling, other IR modelling, term association, entity work, multi-source fusion related. Note that some familiarity with principles of probability theory is assumed on the part of the readers.

2.1 Probabilistic Modelling

2.1.1 "Relevance" vs "Probabilistic Relevance"

In the dictionary, a definition of "relevance" is "closely connected or appropriate to the matter at hand". In the Computer Science (CS) and Library Science (LS) fields, another definition to "relevance" by Hjørland, Birger and Christensen (Hjørland and Christensen) is that "something (A) is relevant to a task (T) if it increases the likelihood of accomplishing the goal (G), which is implied by T".

In the IR field, a widely accepted definition is a relationship that may or may not hold between a document and a user of the IR system who is searching for some information(Crestani et al. 1998). Relevance (R) is a relationship between a document d_j

and a query q. If the user wants the document d_j in terms of her/his query q, then d_j is relevant R to q.

In the probabilistic modelling, the relevance is a computable function of multiple variables based on the document, the user and her/his information need. The probabilistic models introduced in the following sections explain what evidence is available and how to estimate the probability of relevance $P(R|q, d_j)$. Note that every model presents a different definition such that here a general relevance definition above all the models can not be given.

2.1.2 Probability Space

In the IR probabilistic modelling, the probability space is built following the probability theory. The collection is the sample space Ω. The event space is the set of $Q \times D$, in which Q represents a set of queries and D a set of all the documents in the collection. An element in the event space is a query-document pair (d_j, q_k). Every single element (d_j, q_k) is associated with a relevance judgement $r(d_j, q_k) \in R$. Note that all the probabilistic models use the same probability space. The difference among them is that different models use different document and query representations.

A query is a description of an information need. So every single query is treated as a unique event, which means that there are two queries if two different users submit the same query, and the same query has been submitted by the same user at different

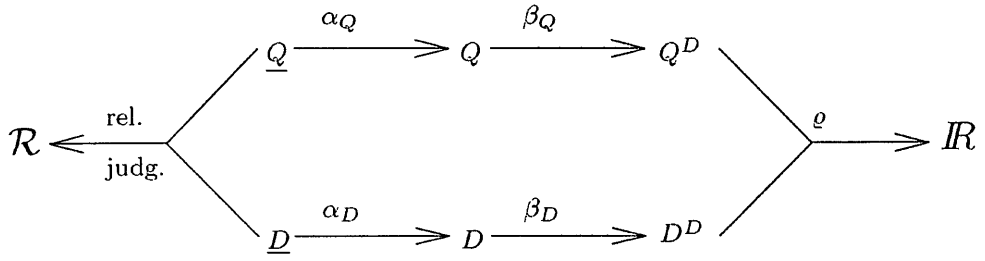

Figure 2.1: A Conceptual Model

occasions. A document is an object containing information, which can be a piece of text, an image, a video or a sound. In this research work, I focus on text such that only text-based IR systems are discussed.

There is a strong assumption lying in most of the traditional probabilistic models (Robertson 1977) that the relevance judgements for different documents with respect to the same query are independent of each other. Another interpretation of the probability of relevance judgement $P(R|d_j, q_k)$ is that the pair (d_j, q_k) has the same representations in the event space.

2.1.3 A Conceptual Model for IR

The importance of a conceptual model is widely recognized in fields such as IR, database and knowledge management. Fuhr's conceptual model (Fuhr 1992) is adopted here to be a conceptual basis for all the probabilistic models in the follwoing sections, since it is simple and general enough.

Fuhr's conceptual models is shown in Figure 2.1. d_j and q_k denote the representations of a raw document $\underline{d_j}$ and a raw query $\underline{q_k}$. As we mentioned in the previous section, every query is unique, no matter whether the queries are same. In the event space, a mapping r for relevance has been defined as $Q \times D \rightarrow R$ (Bookstein and Fouty 1976).

Different IR models may have different representations, such as that the document representation is a list of words and the query representation is a Boolean express in Boolean system. To make the conceptual model more flexible and general to deal with more IR models, a further mapping is introduced as β_Q and β_D, which makes Q and D map to Q^D and D^D respectively.

Therefore, relevance r is further defined as $Q^D \times D^D \rightarrow R$. The task of ranked retrieval IR systems given a query q_k is then to compute $r(d_j^D, q_k^D)$ and rank each d_j and $\underline{d_j}$ in the collection.

2.1.4 The Probability Ranking Principle

The probability ranking principle (PRP) (Robertson 1977) asserts that optimum retrieval can be achieved on probabilistic models when documents are ranked according to their probabilistic relevance to a query. Optimum retrieval is proposed based on the following reasons: (1) it is more suitable compared to perfect retrieval; (2) it can be defined precisely since it has been proved theoretically with respect to representations of documents and query.

Mathematically, the formal definition of the PRP is Equation 2.1.

$$C \cdot P(R|q_k, d_j) + \bar{C} \cdot (1 - P(R|q_k, d_j)) \leq C \cdot P(R|q_k, d_m) + \bar{C} \cdot (1 - P(R|q_k, d_m)) \quad (2.1)$$

where C denotes the cost of retrieving a relevant document and $\bar{C}$ the cost of retrieving an irrelevant document, d_j and d_m are two different document candidates.

The decision rule is that d_j should be retrieved in response to q_k above any document d_m in the collection if Equation 2.1 is satisfied. In other words, a document d_j is retrieved when the expected costs (C and $\bar{C}$) of retrieval are a minimum. For most of time, $C < \bar{C}$ holds, Equation 2.1 is equivalent to

$$P(R|q_k, d_j) \geq P(R|q_k, d_m) \quad (2.2)$$

Therefore, the documents should be ranked according to their decreasing probability of being relevant. The PRP can be extended to deal with multi-valued relevance scales (Bookstein 1985, Cooper).

2.1.5 The Binary Independence Retrieval Model

The binary independence retrieval (BIR) model (Robertson and Jones 1976) is a fairly simple one, whereas its precise assumptions are developed throughout most of the probabilistic models. In the BIR model, a document d_j is judged relevant to a given query q_k concerning to d_j's probability $P(R|d_j, q_k)$. The basic assumption, called the "cluster

hypothesis", is that terms (which are defined as non-trivial words) are distributed differently within relevant R and irrelevant $\bar{R}$ documents. This assumption has been proved experimentally by Rijsbergen and Jones (van Rijsbergen and Jones 1973).

Denote $T = \{t_1, ..., t_n\}$ as the set of terms in the collection, a binary vector $x = (x_1, ..., x_n)$ as the document d_j^T where d_j is represented by the set of terms T with $x_i = 1$, if $t_i \in d_j^T$ and otherwise, $x_i = 0$.

Therefore, the BIR model computes $P(R|x, q_k)$, instead of $P(R|d_j, q_k)$ for a document d_j. Note that different documents having the same term distributions yield the same estimation of probability $P(R|x, q_k)$. Additionally, all the terms in the queries q are assumed to satisfy $q_k^T \subset T$.

Using Bayes's theorem, the probability of relevance is:

$$P(R|x, q_k) = \frac{P(R|q_k) \cdot P(x|R, q_k)}{P(x|q_k)} \tag{2.3}$$

To simplify the notations, q_k is omitted in term of the understanding that evaluations are with respect to a given query q_k. Equation 2.3 becomes

$$P(R|x) = \frac{P(R) \cdot P(x|R)}{P(x)} \tag{2.4}$$

where $P(R)$ is the prior belief of relevance, $P(x|R)$ is the probability of observing the document representation x conditioned on relevance having been observed, and $P(x)$ is the probability that x is observed.

In order to further simply the estimation process, the components of the vector x

are assumed to be stochastically independent when conditionally dependent on R and $\bar{R}$. This is that the joint probability $P(x|R)$ is given by the product of the marginal probability distributions in Equation 2.5.

$$\begin{aligned} P(d_j|R) &= P(x|R) = \prod_{i=1}^{n} P(x_i|R), \\ P(d_j|\bar{R}) &= P(x|\bar{R}) = \prod_{i=1}^{n} P(x_i|\bar{R}), \\ P(R) &+ P(\bar{R}) = 1 \end{aligned} \tag{2.5}$$

This binary independence assumption is the basis of the model proposed by Robertson and Sparck Jones (Robertson and Jones 1976). However, as it has been pointed out by Cooper (Cooper 1995) that the assumption underlying the BIR model is not that binary independence, but rather the weaker assumption of linked dependence of the form:

$$\frac{P(d_j|R)}{P(d_j|\bar{R})} = \frac{P(x|R)}{P(x|\bar{R})} = \prod_{i=1}^{n} \frac{P(x_i|R)}{P(x_i|\bar{R})} \tag{2.6}$$

Equation 2.6 states that the ratio between the probabilities of x occurring in relevant and irrelevant documents is equal to the product of the corresponding ratio of every single term. Upon Equation 2.6, the transformation can be done as

$$\begin{aligned} \frac{P(R|x,q_k)}{P(\bar{R}|x,q_k)} &= \frac{P(R|q_k)}{P(\bar{R}|q_k)} \cdot \frac{P(x|R,q_k)}{P(x|\bar{R},q_k)} \\ &= \frac{P(R|q_k)}{P(\bar{R}|q_k)} \cdot \prod_{i=1}^{n} \frac{P(x_i|R,q_k)}{P(x_i|\bar{R},q_k)} \\ &= \frac{P(R|q_k)}{P(\bar{R}|q_k)} \cdot \prod_{i=1}^{n} \frac{P(x_i=1|R,q_k)}{P(x_i=1|\bar{R},q_k)} \cdot \prod_{i=1}^{n} \frac{P(x_i=0|R,q_k)}{P(x_i=0|\bar{R},q_k)} \end{aligned} \tag{2.7}$$

To further more simplify $\frac{P(R|x,q_k)}{P(\bar{R}|x,q_k)}$, denote $p_{ik} = P(x_i = 1|R, q_k$ and $q_{ik} = P(x_i = 1|\bar{R}, q_k$, and assume $p_{ik} = q_{ik}$ for the terms not occurring in the set q_k^T. Then Equation 2.8 is derived

$$\begin{aligned}\frac{P(R|x,q_k)}{P(\bar{R}|x,q_k)} &= \frac{P(R|q_k)}{P(\bar{R}|q_k)} \cdot \prod_{t_i \in d_j^T \bigcap q_k^T} \frac{p_{ik}}{q_{ik}} \cdot \prod_{t_i \in q_k^T \; d_j^T} \frac{1-p_{ik}}{1-q_{ik}} \\ &= \frac{P(R|q_k)}{P(\bar{R}|q_k)} \cdot \prod_{t_i \in d_j^T \bigcap q_k^T} \frac{p_{ik}(1-q_{ik})}{q_{ik}(1-p_{ik})} \cdot \prod_{t_i \in q_k^T} \frac{1-p_{ik}}{1-q_{ik}}\end{aligned} \tag{2.8}$$

For the parameter estimation of p_{ik} and q_{ik}, relevance feedback is the solution. A scenario is assumed that the user is asked to judge the relevance of those retrieved documents by an IR system given a query q_k. Based on the user's relevance feedback data, the parameters are estimated. Mathematically, the scenario is represented as follows. First, let f denote the number of documents provided to the user, r documents are judged as relevant by the user. Second, f_i is the number among the f documents where the term t_i occurs, and r_i is the number of relevant documents where t_i occurs. Third, the estimates are obtained as

$$\begin{aligned} p_{ik} &\approx \frac{r_i}{r} \\ q_{ik} &\approx \frac{f_i - r_i}{f - r} \end{aligned} \tag{2.9}$$

2.1.6 The Binary Independence Indexing Model

The binary independence indexing (BII) model (Fuhr and Buckley 1991) is a variant of the BIR model. The major difference between the BII model and the BIR model is that

the BII model regards a document in relation to a number of queries, whereas the BIR model regards a single query in relation to the entire document collection.

The motivated idea of the indexing in the BII model comes from the Maron and Kuhns's indexing model (Maron and Kuhns 1960). The indexing weight of a term is evaluated as an estimate of the probability of relevance of that document with respect to queries using that term.

Since the BII model focuses on a set of queries, query representation is more important than document representation. Assume a binary vector z, whose dimension depends on the set of all terms T, and $z_i = 1$ if the term t_i occurs in the query, otherwise $z_i = 0$. Note that the term weights are defined in terms of frequency information derived from queries, and an explicit document representation is not required.

Unlike the BIR model which computes $P(R|x, q_k)$, the BII model seeks to estimate $P(R|z, d_j)$ that the document d_j is judged relevant to the query representation z. x is still employed here to represent the document. Then, Bayes' theorem is applied:

$$P(R|z, x) = \frac{P(R|x) \cdot P(z|R, x)}{P(z|x)} \tag{2.10}$$

where $P(R|x)$ is the probability that the document representation x is judged relevant to a query, $P(z|R, x)$ is the probability that the document is relevant to a query representation z, $P(z|x)$ is reduced to be $P(z)$ since z and x are assumed to be independent on each other.

Inheriting the classic binary independence assumption, the conditional distribution

$P(z|R,x)$ becomes

$$P(z|R,x) = \prod_{i=1}^{n} P(z_i|R,x) \tag{2.11}$$

Similar to the BIR model, q_k is omitted in term of the understanding that evaluations are with respect to a given query q_k, which means z can be omitted as $P(R|z_i,x) = P(R|x)$. Therefore, Equation 2.10 arrives

$$\begin{aligned} P(z|R,x) &= \frac{P(R|x)}{P(z|x)} \cdot \prod_{i=1}^{n} P(z_i|R,x) \\ &= \frac{\prod_i P(z_i)}{P(z)} \cdot \prod_{i=1}^{n} \frac{P(R|z_i,x)}{P(R|x)} \\ &= \frac{\prod_i P(z_i)}{P(z)} \cdot P(R|x) \cdot \prod_{z_i=1} \frac{P(R|z_i=1,x)}{P(R|x)} \cdot \prod_{z_i=0} \frac{P(R|z_i=0,x)}{P(R|x)} \end{aligned} \tag{2.12}$$

2.1.7 The Darmstadt Indexing Model

The Darmstadt indexing approach (DIA) model is a description-oriented indexing approach (Biebricher et al. 1988, Fuhr 1989). Its basic idea is to apply the third learning strategy where features of terms in documents are regarded instead of the term-document pairs. The DIA model attempts to estimate specific index-term features based on the use of index terms in the learning sample. In other words, it aims to estimate $P(R|d_j,q_k)$ from a sample of relevance judgements of query-document/term-document pairs.

The indexing stage is subdivided in two steps, which are a description step and a decision step.

In the description step, relevance descriptions for term-document pairs t_i, d_j are

formed. These relevance descriptions $x(t_i, d_j)$ comprise values of attributes of (t_i, d_j, R). The DIA model makes no assumptions about the structure of the function x or about the choice of attributes. Some possible attributes can be defined as: (1) frequency of occurrence of term t_i in the document d_j; (2) inverse document frequency of t_i in the collection; (3) position information of term t_i in the document; (4) parameters describing the document, such as the document length.

In the decision step, a probabilistic index weight based on the previous relevance descriptions is assigned. This means that $P(R|x(t_i, d_j)$ is estimated, instead of $P(R|t_i, d_j)$. In order to estimate $P(R|t_i, d_j)$, each single document d_j is regarded with respect to all queries containing term t_i. Currently the DIA model regards the set of all query-document pairs in which the same relevance description x occurs in order to compute $P(R|x(t_i, d_j)$.

The probabilistic index term weights $P(R|x(t_i, d_j)$ are derived from a learning example $L \subset \underline{Q} \times \underline{D} \times R$ of query-document pairs for relevance judgements such that $L = \{(d_j, q_k, r_{jk})\}$. By forming relevance descriptions for the terms common to queries and documents for each query-document pair in L, a bag of relevance descriptions with relevance judgements are defined as

$$L^x = [(x(t_i, d_j), r_{jk})|t_i \in q_k^T \cap d_j^T \wedge (d_j, q_k, r_{jk}) \in L] \tag{2.13}$$

With above set, $P(R|x(t_i, d_j)$ can be estimated as the relative frequency of those elements of L^x with the same relevance descriptions. Nevertheless, the technique in the DIA

model makes use of an indexing function, since it provides better estimates through additional assumptions about the indexing function. Fuhr and Buckley (Fuhr and Buckley 1991) used various linear indexing functions estimated by least-squares functions, and they (Fuhr et al. 1993) again attempted a logistic indexing function estimated by maximum likelihood. Their experiments conducted on the standard test collections indicate that the DIA model is superior to other indexing methods.

2.1.8 The N-Poisson Model

The N-Poisson indexing model is an extension to n dimensions of the 2-Poisson model first proposed by Bookstein et al. (Bookstein and Swanson 1974).

The 2-Poisson model is based on the assumption that it is possible to decide if a term should be assigned to a document by determining to which of the following two distributions the term belongs, where the two distributions are: (1) the number of occurrences of a term within a document is different depending on whether the document is relevance; (2) the number of occurrences of the term can be modelled using a know distribution. The 2-Poisson model resulted from a search for the statistical distribution of occurrence of potential index terms in a collection of documents.

To extend the 2-Possion model to the n-dimensional case, suppose there are n classes of documents in which the term t_i appears with different frequencies according to the extent of coverage of the topic related to that specific term.

2.2 Other Related IR Modelling

2.2.1 The Vector Space Model

The vector space model (Salton 1968) ranks the documens according to a measure of similarity. One of the first IR experimental systems is the SMART system (Salton and Lesk 1965). Dcouments d and queries q are represented as term t vectors: $d_j = (t_{1,j}, t_{2,j}, ..., t_{k,j})$ and $q = (t_{1,q}, t_{2,q}, ..., t_{k,q})$. Dimensions depend on k. The basic matching theory is that a document's value in the vector is non-zero, if at least a term occurs in it.

Based on the assumption of document similarity theory, document relevance and ranking are computed by the cosine function of vector d and q in practice, where $d, q \in R^V$ and R^V is the product space of the vocabulary V. The cosine is the inner productin Equation 2.14, in which $|| * ||$ is the norm of the vector. The norm of vector q is $||q|| = \sqrt{\sum_{i=1}^{n} q_i^2}$.

$$, sim(q, d) = \frac{(q, d)}{||q||||d||} \tag{2.14}$$

Equaiton 2.14 shows that all vectors are elementwise nonnegative so that a cosine value of zero means that the query and document vectors are orthogonal and have no match.

Salton at el. (Salton et al. 1983) proposed a classic vector space model, known as term frequency-inverse document frequency (tf-idf) model. Therefore, the weight vector

for document d is

$$w_{t,d} = tf_{t,d} log \frac{|D|}{|\{d' \in D | t \in d'\}|} \tag{2.15}$$

where $tf_{t,d}$ is term frequency of term t in document d, $log \frac{|D|}{|\{d' \in D | t \in d'\}|}$ is inverse document frequency, $|D|$ is the total number of document in the document set and $|\{d' \in D | t \in d'\}|$ is the number of documents containing term t.

Adopting the cosine similarity function of Equation 2.14, the relevance ranking of document d_j and query q can be

$$sim(q, d_j) = \frac{(q, d_j)}{||q|| ||d_j||} = \frac{\sum_{i=1}^{N} w_{i,j} w_{i,q}}{\sqrt{\sum_{i=1}^{N} w_{i,j}^2} \sqrt{\sum_{i=1}^{N} w_{i,q}^2}} \tag{2.16}$$

Length normalization is one of the major limitations of the vector space model, since the longer document receives a smaller similarity score if all term frequencies are the same in both documents. The cosine function for normalization requires heavy computation in implementation. Hence, the BM25 formula abandons the normalization and considers the document length as an explicit random variable of the probability space (Robertson and Walker 1994, Zhou et al. 2011).

2.2.2 Inference Network Retrieval

As a probabilistic formalism for inference networks with uncertainty, Bayesian inference networks have been described by Turtle and Croft in (Turtle and Croft 1990). They applied the use of inference network to support document retrieval. IR is treated to be an

inference or evidential reasoning process in which the probability is estimated such that a user's information need is met given a document as "evidence".

The basic inference network model consists of two components: a document network and a query network. The document network is built once for a given collection and its structure does not change during query processing. Meanwhile, the query network is built for each information need and is modified during query processing as existing queries are refined or new queries are added in an attempt to better characterize the information need. The document and query networks are joined by links between representation concepts and query concepts. All nodes in the inference network take on values from the set {false, true}(Turtle and Croft 1990).

The retrieval inference network tries to capture all of the significant probabilistic dependencies among the variables represented by nodes in the document and query networks. Similar to the probability ranking principle (Robertson and Jones 1976), the inference network treats documents in isolation, instead of looking for the subset of documents which produce the highest probability that meet the information need.

The directed arcs of the document and query networks indicate probabilistic dependence of nodes. The probability of a node being true depends only on the values of its parents.

A link matrix is introduced to be a 2×2^n size for n parents and specify the probability that a node Q takes the value $Q = true$ or $Q = false$ for all combinations of parent

values. Bayesian networks are employed to compute the predictive component of the belief in Q or $P(Q = true)$. Turtle and Croft use the following formula for assigning the posterior probability $P(t_k|d_j)$:

$$\begin{aligned} P(t_k|d_j) &= \gamma + \delta \cdot idfn(t_k) \cdot tfn(t_k, d_j) \\ P(\neg t_k|d_j) &= \delta(1 - idfn(t_k) \cdot tfn(t_k, d_j)) \\ idfn(t_k) &= \frac{log\frac{N}{n_{t_k}}}{logN} \\ \gamma + \delta &= 1 \\ tfn(t_k, d_j) &= \frac{tf(t_k, d_j)}{arg_{d \in D} maxtf(t_k, d)} \end{aligned} \tag{2.17}$$

where t_k is the k^{th} term and d_j is the j^{th} document, $idfn(t_k)$ is the normalized iverse document frequency, $tfn(t_k, d_j)$ is the normalized term frequency of a term in a document.

Beyond the idea of combining evidence from multiple sources, inference network first introduced the use of a non-zero default probability for term weight. Language model holds the same feature.

2.2.3 The Basic Language Model

The first language model approach is proposed by Ponte and Croft (Ponte and Croft 1998), which offers a uniform approach to both indexing and weighting schemes, while the standard probabilistic modelling uses two different models. Unlike the probabilistic

relevance in the BM25 formula defined as a hidden variable, this approach by Ponte and Croft starts from the "raw" maximum likelihood of terms in the given document as the "model" of the language:

$$\begin{aligned} P(t|d) &= P_d(t) = \frac{tf}{l_d} \\ p(q|d) &= \prod_{t \in q} P(t|d) \end{aligned} \tag{2.18}$$

Then, for those terms which do not occur in the document d, their probabilities are set by their raw term frequency $tf_{t,d}$ in the collection D as the default values:

$$P(t|d) = \frac{tf_{t,d}}{tf_{t,D}} \tag{2.19}$$

To sum up all the probabilities in Equation 2.18 and 2.19, since there is a fundamental condition that $\sum_{t \in D} P(t|d) = 1$, a normalization process is needed. Hence, Ponte and Croft adopt a risk probability function $\hat{R}$ in Equation 2.20.

$$\hat{R}_{t,d} = \left(\frac{1}{1 + \bar{f}_{t,d}} \right) \times \left(\frac{\bar{f}_{t,d}}{1 + \bar{f}_{t,d}} \right)^{tf} \tag{2.20}$$

where $\bar{f}_{t,d} = \bar{P}(t|d) \times l_d$ and $\bar{P}(t|d)$ is a probability established with a larger estimation which considers the set D_t of documents only containing term t in Equation 2.21

$$: \bar{P}(t|d) = \frac{1}{n_t} \sum_{d \in D_t} P(t|d) \tag{2.21}$$

Therefore, the normalization can be:

$$P(q|d) = \begin{cases} P_d(t)^{1-\hat{R}_{t,d}} \times \bar{P}(t|d)^{\hat{R}_{t,d}} & \text{if } tf > 0 \\ \frac{tf_{t,d}}{tf_{t,D}} & \text{otherwise} \end{cases} \tag{2.22}$$

The risk function aims to enable a choice of probability close to either the value of the maximum likelihood or to the mean frequency of the term, based on the size of the relative term frequency. If $tf_{t,d}$ is high, the risk $\hat{R}_{t,d}$ is minimal and $P_d(t)$ can be reduced to its maximum. On the contrary, if $tf_{t,d}$ is small, the maximum likelihood estimate is less reliable and then the risk function $\hat{R}_{t,d}$ is high, the probability $P_d(t)$ reduces mainly to the mean frequency in the set D_t where all documents contain term t. Similarly, if the length l_d is large, the maximum likelihood is more reliable and then the risk function $\hat{R}_{t,d}$ becomes small, $P_d(t)$ reduces to its maximum in the document. This also explains that the language model is better for long documents instead of short text.

2.3 Term Association

Modelling and mining term association is important for information retrieval, which allows an IR system given a user's query terms to retrieve relevant documents more precisely.

2.3.1 Term Dependency

Bendersky and Croft (Bendersky and Croft 2012) proposed a retrieval framework that modeled higher-order term dependencies, i.e., dependencies between arbitrary query concepts rather than just query terms. In order to model higher-order term dependencies, the authors represented a query using a hypergraph structure – a generalization of a

graph, where a (hyper)edge connected an arbitrary subset of vertices. A vertex in a query hypergraph corresponded to an individual query concept, and a dependency between a subset of these vertices was modelled through a hyperedge.

Hai et al. (Hai et al. 2012) proposed a generalized approach to opinion feature extraction by incorporating robust statistical association analysis in a bootstrapping framework. This approach started with a small set of feature seeds, on which it was iteratively enlarged by mining feature-opinion, feature-feature, and opinion-opinion dependency relations. Two robust bootstrapping approaches, LRTBOOT and LSABOOT, were utilized on a real-life reviews crawled from the cellphone and hotel domains.

Shi and Nie (Shi and Nie 2010) proposed a dependency model to integrate term dependencies. In their model, pair of terms is assigned a different weight of dependency according to their utility to IR, and each term dependency was weighted according to its strength and possible impact on the retrieval effectiveness. The main task was to determine the strength and impact. A learning process was applied through using a set of features.

Metzler and Croft (Metzler and Croft 2005) developed a general, formal framework for modelling term dependencies via Markov random fields. They not only made used of features based on occurrences of single terms, ordered phrases, and unordered phrases, but also explored full independence, sequential dependence and full dependence variant of the model. In addition, the training data were needed in the model for the parameters.

Their ad hoc retrieval experiments showed improvements by modeling dependencies, especially on the larger collections.

Deerwester et al. (Deerwester et al.) proposed an approach to automatic indexing and retrieval, which was to take advantage of implicit higher-order structure in the association of terms with documents in order to improve the detection of relevant documents on the basis of terms found in queries. The proposed approach tried to overcome the deficiencies of term-matching retrieval by treating the unreliability of observed term-document association data as a statistical problem. They assumed that some underlying latent semantic structure in the data was obscured by the randomness of word choice with respect to retrieval. Then, they use statistical techniques to estimate the latent semantic structure for indexing and retrieval.

Grefenstette (Grefenstette 1992) proposed an extraction technique using coarse syntactic analysis without domain knowledge, which produced word associations as lists of words related to the work appearing in a corpus. Their experimental results confirmed that, when the closest related terms were used in query expansion of a standard information retrieval data set, the results were much better than that given by document co-occurrence techniques, and slightly better than using unexpanded queries.

Hiroyuki Kaji et al. (Kaji et al. 2000) presented a method for automatically generating a corpus-dependent association thesaurus from a text corpus. This method consisted of extracting terms and co-occurrence data from a corpus and analysing the correlation

between terms statistically. They conducted the experiments on a newspaper article corpus, which proved that the thesaurus navigator efficiently explored information through a text corpus when the information needs were vague.

Manna and Gedeon (Manna and Gedeon 2008) proposed a term association model which extracted significant terms as well as the important regions from a single document, which based on the subjective data analysis without predefined knowledge. They claimed that the model overcame the basic drawback of existing language models for choosing significant terms in single documents.

In this work, we propose a term association approach to customize a factor analysis based model to quantify the importance and reliance of term associations. Independent keywords, disordered dependent phases and high-order structure are considered at the same time in the proposed approach. In addition, we focus on the appearance of the terms at the same context statistically but not the distance among the terms.

2.3.2 FA vs PCA vs N-Gram

As a popular analysis method, factor analysis is attractive in IR for two main reasons. One apparent advantage of factor analysis is that users can use it to reduce the dimensionality of the data. The other one is to find the hidden patterns. Mandl (Mandl 1999) discussed methods for dimensionality reduction using factor analysis in IR. Machado at el (Machado et al. 2003) presented a perspective to image retrieval based on multivari-

ate factor analysis to minimize data redundancy and reveal hidden patterns. Mehta at el (Mehta et al. 2006) proposed an approach for cross-system personalization by factor analysis. Their proposed factor analysis method offered an algorithmic improvement over their previous work by taking into account the incompleteness of data. In our proposed approach, factor analysis is applied to discover some hidden common factors as the "eliteness" variables that can be used to estimate the importance of term associations.

Principal Component Analysis (PCA) (Olivas et al. 2009) and factor analysis are two methods that can help reveal simpler patterns within a complex set of variables. In particular, they seek to discover if the observed variables can be explained largely or entirely in terms of factors. The main commonality between PCA and factor analysis is that they both have eigenvectors, eigenvalues, loading factors and scores. The differences are: (1) PCA is often used as a simple starting point in multi- variate analysis; (2) factor analysis is often considered to be "statistical" in nature rather than purely mathematical as in PCA, since PCA eigenvectors cumulatively account for all the variability in the data set whereas factor analysis results include an unresolved component; (3) factor analysis results are often transformed through varimax and other methods to optimize eigenvectors for interpretation. This motivates us to choose factor analysis to compute the importance and reliance of term associations, in order to find the hidden "eliteness" variables.

An n-gram (Manning and Schlztze 1999) is a subsequence of n items from a given sequence. An n-gram model is a type of probabilistic model for predicting the next

item in such a sequence. Some language models built from n-grams are "$(n-1)$-order Markov models". Its grammar is a representation of an n^{th} order Markov model in which the probability of occurrence of a symbol is conditioned upon the prior occurrence of $(n-1)$ other symbols. Probabilistic latent semantic analysis (PLSA) (Hofmann 1999) is a method of latent semantic analysis that uses probabilistic means to obtain the hidden topics and their relationships to terms and documents. In this paper, we use factor analysis to estimate the latent factors and compute the communalities for term associations statistically.

Some related work has been done in the biomedical domain during the past few years. In (Hu and Huang 2010), we concentrated on passage extraction and result combination. Three algorithms are presented for passage extraction to build indices and two result combination methods are proposed to combine the retrieval results from different indices. A naive model using factor analysis was also applied to improve the baselines for result combination, where unigrams and bigrams are considered. We also studied on a Bayesian learning approach to promoting diversity in ranking in (Huang and Hu 2009). In this approach, a re-ranking model computed the maximum posterior probability of the hidden property corresponding to each retrieved passage. Then it iteratively groups the passages into subsets according to their properties. In this paper, we focus on modelling term associations. The latent factors behind term associations reflect the importances and reliance of these term associations. They are decided by the proposed factor analysis

based model and their term appearances in the first round retrieved passages.

2.3.3 GSP

We adopt the Generalized Sequential Pattern (GSP) algorithm (Srikant and Agrawal 1996) as a comparison to our proposed approach, which contains two main steps as candidate generation and support counting. At first, all single items ($1-sequences$) are counted. Then, from the frequent single items, a set of candidates of $2-sequences$ are formed and filtered to identify their frequencies by removing the non-frequent items based on the minimum support. The frequent $2-sequences$ are used to generate the candidates of $3-sequences$. This process is repeated until no more frequent sequences are found. The support counting is based on the minimum support value.

2.4 Entity Ranking

2.4.1 Entity-based Work

Entity retrieval is to examine terms co-occurring with an entity in the context to satisfy the users' information needs. We have done an investigation on the entity-related topics of the papers published in the top conferences such as SIGIR, CIKM and VLDB. There are only 5% papers' topics on entity ranking. The rest of the entity-related papers focus on entity recognition/extraction, query expansion with entity, quantity entity, entity

information management and entity relationships.

In (Rode et al. 2008), the authors ranked the documents according to the relevance of the entities in the documents. (Balog et al. 2006) ranked expert entities according to the relevance sum of the documents containing the entities. (Blanco and Zaragoza 2010) adopted a tf-isf (term-frequency and inverse sentence frequency) method to rank the entities, where tf-isf is similarly defined as tf-idf. In (Na and Ng 2009), an entity ranking method was not mentioned specifically. However, according to the proposed model and algorithm, it ranked the entities according to the relevance of the document as well. (Zaragoza et al. 2007) ranked the entities by the relevance scores of the passages (short documents). Note that this paper ranked the entities instead of the related documents. (Kaptein et al. 2010) scored the entities based on the query type and the document entity type, not by frequency. (Sarmento et al. 2007) sorted the entities by the weighted scores, but not the document relevance scores. (Bron et al. 2010) also ranked the entities by the weighted relevance directly. (Cheng et al. 2007) extracted the entities directly and then put them together for the straightforward view of users. A ranking frame work "entityRank"was proposed for the ranking purpose.

The TREC entity track is targeted to perform entity-oriented search tasks on the World Wide Web (WWW), which makes users' information needs be better answered by specific entities instead of just any type of documents. The track defines entities as typed search results, "things" represented by their homepages on the web. Searching

for entities thus corresponds to ranking these homepages. The track thereby investigates a problem quite similar to the question answering (QA) list task. The entity track has hosted two years as 2009 and 2010. In general, the track's scope is limited to search for instances of the organizations, people, and product entity types (Balog et al. 2010).

2.4.2 Topic Modelling

A topic model is a statistical model for finding the latent "topics" occurred in the documents. An early topic model named probabilistic latent semantic indexing (PLSI) was proposed by Thomas Hofmann in 1999 (Hofmann 1999). Then in 2002, latent Dirichlet allocation (LDA) was developed by David Blei, Andrwe Ng and Michael Jordan (Blei et al. 2003), which allows documents to have a mixture of topics. Many previous work has been applied and extended on the topic model as (Lin and Wilbur 2007, Liu et al. 2010a, Lu et al. 2011, McCallum et al. 2007, Wei and Croft).

Sen (Sen 2012) proposed topic models to keep track of the context of every word in the knowledge base, so that words appearing within the same context as an entity were more likely to be associated with that entity. The author claimed that the proposed topic models utilized all text presented in the knowledge base and helped learn high-quality catalogs. Unlike most previous topic models, the topic models were non-parametric and did not require the user to specify the exact number of groups present in the knowledge base.

Cha and Cho (Cha and Cho 2012) discussed how they extended probabilistic topic models to analyze the relationship graph of popular social-network data, so that they could group or label the edges and nodes in the graph based on their topic similarity. In particular, the authors argued that the existing topic models could not handle popular nodes (nodes with many incoming edges) in the graph very well. Then they proposed possible extensions to the topic models to deal with popular nodes.

Kataria et al. (Kataria et al. 2011) proposed a semi-supervised hierarchical model called Wikipedia-based Pachinko Allocation Model (WPAM) that exploits, based on Latent Dirichlet Allocation (LDA) and its hierarchical variants.

(McCallum et al. 2007) presented a Author-Recipient-Topic (ART) model for social network analysis, which learned topic distributions based on the the direction-sensitive messages sent between entities. The model built on LDA and the Author-Topic (AT) model, adding the key attribute that distribution over topics was conditioned distinctly on both the sender and recipient-steering the discovery of topics according to the relationships between people.

(Wei and Croft) studied how to efficiently use LDA to improve ad-hoc retrieval by proposing an LDA-based document model within the language modeling framework and evaluate it on several TREC collections.

Lin et. al. (Lin and Wilbur 2007) proposed a probabilistic topic-based model for content similarity called pmra that underlies the related article search feature in PubMed.

Unlike previous probabilistic retrieval models, they did not estimate relevance, but rather focus on "relatedness", the probability that a user would want to examine a particular document given known interest in another.

(Liu et al. 2010b) presented an adaptive sentiment analysis model called S-PLSA+ to capture the hidden sentiment factors in the reviews with the capability to be incrementally updated as more data become available.

2.4.3 Proximity

Many previous work has been done on term proximity, such as (Broschart and Schenkel 2008, Büttcher et al. 2006, Hawking and Thistlewaite 1995, Keen 1991, 1992, Lv and Zhai 2009, Metzler and Croft 2005, Petkova and Croft 2007, Rasolofo and Savoy 2003, Song et al. 2008, Tao and Zhai 2007, Tellex et al. 2003, Zhao et al. 2011). Some early researches started to discover the effectiveness of term proximity in IR two decades. Keen's work (Keen 1991, 1992) proposed the use of term proximity by analogy with boolean techniques, where a "NEAR" operator was introduced to quantify the proximity of query terms. Hawking and Thistlewaite (Hawking and Thistlewaite 1995) evaluated "Span" proximity approaches on TREC data sets. Nowadays, most studies heuristically integrated word proximity into probabilistic weighting models, such as (Broschart and Schenkel 2008, Büttcher et al. 2006, Lv and Zhai 2009, Metzler and Croft 2005, Petkova and Croft 2007, Rasolofo and Savoy 2003, Tao and Zhai 2007, Zhao et al. 2011). Petkova

and Croft (Petkova and Croft 2007) retrieved named entities in the context of the documents with a proximity-based document representation. Tao (Tao and Zhai 2007) examined different measures and made a conclusion that the minimum pair-wise distance is most effective. In (Lv and Zhai 2009), the authors integrated the term proximity into the language model (Zhai and Lafferty 2001) as a positional language model. Zhao et. al. (Zhao et al. 2011) introduced a Cross Term to model term proximity for boosting retrieval performance in the BM25 system (Beaulieu et al. 1997).

2.5 Multi-source fusion

A lot of previous work has been done on result combination. In the TREC 2007 Genomics Track, there are more than seven teams which utilize result combination to improve their final submissions in a total of 66 runs by 27 teams. "NLMFusion", submitted by the team of National Library of Medicine (Demner-Fushman et al. 2007), as the top scoring automatic run for all three metrics of the passage2-level, the aspect-level and the document-level, suggested that combining results from different IR models may improve the final score. Here "NLMFusion" is an automatic run obtained by applying fusion to a LHNCBC run, a Terrier run, an NCBI Themes run, an INDRI run and an easyIR run. However, not all teams using fusion/combination achieved the successfully improvements. The teams from University of Neuchatel (Fautsch and Savoy 2007), European Bioinformatics Institute (Jimeno et al. 2007), Kyoto University (Wan et al. 2007) and so

on, showed slight declines in performance from their non-fusion/non-combination runs. Nevertheless, each team who used different methods, for fusing the individual different method runs, may have contributed to the differences in performance.

2.5.1 DFR vs BM25 vs LM

Divergence from randomness (DFR) (Salton et al. 1983), as one of five individual runs used in "NLMFusion", was reported to be the highest scoring subcomponent run in the TREC 2007 Genomics Track. "UniNE3" (Fautsch and Savoy 2007), the fusion run submitted by University of Neuchatel, also gave details of success in using it. Since DFR was often used in fusion as one of the components, such as in 49 automatic submissions in 2007, there was only a run as "UniNE1" from University of Neuchatel (Fautsch and Savoy 2007) which used DFR as a single model but did not combine too many other models.

Okapi BM25, as one of the best well-known probabilistic weighting function, was very popular in the TREC Genomics Tracks. "MuMshFd", the run submitted by University of Melbourne (Stokes et al. 2007), obtained the highest score of the passage2-level, the aspect-level and the document-level in all the BM25 submissions. Other teams who applied the Okapi BM25 model, such as those from York University (Huang et al. 2007) and University of Illinois at Chicago (Zhou and Yu 2007), obtained the performance around the mean MAP on all the evaluation measures. "DUTgen3", submitted by Dalian

University of Technology (Yang et al. 2007), which also used the Okapi BM25 model, however, only slightly hit the median MAP.

Language model, as one of the most well-known statistical model, was also employed popularly by many teams. "AIDrun3" submitted by Arizona State University (Tari et al. 2007), "DUTgen1" and "DUTgen2" submitted by Dalian University of Technology (Yang et al. 2007), "UBexp1" from University at Buffalo (Ruiz et al. 2007) and "kyoto1" from Kyoto University (Wan et al. 2007), achieved better average performance than the Okapi runs, although the individual run is not as good as the Okapi BM25 run, "MuMshFd" submitted by University of Melbourne.

2.5.2 Mutliple Sources

So far, not much previous work on multiple sources has been done in the probabilistic modelling field. However, there is lots of work in the other areas, such as biomedicine, image, video and the multilingual information retrieval.

Yuan et al. (Yuan et al. 2012) presented a problem of incomplete data when integrating large-scale brain imaging data sets from different imaging modalities. The authors claimed to address this problem by proposing two learning methods where all the samples (with at least one available data source) could be used. In the first method, they divided the samples according to the availability of data sources, and learnt shared sets of features with state-of-the-art sparse learning methods. Their second method learnt a

base classifier for each data source independently, then estimated the missing prediction scores. Finally, a multi-source fusion model was built.

Chattopadhyay et al. (Chattopadhyay et al. 2011) proposed a transfer learning framework based on the multi-source domain adaptation methodology for detecting different stages of fatigue using surface electromyography (SEMG) signals, in which the SEMG data of a subject represented a domain, data from multiple subjects in the training set form the multiple source domains and the test subject data form the target domain. SEMG signals were predominantly different in conditional probability distribution across subjects. The authors concluded that their key feature of the proposed framework was their weighting scheme that addressed the conditional probability distribution differences across multiple domains (subjects).

Knoblock et al. (Knoblock et al. 2001) have developed the Heracles framework for building Web-based information assistants. Their framework provided the infrastructure to rapidly construct applications that extracted information from multiple Web sources and interactively integrated the data using a dynamic, hierarchical constraint network. The authors described their core technologies that comprised the framework, including information extraction, hierarchical template representation, and constraint propagation.

Ribeiro and Matos (Ribeiro and de Matos 2008) deployed multi-document fusion in speech-to-text summarization systems. they proposed the inclusion of related, solid background information to cope with the difficulties of summarizing spoken language

and the use of multi-document summarization techniques in single document speech-to-text summarization. They also explored the possibilities offered by phonetic information to select the background information and conducted a perceptual evaluation to better assess the relevance of the inclusion of that information.

3 Term Association

This chapter is organized as follows. First, a term association approach is systematically and consistently presented, followed by a factor analysis based model and a corresponding algorithm, including a recursive re-ranking algorithm. Second, the IR environment is briefly described, including the data sets, queries, evaluation measures, the IR system and indices. After that, the experimental results and discussions are reported in the results and discussion section, which includes the analysis for the baselines, the proposed term association, the influence of different indices and $\mathbf{k}$ for the recursive re-ranking algorithm, the comparisons to the GSP algorithm and the official submissions.

3.1 Observations

In the traditional IR systems, keywords extracted from the queries are used to retrieve documents/passages with some weighting functions. term associations among keywords are examined to improve information retrieval performance. For example, there are n keywords extracted from a query, and the system gives N passages for each retrieval

Table 3.1: Sample of retrieval passage list

(1) The top 1000 passages are shown for each topic; (2) each document contains multiple passages such that some retrieved passages share the same document ID, but with different offset numbers and lengths; (3) the rank is given by the weights; (3) the weights are given by the IR system; (4) the offset is the starting point of the passage, and the length is the passage length, where the stopping words have been removed in the indexing stage; (5) the label is used to identify a specific run.

Topic #	Document ID	Rank	Weight	Offset	Length	Label
200	12595615	1	48.63	28426	295	yorkuga1
200	12595615	2	46.25	3839	339	yorkuga1
200	15814577	3	43.338	5656	125	yorkuga1
...	...	...	...	...	...	...

baseline result. Term associations among these n keywords are extracted and used for re-ranking the N passages.

Two main observation files from the system are: 1) the baseline list retrieved by the system with N passages for each query; 2) the corresponding term file which displays how many and which keywords are retrieved in each passage. The sample data are presented in Table 3.1 and 3.2.

Taken n keywords as a sequence, 1-keyword subsequence, 2-keyword subsequence and 3-keyword subsequence are studied as unigram, bigram and trigram term associa-

Table 3.2: Sample of the corresponding term file

(1) Specific retrieved terms are displayed, instead of a whole document; (2) the term file is a temporary file generated by the IR system; (3) the passage number is corresponded to the rank of the baseline, such as "passage #1" corresponds to the document ID "12595615" in Table 3.1.

Topic #200
passage #1: 4 of the 9 terms was found >> activ associ diseas lupus
passage #2: 4 of the 9 terms was found >> activ associ diseas lupus
passage #3: 3 of the 9 terms was found >> activ diseas lupus
... ...

tions. If one term is appeared in a passage, it scores 1; if not, it scores 0. Therefore each passage can be presented as a 1-0 vector as shown in Table 3.3.

3.2 A Factor Analysis Based Model

Factor analysis is a method for investigating whether a number of variables of interest $T_1, T_2, ..., T_n$, are linearly related to a smaller number of unobservable factors $F_1, F_2, ..., F_m$.

Based on the observation data, it is suggested that the observations are functions of a number of common underlying factors. The underlying factors, tentatively and rather loosely describe the unobservable features of the retrieval passages. The score over all

Table 3.3: Observation of keyword associations

(1) 1-keyword subsequences are unigrams, 2-keyword subsequences are bigrams, 3-keyword subsequences are trigrams; (2) all unigrams, bigrams and trigrams are defined as independent terms; (3) a passage scores 1 if a term is appeared in it, otherwise it scores 0; (4) a passage is represented as a 1-0 vector.

#	unigram			bigram			trigram		
	T_1	..	T_n	T_{n+1}	..	$T_{C_n^1+C_n^2}$	$T_{C_n^1+C_n^2+1}$	..	$T_{C_n^1+C_n^2+C_n^3}$
	k_1	..	k_n	k_1k_2	..	$k_{n-1}k_n$	$k_1k_2k_3$	..	$k_{n-2}k_{n-1}k_n$
1	1	..	1	0	..	1	0	..	1
2	1	..	1	1	..	1	1	..	1
..		.	.		.	.		.	.
N	0	..	0	0	..	1	0	..	1

term associations is the sum of a constant times a common factor, i.e., it is a linear combination of those common factors in Equation 3.1.

$$\sum_{i=1}^{m} \ell_i \dot{\times} CommonFactor_i \tag{3.1}$$

where m stands for the count of common factors, $m \leq n$. The numbers $\ell_1, .., \ell_m$ are the factor loadings associated with this term association.

Term associations contain unigrams, bigrams and trigrams. Then, the data applied by the factor analysis based model would be $C_n^1 + C_n^2 + C_n^3$ associations and N passages

for each query, which is a $(C_n^1 + C_n^2 + C_n^3) \times N$ matrix. The factor loadings and the common factors for each query must be inferred from the data. Here I use n' to denote $C_n^1 + C_n^2 + C_n^3$.

In order to compute the reliance of the associations, communality is defined for the n' associations as

$$h_i^2 = \ell_{i,0}^2 + \ell_{i,1}^2 + .. + \ell_{i,m}^2 \tag{3.2}$$

The rule is that the larger of the communalities h_i^2 are, the more important of common factors are to represent the keywords.

It is assumed that each term association is related to m factors. Therefore, the mathematical model for the above example can be written as follows.

$$\begin{cases} T_1 & = \ell_{1,0} + \ell_{1,1} f_1 + .. + \ell_{1,m} f_m + \varepsilon_1 \\ & \cdots \\ T_k & = \ell_{k,0} + \ell_{kN,1} f_1 + .. + \ell_{k,m} f_m + \varepsilon_k \\ & \cdots \\ T_{n'} & = \ell_{n',0} + \ell_{n',1} f_1 + .. + \ell_{n',m} f_m + \varepsilon_{n'} \end{cases} \tag{3.3}$$

where T_k is the score of the k^{th} term association, with $k = 1, .., n'$; $< f_1, .., f_m >$ is the unobserved common factor vector for the k^{th} term association; $< \ell_{k,0}, \ell_{k,1}, .., \ell_{k,m} >$ are the factor loading vector of the k^{th} term association; ε_k is the error term, which serves to indicate that the hypothesized relationships are not exact.

In matrix notation, Equation 3.3 becomes

$$\mathbf{T} = \mathbf{LF} + \varepsilon \tag{3.4}$$

where $\mathbf{T}$ is an $n' \times N$ matrix of observable data; $\mathbf{L}$ is an $n' \times (m+1)$ matrix of factor loadings, which are unobservable constants; $\mathbf{F}$ is an $n' \times m$ matrix of unobservable common factors; ε is an $n' \times N$ matrix of unobservable error variables.

Observe that by doubling the scale on which f_1 of $\mathbf{F}$ is measured, and simultaneously halving the factor loadings for $f_j(j = 2..m)$ makes no differences to the model. Thus, no generality is lost by assuming that the standard deviation of $f_j(j = 2..m)$ is 1. Likewise for f_1. Moreover, for similar reasons, no generality is lost by assuming every two factors f_i and f_j $(i \neq j)$ are uncorrelated with each other. The "errors" ε are taken to be independent of each other. The variances of the "errors" associated with the n' different associations are not assumed to be equal. The values of the factor loadings $\mathbf{L}$ and the variances of the "errors" ε can be estimated given the observed data $\mathbf{T}$.

3.3 A Factor Analysis Based Algorithm

A factor analysis based algorithm is proposed in Figure 3.1, in which seven phases are included. The phase of **Initialization** gives the initial values for this algorithm, such as $N = 1000$. The phase of **Matrices generation** creates the matrices of the associations, where unigrams, bigrams and trigrams are considered. The communalities are calculated

for all the associations at the phase of **Communality calculation**. Finally, the phase of **Re-ranking** is using a recursive re-ranking algorithm to re-rank the original result in Figure 3.2.

Keywords are directly extracted from the queries. There is a term file which displays how many and which keyword terms are retrieved for each passage by the system. In other words, all the retrieved passages can be labelled by the keywords. Furthermore, for the keywords in the queries, no query expansion but stemming is applied. For example, "change" can have several expressions such as "changeless", "changing", "changeable", and so on. So the system deals with "change" as "chang". The process is done automatically in the system (Huang et al. 2005b, Zhong and Huang 2006).

According to the keyword sequence, unigrams, bigrams and trigrams are generated as term associations for each query, which makes a $C_n^1 + C_n^2 + C_n^3 \times N$ matrix.

The factor analysis model is set up after generating the matrices. Through sorted the communalities, term associations are ranked by their importances. The larger the communalities are, the more important the corresponding associations are. Finally, the passages is recursively re-ranked as the output results using the recursive re-ranking algorithm in Figure 3.2.

0. Input

The baseline result for the queries on each data set.

The term file corresponding to the baseline result.

1. Output

A re-ranking result for the queries on each data set.

2. Initialization

$N = 1000$; $k = 10$;

3. Keyword Extraction

Read the term file;

For each query {Get the keyword sequence; Get the value of n;}

4. Matrices Generation

For each query {For $(i = 1; i <= 3; i++)$ { Generate the i-keyword subsequences;

For all the i-keyword subsequences {Search the subsequence in the term file;

If the subsequence exits, it scores 1;Else it scores 0;}}

A $(C_n^1 + C_n^2 + C_n^3) \times N$ matrix is generated. }

5. Communality Calculation

For each query {Mathematically set up the factor analysis model;

Estimate the factor loadings for the common factors; Compute the communalities; }

Sort the associations according to their communalities;

Get top k associations as the ranking association list.

6. Re-ranking

For each query {Call the recursive re-ranking algorithm in Section 3.4; }

A final re-ranking result is generated.

Figure 3.1: A Factor Analysis Based Algorithm

3.4 A Recursive Re-ranking Algorithm

A recursive re-ranking algorithm is called for the phase of **Re-ranking** in the previous factor analysis based algorithm. The pseudo codes in Figure 3.2 show how this recursive

0. Input

The baseline result for the queries on each data set.

The ranking association list generated by the factor analysis based model.

1. Output

A re-ranking passage list for the queries on each data set.

2. Initialization

$k = 10$ which will be discussed in Section 3.6.3;

3. Recursive Division

For the 1st association T_1' in the ranking association list,

The result list is divided into 2 parts: P_1^1 and P_2^1, where P_1^1 contains T_1' and P_2^1 does not.

For each part, the passages are sorted by their given weights.

For the 2nd association T_2' in the list,

4 parts: P_1^2, P_2^2 and P_3^2, P_4^2, where P_1^2 and P_3^2 contain T_2', while P_2^2 and P_4^2 do not.

Repeat to re-rank the result list for k times.

2^k parts, where the odd ones contain the associations while the even ones do not.

4. Re-ranking

For $(i = 1; i \leq 2^k; i++)$ {Sort the passages in P_i^k according to their weights; $q(i)$ = # of P_i^k}

Let RL is the final re-ranking list, v is the size of RL;

The first $q(1)$ passages in RL are the passages in P_1^k;

For P_{i+1}^k { $v(i)$ = # of RL;

Passages from $(v(i) + 1)$ to $(v(i) + q(i + 1))$ in RL are passages in P_{i+1}^k; $i = i + 1$; }

5. Final Result Generation

A re-ranking result list is generated.

Figure 3.2: A Recursive Re-ranking Algorithm

re-ranking algorithm works.

There are three main phases. The phase of **Initialization** gives the initial values

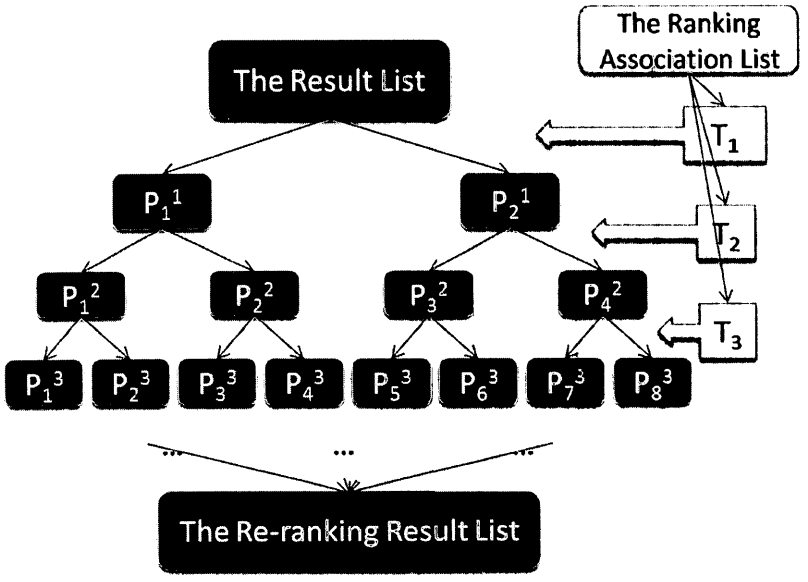

Figure 3.3: Recursive Division for Recursive Re-ranking

i.e. $k = 10$, in which a deep discussion is given in the section of influence of K for recursive re-ranking. The phase of **Recursive division** divides the passages into the base cases, according to the ranking association. This procedure is displayed in Figure 3.3, which is very similar to a binary tree. For example, the factor analysis based model gives a ranking list of terms as $\{T_1, T_2, T_3\}$ for re-ranking: (1) the baseline results are re-ranked by T_1, where P_1^1 are results containing T_1 and P_2^1 are results not containing T_1; (2) P_1^1 and P_2^1 are recursively re-ranked by T_2. P_1^2 are the results containing T_2 and T_1, while P_2^2 are those not containing T_2 but containing T_1. P_3^2 are the results containing T_2 but not containing T_1, while P_4^2 are those not containing T_2 and T_1; (3) similarly, $P_i^2 (i = \{1, 2, 3, 4\}$ are re-ranked by T_3. The phase of **Re-ranking** gets the passages in

the base cases $P_i^k (i = 1, .., 2^k)$. Finally, a recursive result list for re-ranking is generated.

3.5 Experimental Environment

Before the experimental results are reported, the experimental environment is briefly introduced. The data sets, queries, evaluation measures, gold standard and indexing are presented. More details are attached as the appendices.

3.5.1 Data Sets and Queries

The proposed model and algorithms are evaluated on the TREC 2004-2007 Genomics data sets and the TREC 2004 HARD data set.

TREC 2007 and 2006 Genomics data sets provide a test collection of 162,259 full-text documents assembled with 36 queries in 2007 and 28 queries in 2006. The TREC 2007 queries are in the form of questions asking for lists of specific entities. The definitions for these entity types are based on controlled terminologies from different sources, with the source of the terms depending on the entity type Hersh et al. (2007). The TREC 2006 queries are derived from the set of biologically relevant questions based on the Generic Topic Types (GTTs) (Hersh et al. 2005). Sample raw data are presented in Appendix C and all queries are listed in Appendix B.

TREC 2005 and 2004 Genomics data sets consists of a document collection for the ad hoc retrieval task which is a 10-year subset of MEDLINE with completed citations

from the database inclusive from 1994 to 2003. This provides a total of 4,591,008 records (Hersh et al. 2005). Each record is an abstract of a document. Then in this paper, I take an abstract as a passage. There are 50 queries for each year respectively. Sample raw data are presented in Appendix C and all queries are listed in Appendix B.

TREC 2004 HARD data set consists of entirely of English text, such as the Agence France Press (AFP), Associated Press (APW), Central News Agency (CNA), LA Times/Wash Post (LAT), New York Times (NYT), Salon.com (SLN), Ummah Press (UMM), Xinhua English (XIN) with the total collection of 652,710 documents. In our research, I parse the documents into passages (Allan 2004). Sample raw data are presented in Appendix C and all queries are listed in Appendix B.

3.5.2 Evaluation Measures

The TREC Genomics Track has three evaluation measures that are the document-level, the aspect-level and the passage2-level (a new measure for the TREC 2007 queries) (Hersh et al. 2007). Each of these provides insight into the overall performance for a user trying to answer the given queries and measured by some variant of mean average precision (MAP), which are briefly described as follows.

Document-level This is a standard IR measure. The precision is measured at every point where a relevant document is obtained and then averaged over all relevant docu-

ments to obtain the average precision for a given query. For a set of queries, the mean of the average precision for all queries is the mean average passage precision of that IR system.

Passage-level As described in (Hersh et al. 2006), this is a character-based precision calculated as follows. For each relevant retrieved passage, precision will be computed as the fraction of characters overlapping with the gold standard passages divided by the total number of characters included in all nominated passages from this system for the topic up until that point. Similar to regular MAP, relevant passages that are not retrieved will be added into the calculation as well, with precision set to 0 for relevant passages not retrieved. Then the mean of these average precisions over all topics will be calculated to compute the mean average passage precision.

Passage2-level This is a new character-based MAP measure which is added to compare the accuracy of the extracted answers and modified from the original measure Passage MAP. Passage2 treats each individually retrieved character in published order as relevant or not, in a sort of "every character is a mini relevance-judged document" approach (Hersh et al. 2007). This is done to increase the stability of the passage MAP measure against arbitrary passage splitting techniques.

3.5.3 Gold Standard

A gold standard is created by extracting out the relevance passages and entities for each topic. Judges for the relevant passages and entities are recruited from the institutions of track participants and other academic or research centres. They are required to have significant domain knowledge, typically in the form of a PhD in a life science. In summary, judges are given the following three instructions. First, reviewing the topic question and identifying key concepts. Second, identifying relevant paragraphs and selecting minimum complete and correct excerpts. Third, developing controlled vocabulary for entities based on the relevant passages and coding entities for each relevant passage based on this vocabulary (Hersh et al. 2006).

3.5.4 System

Okapi BSS (Basic Search System) is used as the main search system. Okapi is an information retrieval system based on the probability model of Robertson and Sparck Jones (Beaulieu et al. 1997, Huang et al. 2005a, 2006, Robertson and Jones 1976, Robertson and Walker 1994, Yin et al. 2010, Zhong and Huang 2006). The retrieval documents are ranked in the order of their probabilities of relevance to the query. Search term is assigned weight based on its within-document term frequency and query term frequency. The weighting function used is BM25.

$$w = \frac{(k_1+1)*tf}{K+tf} * \log \frac{(r+0.5)/(R-r+0.5)}{(n-r+0.5)/(N-n-R+r+0.5)} * \frac{(k_3+1)*qtf}{k_3+qtf} \oplus k_2 * nq * \frac{(avdl-dl)}{(avdl+dl)} \tag{3.5}$$

where N is the number of indexed documents in the collection, n is the number of documents containing a specific term, R is the number of documents known to be relevant to a specific topic, r is the number of relevant documents containing the term, tf is within-document term frequency, qtf is within-query term frequency, dl is the length of the document, $avdl$ is the average document length, nq is the number of query terms, the k_is are tuning constants (which depend on the database and possibly on the nature of the queries and are empirically determined), K equals to $k_1 * ((1-b) + b * dl/avdl)$, and $\oplus$ indicates that its following component is added only once per document, rather than for each term.

In the experiments, the tuning constant parameters k_1 and b are set to be different values. k_2 and k_3 are set to be 0 and 8 respectively.

3.5.5 Indexing

An important issue that most IR systems have to deal with is the size of the retrieved passages and the granularity of the indexed information. In the context of text retrieval, the granularity of the indexed text can be defined as the length of the indexed text unit and the size can be defined as the length of the retrieved passage. Here an indexed text

unit is defined as a *passage* (Hu and Huang 2008, 2010).

Three indices are built on the 2007 and 2006 Genomics data sets according to three passage extraction methods and a paragraph-based index is built on the 2005 and 2004 Genomics data sets (Hu and Huang 2010). A paragraph-based index is set up on the 2004 HARD data set as well. The sentence-based indexing is based on passages each of which has up to 3 sentences. The paragraph-based indexing is generated on passages each of which is a paragraph. Here a paragraph is defined as the sequence of sentences between the and tags from the HTML data set. The word-based indexing forms passages using a dynamic window (Hu and Huang 2008, 2010).

3.6 Experimental Results

The baseline results are reported in Table 3.4, which shows the performance under five parameter settings with three different indices in terms of the document-level, the passage-level and the passage2-level on the genomics 2004-2007 data sets and HARD 2004 data set respectively. Five groups have been set for the parameters of (k_1, b) with there indices. Therefore, there are 15 runs on all five TREC data sets. Note that only a paragraph-based index is set up for the TREC 2005 and 2004 Genomics data sets and the TREC 2004 HARD data set.

Corresponding to the baseline results, the results of the term association approach are generated using the proposed algorithms. The performance and improvements are pre-

Table 3.4: Performance of baselines

(1) Five parameter settings are set for (k_1, b) at the first and second columns; (2) "word" stands for the word-based index, "sentence" for the sentence-based index and "paragraph" for the paragraph-based index; (3) "documen", "passage" and "passage2" are three evaluation measures as the document-level, the passage-level and the passage2-level; (4) "Genomics 2007", "Genomcis 2006", "Genomics 2005", "Genomics 2004" and "'HARD 2004" are five TREC data sets; (5) only a paragraph-based index is set up for the TREC 2005 and 2004 Genomics data sets and the TREC 2004 HARD data set, as mentioned in the section of indexing.

k_1	b	Indices	Genomics 2007			Genomics 2006		Genomics 2005	Genomics 2004	HARD 2004	
			document	passage	passage2	document	passage	document	document	document	passage
		word	0.1584	0.0675	0.0267	0.2662	0.0532	-	-	-	-
0.4	2.0	sentence	0.1368	0.0406	0.0154	0.2378	0.0398	-	-	-	-
		paragraph	0.1086	0.0170	0.0094	0.2036	0.0192	0.1964	0.2952	0.2449	0.2635
		BEST	0.1584	0.0675	0.0267	0.2662	0.0532	0.1964	0.2952	0.2449	0.2635
		word	**0.2108**	**0.0963**	0.0364	0.3140	**0.0718**	-	-	-	-
0.5	1.3	sentence	0.1805	0.0700	0.0350	0.3030	0.0550	-	-	-	-
		paragraph	0.1588	0.0452	0.0333	0.3109	0.0369	0.2602	0.3404	0.2802	**0.2985**
		BEST	0.2108	0.0963	0.0364	0.3140	0.0718	0.2602	0.3404	0.2802	0.2985
		word	0.1556	0.0434	0.0328	0.3097	0.0659	-	-	-	-
1.0	1.0	sentence	0.1809	0.0758	0.0350	0.2918	0.0521	-	-	-	-
		paragraph	0.1902	0.0893	0.0327	0.2916	0.0337	0.2547	0.3425	0.2522	0.2718
		BEST	0.1902	0.0893	0.0350	0.3097	0.0659	0.2547	0.3425	0.2522	0.2718
		word	0.1809	0.0780	0.0295	0.3045	0.0651	-	-	-	-
1.2	0.75	sentence	0.1987	0.0814	0.0394	0.3202	0.0522	-	-	-	-
		paragraph	0.2013	0.0648	0.0578	0.3381	0.0362	**0.2874**	**0.3584**	0.2617	0.2758
		BEST	0.2013	0.0814	0.0578	0.3381	0.0651	0.2874	0.3584	0.2617	0.2758
		word	0.1953	0.0844	0.0317	0.3152	0.0637	-	-	-	-
2.0	0.4	sentence	0.2084	0.0758	0.0401	**0.3529**	0.0490	-	-	-	-
		paragraph	0.2025	0.0633	**0.0641**	0.3476	0.0362	0.2779	0.3483	**0.2810**	0.2895
		BEST	0.2084	0.0844	0.0641	0.3529	0.0637	0.2779	0.3483	0.2810	0.2895

sented in Table 3.5. The values in the parentheses are the relative rates of improvement over the original results.

3.6.1 Influence of Parameter Settings and Indices

In order to investigate the influence of different indices and parameter settings, the experimental results are deeply analysed. First, taken the TREC Genomics 2007 and 2006 data sets as an example, the max, min, mean and sample standard deviation of the baselines are computed in the table 3.6. From this table, it can be seen how these settings effect the results, since there is a disparity between the max and the min values under all the measures. Focused on the sample standard deviation, the SSD values are calculated as a sample standard deviation of a discrete random variable. Compared to the mean, the SSD also shows the influence of the different indices and parameter settings.

To illustrate the results in Table 3.4 graphically, these data in Figure 3.4 and 3.5 are replotted. The performance of the baseline results is shown in terms of the document-level, the passage-level and the passage2-level. The x-axis represents the evaluation measures, where "word", "sen" and "par" stand for the word-based, the sentence-based and the paragraph-based indices. The y-axis shows the MAP performance. This figure shows that the sentence-based index produces the best results in terms of the document-level, the word-based index for the best results in terms of the passage-level and the paragraph-based index for the best results in terms of the passage2-level. This finding also confirms

Table 3.5: Performance of the term association approach

(1) All the runs are under the same settings as the baselines in Table 4.2, including the parameter settings of (k_1, b), the indices and the evaluation measures; (2) the values in the parentheses are the relative rates of improvement over the original results.

k_1	b	Indices	Geno 2007			Geno 2006		Geno 2005	Geno 2004	HARD 2004	
			document	passage	passage2	document	passage	document	document	document	passage
		word	0.2060	0.1296	0.0526	0.2790	0.0765	-	-	-	-
			(30.06%)	(92.00%)	(96.87%)	(4.80%)	(43.87%)	-	-	-	-
0.4	2.0	sentence	0.1710	0.0955	0.0330	0.2477	0.0698	-	-	-	-
			(25.01%)	(135.26%)	(114.11%)	(4.14%)	(75.50%)	-	-	-	-
		paragraph	0.1508	0.0726	0.0336	0.2161	0.0365	0.2156	0.3001	0.2458	0.2683
			(38.90%)	(326.77%)	(256.97%)	(6.15%)	(90.16%)	(9.78%)	(1.66%)	(0.37%)	(1.82%)
		word	0.2668	**0.1611**	0.0650	0.3445	**0.1010**	-	-	-	-
			(31.31%)	(67.31%)	(78.66%)	(9.71%)	(40.64%)	-	-	-	-
0.5	1.3	sentence	**0.2724**	0.1392	0.0619	0.3376	0.0889	-	-	-	-
			(45.38%)	(98.81%)	(76.99%)	(11.43%)	(61.59%)	-	-	-	-
		paragraph	0.1953	0.1040	0.0638	0.3270	0.0579	0.2879	0.3459	0.2843	**0.3031**
			(23.01%)	(130.18%)	(91.48%)	(5.17%)	(56.88%)	(10.65%)	(1.62%)	(1.46%)	(1.54%)
		word	0.2385	0.1425	0.0526	0.3428	0.0975	-	-	-	-
			(53.28%)	(228.44%)	(60.23%)	(10.67%)	(47.89%)	-	-	-	-
1.0	1.0	sentence	0.2251	0.1345	0.0551	0.3202	0.0842	-	-	-	-
			(24.43%)	(77.44%)	(57.29%)	(9.73%)	(61.68%)	-	-	-	-
		paragraph	0.1955	0.0969	0.0564	0.3069	0.0529	0.2777	0.3498	0.2594	0.2801
			(2.80%)	(8.50%)	(72.57%)	(5.25%)	(57.04%)	(9.03%)	(2.13%)	(2.85%)	(3.05%)
		word	0.2469	0.1381	0.0547	0.3221	0.0881	-	-	-	-
			(36.49%)	(77.03%)	(84.33%)	(5.77%)	(35.28%)	-	-	-	-
1.2	0.75	sentence	0.2698	0.1483	0.0667	0.3457	0.0823	-	-	-	-
			(35.77%)	(82.18%)	(69.35%)	(7.95%)	(57.74%)	-	-	-	-
		paragraph	0.2348	0.1155	0.0778	0.3483	0.0444	**0.3085**	**0.3606**	0.2659	0.2812
			(16.64%)	(78.31%)	(34.56%)	(3.01%)	(22.64%)	(7.34%)	(0.61%)	(1.60%)	(1.96%)
		word	0.2450	0.1355	0.0568	0.3228	0.0763	-	-	-	-
			(25.44%)	(60.53%)	(79.09%)	(2.42%)	(19.71%)	-	-	-	-
2.0	0.4	sentence	0.2605	0.1327	0.0622	**0.3549**	0.0697	-	-	-	-
			(25.01%)	(75.08%)	(55.20%)	(0.58%)	(42.27%)	-	-	-	-
		paragraph	0.2308	0.1084	**0.0762**	0.3533	0.0521	0.2889	0.3502	**0.2845**	0.2956
			(13.95%)	(71.30%)	(18.86%)	(2.78%)	(44.01%)	(3.96%)	(0.55%)	(1.25%)	(2.11%)

Table 3.6: MAX, MIN, mean and SSD of the genomics 2007 and 2006 baselines

(1) The parameter settings of (k_1, b) and the different indices effect the baseline results greatly, since there is a disparity between the max and the min values under all the measures; (2) the SSD values are calculated as a sample standard deviation of a discrete random variable; (3) the values in the parentheses are the relative rates of improvement over the means; (4) the SSD also reflects the influence of the different indices and parameter settings of (k_1, b).

	Genomics 2007			Genomics 2006	
	document	passage	passage2	document	passage
MAX	0.2108	0.0963	0.0641	0.3529	0.0718
MIN	0.1086	0.017	0.0094	0.2036	0.0192
Mean	0.1778	0.0662	0.0346	0.3005	0.0487
SSD	0.0291	0.0214	0.0136	0.0397	0.0147
	(-16.37%)	(-32.33%)	(-39.17%)	(-13.20%)	(-30.18%)

the motivation for building up different indices for different information needs.

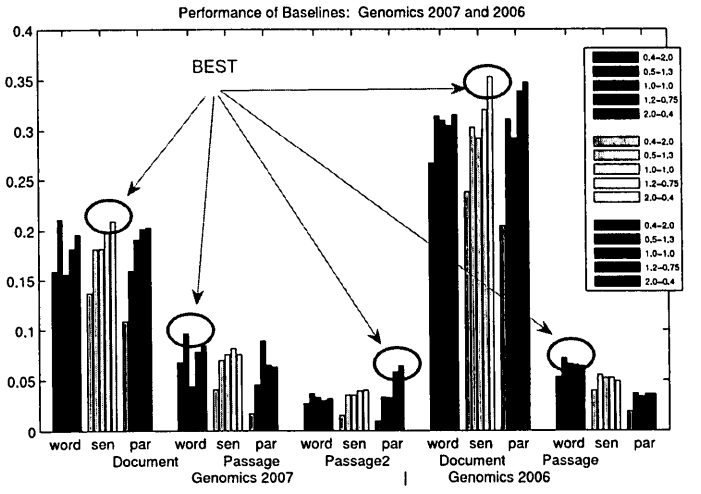

Figure 3.4: Performance of baselines, Genomics 2007 and 2006

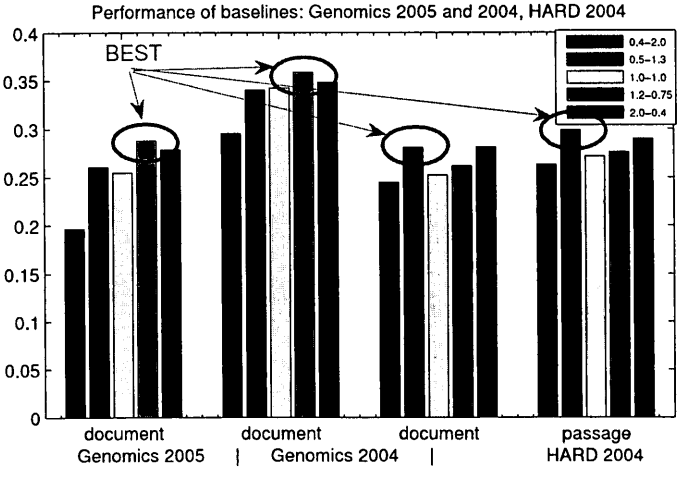

Figure 3.5: Performance of Baselines, Genomcis 2005 and 2004, HARD 2004

3.6.2 Influence of Term Association

In order to illustrate the term association results in Table 3.5, the data are re-plotted graphically in Figure 3.6 and 3.7. It clearly shows that, for all the measures on five TREC data sets, the term association approach always outperforms the baselines. The improvements in the parentheses explain the significance evidently. More interesting, the figures of the factor analysis results almost have the same distributions as the figures of baselines. The best factor analysis results always come from the best baseline results. The sentence-based index produces the best factor analysis results in terms of the document-level, the word-based index for the best factor analysis results in terms of the passage-level and the paragraph-based index for the best factor analysis results in terms of the passage2-level.

The improvements of term association in Table 3.5 are illustrated in Figure 3.8, 3.9, 3.10 and 3.11. There are two observations as follows. First, the positive values of the improvements notify that term association carries important weight on the retrieval results, which is much better than the baselines that only consider the unigram keywords independently. In other words, those bigram and trigram associations have more influential in the retrieval results than the unigram keywords. Second, the influence in terms of the passage levels (the passage2-level and the passage-level) is greater than that in terms of the document-level. It can be see in 3.8, 3.9 and 3.11, that the absolute values of im-

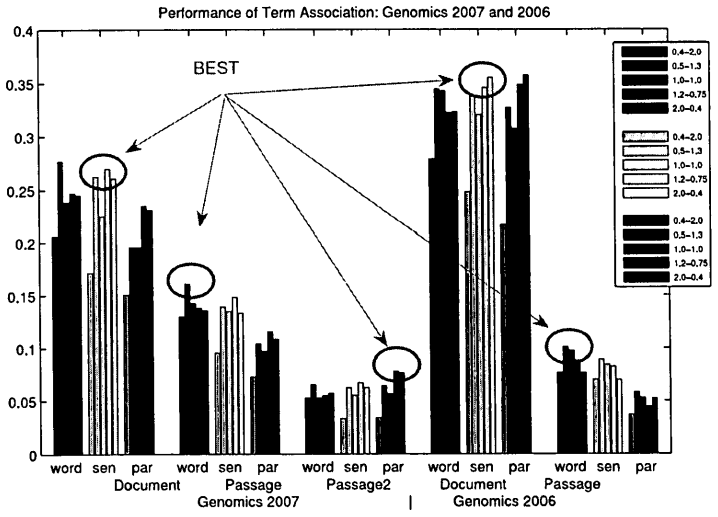

Figure 3.6: Performance of the Term Association Approach, Genomics 2007 and 2006

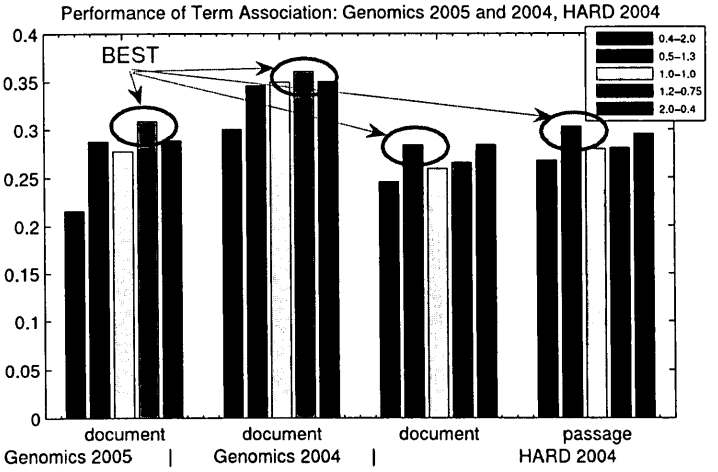

Figure 3.7: Performance of the Term Association Approach, Genomcis 2005 and 2004, HARD 2004

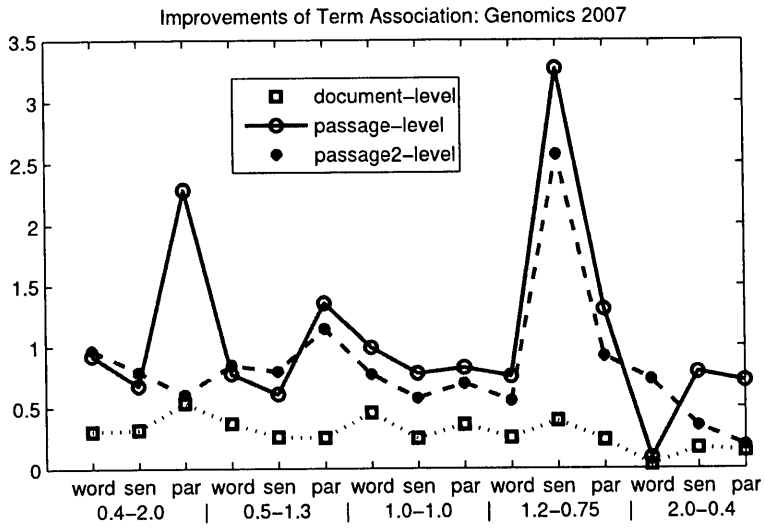

Figure 3.8: Improvements of the Term Association Approach, Genomcis 2007

provements on the passage-level are much higher than those on the document-level. This can be explained that term association is more efficient to be applied in the sentences or paragraphs compared to the documents.

3.6.3 Influence of K for Recursive Re-ranking

The depth in the recursive re-ranking algorithm is initialized as $k = 10$. The number k stands for the top k term associations weighted by the factor analysis based model. The retrieved passages are recursively re-ranked according to whether the passages contain the top k term associations or not. A series of experiments have been conducted with different settings of k values in order to investigate the influence of value k and find

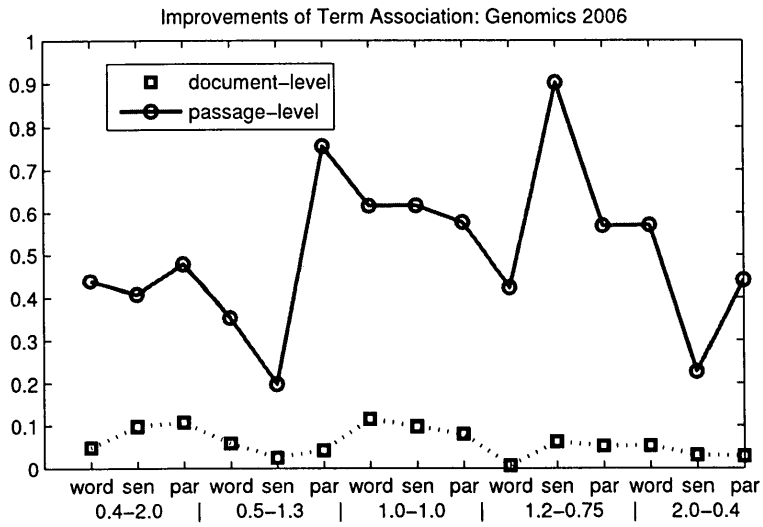

Figure 3.9: Improvements of the Term Association Approach, Genomcis 2006

a local optimization value for the proposed algorithm. Five original baselines are first randomly chosen from the five data sets respectively, namely Genomics 2007, Genomics 2006, Genomics 2005, Genomics 2004 and HARD 2004. Then, the factor analysis model is applied on the baselines. Five numbers such as 1, 5, 10, 20, 100, are tested and the performance is shown in Table 3.7. The number k affects the performance greatly when k is smaller than 10. However, when k becomes larger than 10, the final performance almost has no change. Therefore, this local optimization number is obtained as 10 for k in the recursive re-ranking algorithm for all the runs.

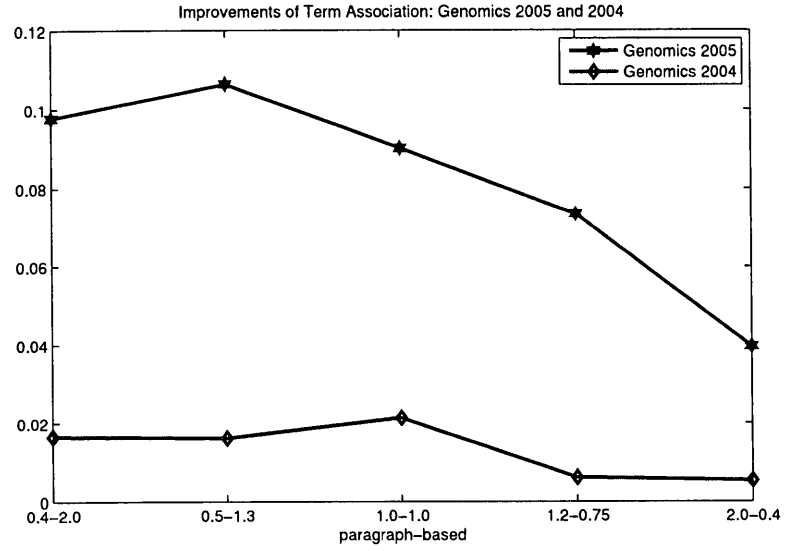

Figure 3.10: Improvements of the Term Association Approach, Genomcis 2005 and 2004

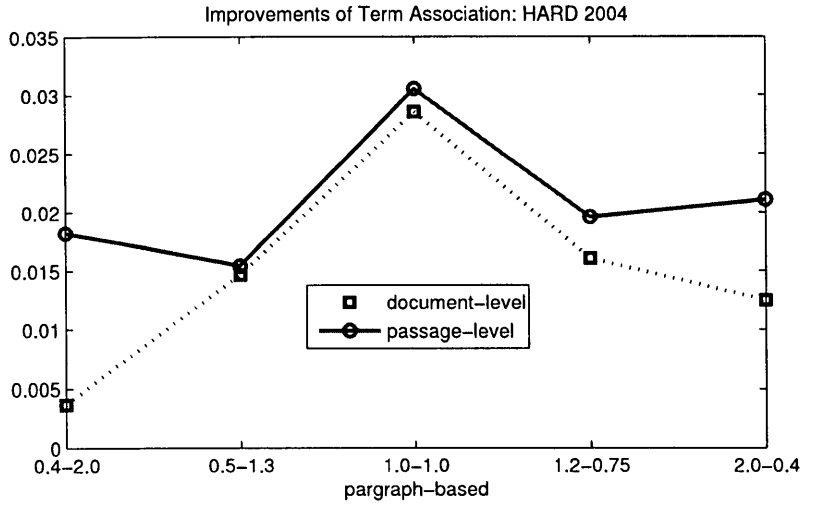

Figure 3.11: Improvements of the Term Association Approach, HARD 2004

Table 3.7: Number **k** discussion

	n	document	passage	passage2
Genomics 2007	1	0.3012	0.0918	0.1436
	5	0.3349	0.1400	0.1588
	10	0.3438	0.1422	0.1635
	20	0.3438	0.1422	0.1635
	100	0.3438	0.1422	0.1635
Genomics 2006	1	0.3974	0.1401	-
	5	0.4049	0.1445	-
	10	0.4087	0.1467	-
	20	0.4083	0.1466	-
	100	0.4083	0.1466	-
Genomics 2005	1	0.3012	-	-
	5	0.3116	-	-
	10	0.3123	-	-
	20	0.3123	-	-
	100	0.3123	-	-
Genomics 2004	1	0.3470	-	-
	5	0.3555	-	-
	10	0.3584	-	-
	20	0.3584	-	-
	100	0.3584	-	-
HARD 2004	1	0.2015	0.2005	-
	5	0.2223	0.2197	-
	10	0.2250	0.2208	-
	20	0.2248	0.2208	-
	100	0.2248	0.2208	-

3.6.4 Comparison with GSP Algorithm

The GSP algorithm is adopted as a comparison to the proposed approach. In order to map the GSP algorithm to the research problem, the keywords extracted from the queries are treated as the singleton items and N passages retrieved by the system for each query are as the transaction database. Therefore, the candidates of 1-sequences are all the keywords, the k-sequences candidates are generated on the frequent (k-1)-sequences. For the support counting, the minimum support value is defined corresponding to each query as follows. First, the counts of candidates are automatically calculated by the modified GSP algorithm, including all k-sequences. Then, the counts is simulated as a non-parametric distribution. Third, the 95% confidence interval of this distribution is computed, where the lower bound is the minimum support value for this GSP algorithm.

A study is performed on how the GSP algorithm performs on the five data sets. Table 3.8 shows the experimental results with the paragraph index under five parameter settings. Furthermore, the best results of the GSP algorithm are compared with the results of the baselines and the proposed term association approach.

An interesting finding is drawn from the results of the GSP algorithm. The GSP algorithm works very well in terms of the passage-level and the passage2-level, while it is not good for the document-level. This can be explained by the following scenario. The frequent $3-sequence$ $T_1T_3T_4$ is found in the documents D_1 and D_2. In D_1, $T_1T_3T_4$ is

Table 3.8: Performance of GSP algorithm

(1) The candidates of 1-sequences are all the keywords, the k-sequences candidates are generated on the frequent (k-1)-sequences, after mapped the GSP algorithm to our research problem; (2) the counts of candidates are simulated as a non-parametric distribution, where the lower bound of the 95% confidence interval is the minimum support value for this GSP algorithm; (3) only the paragraph index under five parameter settings of (k_1, b) is considered; (4) the best results of the GSP algorithm are compared with the best of the baselines and the proposed term association approach; (5) "TA" stands for term association; (6) the values in the parentheses are the relative rates of improvement over the original baselines.

	(k_1, b)	Geno 2007			Geno 2006		Geno 2005	Geno 2004	HARD 2004	
		document	passage	passage2	document	passage	document	document	document	passage
GSP	(0.4,2.0)	0.1066	0.0338	0.0149	0.1892	0.0242	0.1867	0.2723	0.2358	0.2639
		(-1.87%)	(-98.75%)	(-58.28%)	(-7.09%)	(-25.95%)	(-4.96%)	(-7.74%)	(-3.72%)	(-0.15%)
	(0.5,1.3)	0.149	0.0843	0.0456	0.2855	0.0466	0.2423	0.3165	0.2562	0.3001
		(-6.18%)	(-86.59%)	(-36.85%)	(-8.17%)	(-26.31%)	(-6.88%)	(-7.01%)	(-8.57%)	(-0.54%)
	(1.0,1.0)	0.1839	0.0898	0.0357	0.2757	0.0402	0.2385	0.3166	0.2501	0.2842
		(-3.32%)	(-0.60%)	(-9.21%)	(-5.46%)	(-19.40%)	(-6.36%)	(-7.55%)	(-0.83%)	(-4.56%)
	(1.2,0.75)	0.1905	0.0714	0.0658	0.3174	0.0404	0.2655	0.3293	0.2589	0.2776
		(-5.35%)	(-10.11%)	(-13.79%)	(-6.11%)	(-11.65%)	(-7.62%)	(-8.11%)	(-1.07%)	(-0.65%)
	(2.0,0.4)	0.1931	0.0657	0.0667	0.3203	0.0403	0.2588	0.3206	0.2567	0.2916
		(-4.62%)	(-3.79%)	(-4.02%)	(-7.85%)	(-11.40%)	(-6.89%)	(-7.96%)	(-8.65%)	(-0.73%)
	Best	0.1931	0.0898	0.0667	0.3203	0.0466	0.2655	0.3293	0.2589	0.3001
Baselines	Best	0.2108	0.0963	0.0641	0.3529	0.0718	0.2874	0.3584	0.281	0.2985
TA	Best	0.2724	0.1611	0.0762	0.3549	0.101	0.3085	0.3606	0.2845	0.3031

contained in a short passage so that D_1 earns good MAP results on the document-level and the passage-level. In the document D_2, the situation is that T_1, T_3 and T_4 are found in different passages respectively. Since $T_1T_3T_4$ is still found as a sequence based on the definitions, D_2 is given a high weight and is going to earn good performance at least on the document-level. However, the standard evaluation does not think D_2 is qualified to be a relative document so that D_2 decreases the performance of the document-level.

Compared to the GSP algorithm, the proposed term association approach outperforms the baselines and the GSP results on all the measures. The factor analysis based model considers not only the concurrence of the terms, but also the dependency, especially in the high order structure. In the GSP algorithm, the document D_2 is given a good score. However, in the factor analysis based model, the factor loadings of $T_1T_3T_4$ in D_2 are very small, since $T_1T_3T_4$ is not treated as a trigram term association. T_1, T_3 and T_4 are three unigram terms, while $T_1T_3T_4$ is a frequent $3-$ *sequence* in the GSP algorithm. So the proposed approach avoids assigning a high weight to the document D_2.

The major difference among the proposed approach, N-gram and PLSA, is that term associations are not dependent on the previous associations, whose reliance and importance are decided by the dependencies among the keywords in the passages, not by their probabilities upon the previous terms. For example, another interesting finding using factor analysis in this work, is that the bigram $k_1k_j(j \neq 1)$ might have the highest reliance, even though their previous unigram term k_1 or k_j is not the most important for

a query in some IR systems. And the experiment confirms that $k_1 k_j$ plays an important role in the improved re-ranking result. Therefore, one of the major contributions of the proposed approach is to extract subsequences as term associations from a query without preliminary knowledge. This promotes me to employ the GSP algorithm as a comparison to evaluate the proposed approach statistically, but not to compare this approach with PLSA and PCA.

3.6.5 Comparison with Official Submissions

In order to further evaluate the term association approach to improving performance, the performance of the term association approach is compared to the official submissions at the best and mean values on the five TREC data sets in Table 3.9. Since the submissions of the 2004 HARD data set are not officially released, only the genomics data sets are presented. For the mean performance, term association outperforms baselines and the official submissions. For some best performance, term association makes improvements on baselines, but is not as good as the official submissions. However, based on the discussion upon the influence of term association in the section of influence of term association, higher performance can be achieved if there are better baselines.

Table 3.9: Comparisons of baselines, term associations and official submissions

(1) All the results are compared at the best values and the mean values; (2) the submissions of the TREC 2004 HARD data set are not officially released; (3) "TA" stands for term association and "official" for official submissions.

		Geno 2007			Geno 2006		Geno 2005	Geno 2004	HARD 2004	
		document	passage	passage2	document	passage	document	document	document	passage
Baselines	Best	0.2108	0.0963	0.0641	0.3529	0.0718	0.2874	0.3584	0.2810	0.2985
	Mean	0.1778	0.0662	0.0346	0.3005	0.0487	0.2553	0.3370	0.2640	0.2798
TA	Best	0.2724	0.1611	0.0762	0.3549	0.1010	0.3085	0.3606	0.2845	0.3031
	Mean	0.2273	0.1236	0.0579	0.3182	0.0719	0.2757	0.3413	0.2680	0.2857
Official	Best	0.3105	0.0976	0.1097	0.5439	0.1486	0.3020	0.4075	-	-
	Mean	0.1891	0.0582	0.0421	0.2887	0.0392	0.1968	0.2074	-	-

3.6.6 A Case Study

Topic 200 of the TREC 2007 queries is taken as an example. The description for Topic 200 is "What serum [PROTEINS] change expression in association with high disease activity in lupus?". Nine keywords are extracted as serum, proteins, change, expression, association, high, disease, activity and lupus. The rest words are removed by the system as the stop words. The system stems the keywords as serum, protein, chang, express, associ, high, diseas, active and lupus.

Table 3.10 shows the baseline whose parameters are set as $(k_1, b) = (2.0, 0.4)$ with the paragraph-based index. The information of its keywords, the term count, the fre-

Table 3.10: Topic 200: keyword frequency rank

(1) Terms are extracted with stemming; (2) term counts are obtained from the first round retrieved passages, which are the top 1000 retrieved passages as the baseline; (3) the percentage is calculated based on 9 terms; (4) the rank depends on the term counts; (4) the parameters for this baseline are $(k_1, b) = (2.0, 0.4)$ with the paragraph-based index.

#	Term	Term count	Percentage	Rank
1	Lupus	869	23.80%	1
2	Diseas	753	20.70%	2
3	Activ	496	13.60%	3
4	Associ	476	13.10%	4
5	Serum	294	8.10%	5
6	High	274	7.50%	6
7	Protein	195	5.40%	7
8	Express	179	4.90%	8
9	Chang	108	3.00%	9

quency and rank are presented for Topic 200. The parameters for this baseline are $(k_1, b) = (2.0, 0.4)$ with the paragraph-based index. There are totally $(C_9^1 + C_9^2 + C_9^3)$ term associations generated by the proposed approach. Table 3.11 presents the top 10 term associations after applying the factor analysis based model, where terms, term count and their communalities are presented. Then in Table 3.12, the performance of term association is compared with the performance of baseline of Topic 200 in terms of the document-level, the passage-level and the passage2-level.

First of all, no unigram is in the ranking association list. All the term associations in Table 3.11 are bigrams and trigrams. Since the term association improved result outperforms the baseline, it means that term association works very well on all the measures. Therefore, term association is better than only considering the keywords independently. Second, the trigram "high lupus serum" has the higher reliance than the bigram "activ serum", although the trigram's term count is only 7, which is much less than the bigram's term count as 118. This tells that the term frequency might not make sense when compared to term association.

Table 3.11: Topic 200: Ranking Term Associations

(1) The top 10 term associations are shown among $(C_9^1 + C_9^2 + C_9^3)$ term associations generated by the proposed approach; (2) term count are computed from the top 1000 documents of the baseline; (3) the communality of each term association is calculated by the factor analysis based model; (4) the rank of term associations is given by their communalities.

Rank	Term Association	Term Count	Communalities
1	high lupus serum	33	69.4
2	lupus protein serum	47	62.7
3	activ lupus serum	7	61.2
4	activ serum	118	60.0
5	activ associ diseas	124	59.9
6	associ diseas lupus	162	59.5
7	activ associ high	7	59.2
8	lupus serum	116	58.5
9	diseas high lupus	90	58.0
10	associ protein	20	58.0

Table 3.12: Topic 200: Performance Comparison

(1) The values in the parentheses are the relative rates of improvement over the baselines;

(2) term association outperforms baseline.

	document	passage	passage2
Baseline	0.3752	0.1546	0.0688
Term Association	0.4238	0.2157	0.0811
Improvements	(12.95%)	(39.52%)	(17.88%)

4 Entity Ranking

An overview of the proposed entity ranking approach is presented, where the research problem is first defined, then the entity view and the document view are introduced, entity relevance and entity popularity are proposed mathematically. Section 4.5 presents a kernel-based proximity model as the local relevance to compute the entity-based proximity. Section 4.6 describes the global enhancer of how topic model works. In Section 4.7, the experimental environment is set up and the experimental results are presented, followed by the discussion and analysis in Section 4.8.

4.1 Problem Definition

Throughout this work, the terms are formally defined as (1) an *entity* instance is a type of distinct, separate existence, such as product name, people, location and organization, denoted by e_i; (2) an *entity document* is a sequence of p entities denoted by $Ed = (e_1, e_2, ..., e_p)$, where e_i is the i^{th} entity in the sequence; (3) a *corpus* is a collection of m documents denoted by $\mathfrak{D} = \{d_1, d_2, ..., d_m\}$, where $d_j = (w_{j,1}, w_{j,2}, ..., w_{j,m'})$; and (4),

an *entity corpus* is a collection of entities, denoted by $\mathfrak{E} = \{e_1, e_2, ..., e_e\}$.

As described in the motivation, the following majority issues are proposed to solve: (1) relevance: why and how entities can be related to the given query; (2) popularity: how frequent an entity can be to the given query; (3) disambiguation: which aspects of entities should be enhanced to the given query and (4) ranking: how to sort entities to a given query. The framework mathematically is proposed in Equation 4.1.

$$rank(e, q) = relevance(e, q) \oplus popularity(e, q) \tag{4.1}$$

Given

Entity collections $\mathfrak{E} = \{e_1, ..., e_e\}$ over Web page collections

$\mathfrak{D} = \{d_1, ..., d_m\}$;

Input

Entity query $q = \{e_{q1}, ..., e_{qk}\}$ where e_{qi} is an entity keyword.

Output

Ranked named entity list of $E_q = \{e_{q1}, ..., e_{qr}\}$

with the ranking weight $rank(e_i, q)$ given query q,

Figure 4.1: The Entity Retrieval Problem Definition

4.2 Entity View vs Document View

EntityCube is definitely entity aware and is working on the entity-based querying. However, all the entity information is from the documents. Upon this situation, EntityCbue first moves the entity view to the document view, then back from the document view to the entity view.

Here the proposed research problem is transferred from entity awareness to document awareness, where Equation 4.1 becomes

$$rank(d, q) = relevance(d, q) \oplus popularity(d, q) \tag{4.2}$$

Then, taken the entity view back from the document view through accumulating the entity ranking on the document computation, the entity relevance is obtained in Equation 4.3, where the document set D_{e_i} are composed by all the documents $\{d_{1,e_i}, .., d_{n',e_i}\}$ containing entity e_i, and e_i's frequency information is represented by the vector $TF_{e_i=\{tf_{1,e_i},..,tf_{n',e_i}\}}$.

$$relevance(e_i, q) = \sum_{d_{j,i} \in D_i} tf_{j,e_i} \times relevance(d_{j,i}, q) \tag{4.3}$$

4.3 Relevance

Two features are defined for relevance as follows: (1) the local relevance $rel_p(q, e)$ by entity proximity; (2) the global enhancer $rel_T(q, e)$ by entity topic model. In order to integrate these two features, a proximity-based entity search with entity topic model is proposed as an important contribution.

Towards the local relevance $rel_p(q, e)$, an entity proximity-based model $\omega_p(q, e)$ is presented in Section 4.5 to consider the effectiveness of entity positions given a query. An embedded N-gram model is proposed in Section 4.5.2 while a query contains more than an entity. The proximity is presented as

$$\begin{aligned} rel_p(q, e) &= \omega_p(q, e) \\ &= \frac{1}{\mathbf{K}} \sum_{i=1}^{N} \delta_d(i, q) ker(q, e) \end{aligned} \tag{4.4}$$

where $\mathbf{K} = \sum_{\mathbf{i=1}}^{\mathbf{N}} \mathbf{ker(q, e)}$ and the details are introduced in Section 4.5.

In order to enhance the local relevance, an entity-based topic model is developed as the global enhancer. The hidden variables z are estimated as "topics" or "mixed topic distributions" over the documents in the corpus. Equation 4.5 presents the basic topic definition and the detailed model is proposed in Section 4.6. Finally, $P(z|e)$ is obtained for all the documents in the corpus.

$$\begin{aligned} P(e, d) &= P(e)P(d|e) \\ &= P(e) \sum_{z} P(d|z)P(z|e) \\ &= \sum_{z} P(d|z)P(e|z)P(z) \end{aligned} \tag{4.5}$$

Towards the entity-based queries, queries are put into topic model as well. $P(q|z)$ is calculated to get $rel_T(q, e)$ given $P(z|e)$ in

$$rel_T(q, e) =< P(q|z), P(z|e) > \tag{4.6}$$

Based on the local relevance and the global enhancer, the two features are integrated as an expectation in Equation 4.7. Here the traditional linear combination is tested in the experiments. However, there is no obvious difference among the results while tuning the parameters. Then the expectation is adopted in the method.

$$\begin{aligned} relevance(d,q) &= rel_T(q,e) \oplus rel_p(q,e) \\ &= Exp(P(q,e)) \\ &= \sum_z \omega_p(q,e) P(q|z) P(e|z) P(z)/p(e) \end{aligned} \tag{4.7}$$

where Exp stands for the expectation of all the documents.

4.4 Entity Popularity

The popularity of a Web page is generally expressed by the number of clicks to that page and the number of external links to that page. A lot of previous work, such as TrustRank (Gyöngyi et al. 2004), PageRank (Tomlin 2003) and HITS (Kleinberg 1999), has been deeply introduced a popularity-based ranking. Web page popularity can be concluded as (1) the number of hits; (2) the number of visitors; (3) the number of page views.

Intuitively, entity popularity is defined as (1) the importance of an entity in each document; (2) the frequency of an entity in each document and (3) the number of documents which contain this entity. Here I first argue that both the associations among the entities and the documents are considered by relevance, instead of the hyperlinks and authorities

Input

Query $q = \{e_{q1}, ..., e_{qk}\}$ over Web pages collections

$\mathfrak{D} = \{d_1, ..., d_m\}$;

Output

Ranked named entity list of $E_q = \{e_{q1}, ..., e_{qr}\}$ given query q.

Relevance

Proposed: $relevance(q, e) = rel_T(q, e) \oplus rel_p(q, e)$;

Global: $rel_T(q, e) =< P(q|z), P(z|e) >$ where $P(q|z)$ and

$P(z|e)$ are computed by the proposed topic model

in Section 4.6 respectively;

Local: $rel_p(q, e) = \omega_p(q, e) = \frac{1}{\mathbf{K}} \sum_{i=1}^{N} \delta_d(i, q) ker(q, e)$

with $\mathbf{K} = \sum_{\mathbf{i=1}}^{\mathbf{N}} \mathbf{ker(q, e)}$ and four kernel functions

will be applied in Section 4.5.

Figure 4.2: Relevance Integration Function

among the documents. Second, the frequency of an entity occurrence reflects the popularity of the entity in a document. Third, the hits of an entity are counted by how many documents contain this entity.

Entity importance is contributed by entity relevance, which is measured by the local

relevance (the entity-based proximity model) and the global enhancer (the entity-based topic model). For entity frequency, the term frequency is calculated in Equation 4.3 by obtaining an e_i-contained document set $D_{e_i} = \{d_{1,e_i}, .., d_{n',e_i}\}$ and its frequency vector $TF_{e_i=\{tf_{1,e_i},..,tf_{n',e_i}\}}$. Furthermore, no normalization has been made when the relevance in Equation 4.7 is computed, which shows that the entities with more documents will have more opportunity to be ranked with higher weights. In summary, the proposed approach integrate entity popularity and entity relevance beautifully.

4.5 Entity-based Proximity

Here a probabilistic proximity solution is defined to weight the local relevance between query entities and the entities in the collection. Beyond the simple occurrence matching of query terms and documents, entity position information is taken into account, i.e. the distance to compute the dependency. It is easy to calculate the proximity, when the query only contains an entity. Therefore, an embedded N-gram model is proposed for multiple entities in a query. This is the most unique part in the entity-based proximity model.

4.5.1 Entity-centred Representation

Since focused on the entity ranking problem, and in order to reduce the complexity of the proximity algorithm, an entity-centred representation is formulated. The distance between query entities and the extracted entities is computed in the documents. Then,

the distance $|q - e|$ between the positions where the query term and the entity occur can be:

$$\omega_p(q|e,d) = \frac{1}{\mathbf{K}} \sum_{i=1}^{N} \delta_d(i,q) ker(q,e)$$
$$\mathbf{K} = \sum_{i=1}^{N} ker(q,e) \tag{4.8}$$

4.5.2 Embedded N-gram Model

An embedded N-gram model is proposed for multiple entities in the query, based on the formula in Equation 4.8. For Query $q = \{e_{q1}, ..., e_{qk}\}$, the unigram, bigram and trigram entities are centered around, according to the previous research in (Hu et al. 2012a) where term associations among the query terms are discussed.

The entity sequences of $q = \{e_{q1}, ..., e_{qk}\}$ is represented as $e^n_{q_i} = e_{q_i} e_{q_{i+1}}, ..., e_{q_{i+n-1}}$, where $n = \{1, 2, 3\}$. Their chain rule of probability is in Equation 4.9.

$$\begin{aligned} P(e^n_{q_i}) &= P(e_{q_i}) P(e_{q_{i+1}}|e_{q_i}) P(e_{q_{i+2}}|e^2_{q_i}) ... P(e_{q_{i+n-1}}|e^{n-1}_{q_i}) \\ &= \prod_{j=1}^{n} P(e_{q_{ij}}|e^{j-1}_{q_i}) \end{aligned} \tag{4.9}$$

The bigram and trigram approximations are computed by their appearances in the document set $D_{e^n_{q_i}}$ containing $e_{q^n_i}$ as Equation 4.10, instead of using those complicated models such as Markov model (Andrieu et al. 2003) , in order to satisfy the response time of the system.

$$P(e^n_{q_i}) = \frac{tf^n_{e_{q_i}}}{\sum_{d_{j,i} \in D_{e^n_{q_i}}} tf_{j,e^n_{q_i}}} \tag{4.10}$$

The corresponding entity-centred representation is rewritten in Equation 4.11, where the distance $|q-e|$ is redefined as well and n equals $\{1,2,3\}$.

$$\omega_p(e^n_{q_i}|e,d) = \frac{1}{\mathbf{K}}\sum_{i=1}^{N}\delta_d(i,e^n_{q_i})ker(e^n_{q_i},e)$$
$$\mathbf{K} = \sum_{i=1}^{N} ker(e^n_{q_i},e) \tag{4.11}$$
$$|q-e| = |e^n_{q_i}-e| = P(e^n_{q_i})\sum_{j=1}^{n}(|e_{q_{i+j-1}}-e|)/n$$

4.5.3 Kernel Functions

The previous work of (Zhao et al. 2011), (Lv and Zhai 2009) and (Petkova and Croft 2007) suggested various kernel functions for different data sets. Here four kernel functions are employed to compute the entity pair (q_i, e_j) relationships. The kernel functions satisfy continuous and symmetric properties (Petkova and Croft 2007) such that four kernel functions are presented as Gaussian kernel, Triangle kernel, Circle kernel and Cosine kernel in Equation 4.12, 4.13, 4.14 and 4.15. They are also presented graphically in Figure 4.3.

Gaussian kernel

$$ker(q,e) = exp[\frac{-\mu^2}{2\sigma^2}] \tag{4.12}$$

where the mean μ is the distance of $|q-e|$ and the variance σ can be interpreted as the distance within which I expect to find words that describe the entity in a

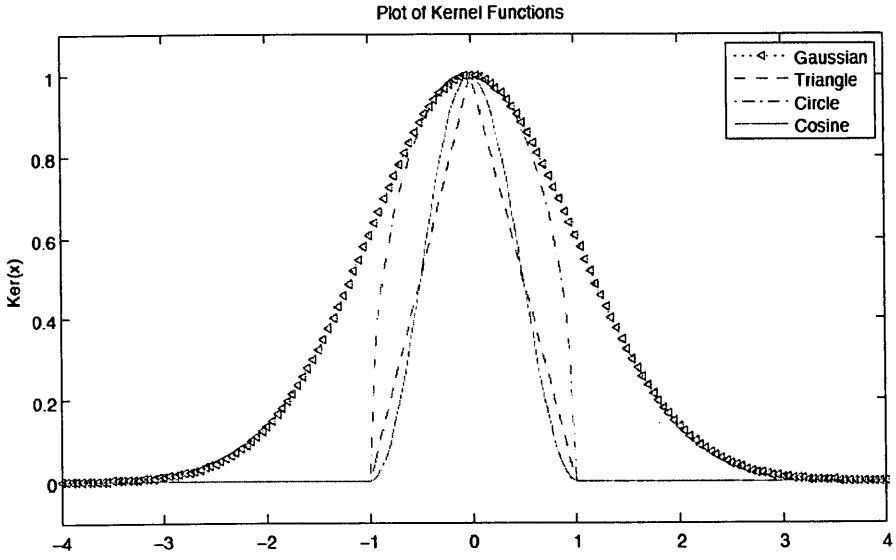

Figure 4.3: Plots of Four Kernel Functions

reliable way, since all positions are assigned non-zero probability and the interval is varied by tuning σ. The variance σ controls how quickly the curve tails off.

Triangle kernel

$$ker(q, e) = (1 - \frac{\mu}{\sigma}) \cdot \mathbf{1}_{\mu \leq \sigma} \tag{4.13}$$

where μ is the distance of $|q - e|$ and σ is the parameter to tune which controls the spread of kernel curves. $\mathbf{1}_{\mu \leq \sigma}$ is an indicator function as

$$\mathbf{1}_{\mu \leq \sigma} = \begin{cases} 1, & \text{if } \mu \leq \sigma; \\ 0, & \text{otherwise.} \end{cases}$$

Circle kernel

$$ker(q, e) = \sqrt{1 - (\frac{\mu}{\sigma})^2} \cdot \mathbf{1}_{\mu \leq \sigma} \tag{4.14}$$

where μ is the distance of $|q - e|$ and σ is the parameter to tune which controls the spread of kernel curves. $\mathbf{1}_{\mu \leq \sigma}$ is the same indicator function as that in triangle kernel.

Cosine kernel

$$ker(q, e) = \frac{1}{2}[1 + cos(\frac{\mu\pi}{\sigma})] \cdot \mathbf{1}_{\mu \leq \sigma} \tag{4.15}$$

where μ is the distance of $|q - e|$ and σ is the parameter to tune which controls the spread of kernel curves. $\mathbf{1}_{\mu \leq \sigma}$ is the same indicator function as that in triangle kernel.

In general, the optimal setting of σ for query terms and entities may vary and also depend on different queries, since terms presumably would have various semantic scopes in documents. These options are not explored, although in principle it could be allowed. Hence, here σ is set as a normalization parameter for the queries and entities in the experiments.

4.6 Entity-based Topic Model

An entity-based topic model is proposed as a global enhancer to the local proximity. In order to estimate the hidden variables, an EM algorithm is adopted to fit the topic model

.

4.6.1 Topic Model

PLSA (Hofmann 1999) is a generative statistical latent class model for general co-occurrence data which associates an unobserved latent variable $z \in \mathrm{Z} = \{z_1, ..., z_k\}$ with each observation of an entity in a document. Formally, PLSA can be defined as (1) select a document d with probability $P(d)$; (2) pick a letent variable z with probability $P(z|d)$ and (3) generate a entity e with probability $P(e|z)$, where

$$P(e|d) = \sum_z P(e|z)P(z|d) \tag{4.16}$$

Therefore, for an observed pair (d, e), the joint probability model results in Equation 4.17.

$$\begin{aligned} P(d,e) &= P(d)P(e|d) \\ &= P(d)\sum_z P(e|z)P(z|d) \\ &= \sum_z P(e|z)P(d|z)P(z) \end{aligned} \tag{4.17}$$

Following, the likelihood principle, one determines $P(d)$, $P(e|z)$ and $P(z|d)$ by maximizing the log-likelihood function

$$\begin{aligned}\mathcal{L} &= \sum_{d \in D} \sum_{e \in E} n(d,e) log P(d,e) \\ &= \sum_{d \in D} \sum_{e \in E} n(d,e) log \sum_{z} P(e|z) P(d|z) P(z)\end{aligned} \tag{4.18}$$

where $n(d,e)$ denotes the term frequency of the entitiy e occurred in the document d.

4.6.2 Topic Model Fitting with EM

The standard procedure for maximum likelihood estimation in latent variable models is the Expectation Maximization (EM) algorithm (Gupta and Chen 2011). EM alternates two steps as (1) an expectation (E) step in Equation 4.19, where posterior probabilities are computed for the latent variables z, based on the current estimates of the parameters; (2) an maximization (M) step in Equations 4.20, 4.21 and 4.22, where parameters are updated for given posterior probabilities computed in the previous E-step.

$$P(z|d,e) = \frac{P(z)P(d|z)P(e|z)}{\sum_{z'} P(z')P(d|z')P(e|z')} \tag{4.19}$$

$$P(e|z) = \frac{\sum_{d} n(d,e)P(z|d,e)}{\sum_{e'} \sum_{d} n(d,e')P(z|d,e')} \tag{4.20}$$

$$P(d|z) = \frac{\sum_{e} n(d,e)P(z|d,e)}{\sum_{d'} \sum_{e} n(d',e)P(z|d',e)} \tag{4.21}$$

$$P(z) = \frac{1}{R}\sum_{d}\sum_{e} n(d,e)P(z|d,e),$$
$$R = \sum_{d}\sum_{E} n(d,e) \tag{4.22}$$

4.7 Experiments

The empirical results of the proposed entity retrieval approach have been reported with different configurations in EntityCube. The data sets, queries and evaluation measures are first described. Then, the proposed model (proximity + topic model) are compared with both the non-proximity topic model (TF-IDF) and the proximity model without applying the topic model (proximity). The advantages of the proposed approach are shown in EntiyCube.

4.7.1 Data Sets and Queries

EntityCube contains more than 3 billion Web pages, where 207,701,938 pages are in English(Liu et al. 2010a, Zhu et al. 2009). In the pre-processing stage, a vision-based page segmentation algorithm (VIPS) (Cai et al. 2003) is applied to parse the Web pages and keep informational blocks instead of the whole Web page (Nie et al. 2005) for building the index. The index contains word term frequency (TF), document frequency (DF), word position information, inverted index (from word to page) and so on. Entity information is also included in the index. Most of the real-world entities have been labelled

with types, such as people, locations and organization. Note that entity type frequency is calculated, for example, "Washington" is extracted as a person by 2000 times and as a location by 1000 times. More details have been described in several related papers published in KDD and WWW (Liu et al. 2010a, Zhu et al. 2009).

In order to evaluate the robustness of our proposed approach, the queries of the TREC 2009 and 2010 Entity Tracks are adopted, instead of our own queries. Table 4.1 presents an example as follows.

Table 4.1: Query Example

⟨query⟩
⟨num⟩1⟨/num⟩
⟨entity_name⟩Blackberry⟨/entity_name⟩
⟨target_entity⟩organization⟨/target_entity⟩
⟨narrative⟩Carriers that Blackberry makes phones for.
⟨/narrative⟩
⟨/query⟩

4.7.2 Evaluation Measures

The Mean Average Precision (MAP) (Balog et al. 2010) is used to evaluate the experimental results. Note that all MAP results stand for the scores over top 1000 entities in this

part. To emphasize on the top retrieved entities, $P@5$, $P@10$ and $P@20$ are highlighted as the evaluation measures. All statistical tests are based on Wilcoxon Matched-pairs Signed-rank test.

4.7.3 Results

The experimental results are presented in Table 4.2. First, term frequency and inverse document frequency (TF-IDF) is applied as a comparison to proximity, in order to investigate the influence of proximity in EntityCube. Second, there are four runs for proximity without topic model and eight runs for proximity with topic model, since four kernel functions (Gaussian, triangle, circle and cosine) are applied in proximity in Section 4.5 and two topic numbers (topics=10 and topics=50) are tested in topic model. Third, all the results are evaluated by MAP, $P@5$, $P@10$ and $P@20$. Fourth, the values in the parentheses are the relative rates of improvement over TF-IDF. Fifth, the best result obtained on each run is marked bold.

The above table shows that the proposed approach outperforms proximity without applying topic model. The advantages of proximity with topic model are especially on $P@5$, $P@10$ and $P@20$. Furthermore, the Gaussian kernel generates the better results than the rest of three kernels.

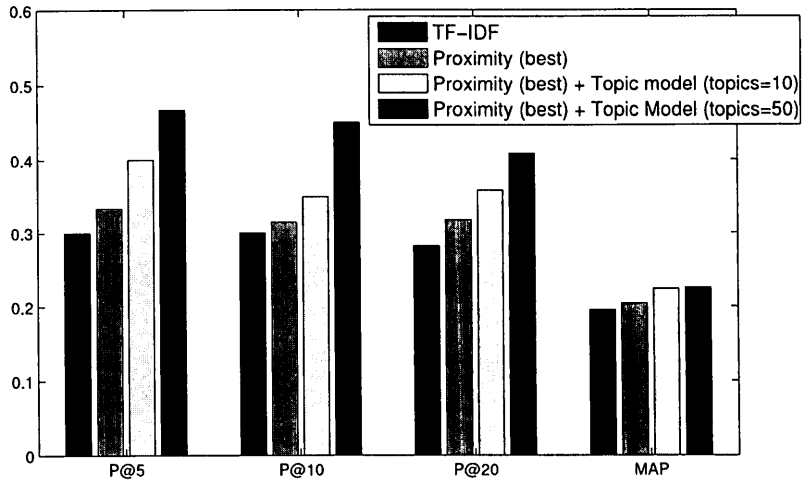

Figure 4.4: Results Comparisons of TF-IDF, Proximity with Gaussian Kernel, Proximity with Gaussian Kernel and Topic Model

4.8 Discussion and Analysis

The experimental results are further discussed and analyzed how they support the proposed approach. First, the influence of proximity is analyzed without considering topic model in Section 4.8.1, followed by the influence of topic model in Section 4.8.2, where a case will be presented to show how proximity and topic model work in details. Then, the investigations are conducted on four kernels in proximity.

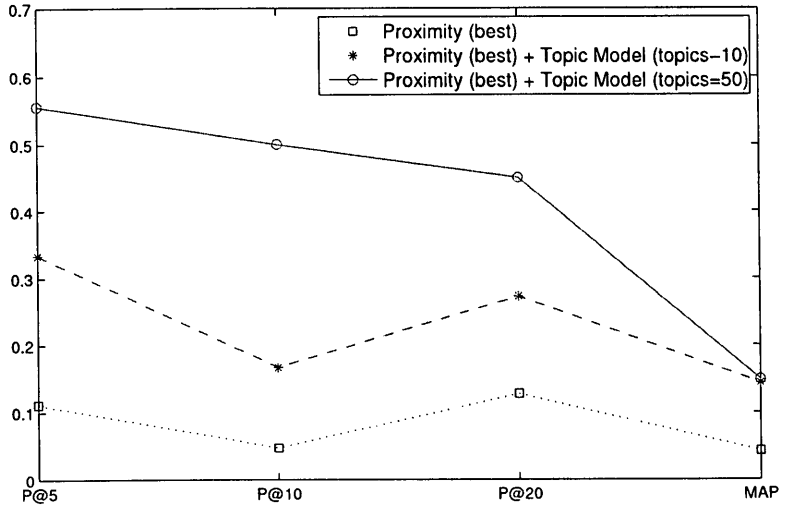

Figure 4.5: Improvements of Proximity with Gaussian Kernel, Proximity with Gaussian Kernel and Topic Model, over TF-IDF

4.8.1 Influence of Proximity

In order to investigate the influence of proximity, the experimental results of proximity without topic model is discussed in Section 4.7.3. TF-IDF is adopted as a comparison.

To illustrate the results in Table 4.2 graphically, the data are re-plotted in Figure 4.4. The x-axis represents the evaluation measures, P@5, P@10, P@20 and MAP over top 1000 entities. The y-axis represents the numerical performance of TF-IDF, the best run of proximity without topic model and the best one of proximity with topic model. It is observable that proximity outperforms TF-IDF.

The better performance of proximity than that of TF-IDF could be interpreted through their definitions. TF-IDF is a numerical statistic which can reflect the importance of a term to a document in a collection or corpus, where TF presents the number of times a term occurs in a document, and IDF helps control for the fact that some terms are generally more common than others. According to the definitions of relevance in Equation 4.3 and proximity in Section 4.5, TF has been counted in Equation 4.3, whereas IDF is not taken into account because of the specific characteristics of entity retrieval compared to text retrieval. Entities are not that common, unlike some words (e.g. "the"). In Table 4.1, the query sample is well defined and very supportive. Furthermore, proximity gets rewards if two entities are close to each other within a defined window (σ).

4.8.2 Influence of Topic Model

In Figure 4.4, another conclusion is drawn that the performance of proximity with topic model is better than that of proximity without applying topic model. Inherently, proximity with topic model also outperforms TF-IDF.

In order to demonstrate our conclusion clearly, a query "Jefferson Airplane" is taken as an example to show how proximity and topic model work. Table 4.3 presents the query descriptions, followed by the top 5 entities retrieved by TF-IDF, proximity without topic model and proximity with topic model. Here the Gaussian kernel is applied. "Y" stands for relevant to the query and "N" for non-relevant. First, all the listed people are

famous and active during the same period (1965-present), contribute a lot on rock, folk rock and other related fields. Second, no wonder that "Bob Dylan" achieves top 1 in TF-IDF, because of his reputation and the huge number of his global fans. Third, proximity computes term associations within a window (the local relevance), which explains the existences of "David Bowis" and "Neil Young", since the members of "Jefferson Airplane" are always mentioned at the same breath with them. Fourth, Topic model considers terms associations over documents (the global enhancer). Under this case, topic model groups "Kantner", "Slick", "Kaukonen" and "Balin" under a hidden "topic" which it can be explained as "members". Unfortunately, topic model still fails to get rid of "Bob Dylan" as a noise.

4.8.3 Influence of Kernels

Four kernel functions as Gaussian, triangle, circle and cosine, are applied in proximity. The kernel parameter σ controls the range of a query's impact. Only a few terms (e.g. close neighbors) are interactive if σ is small. The impact is decreased if σ is large. Hence, σ is chosen to be 20, based on the research by Zhao et.al. (Zhao et al. 2011).

Table 4.2 is re-plotted graphically in Figure 4.6 and 4.7. They show that Gaussian always achieves the best results on $P@5$, $P@10$ and $P@20$, under the parameter setting. The same conclusion has been drawn by Lv and Zhai (Lv and Zhai 2009). However, there is almost no difference among four kernels on MAP over top 1000 entities.

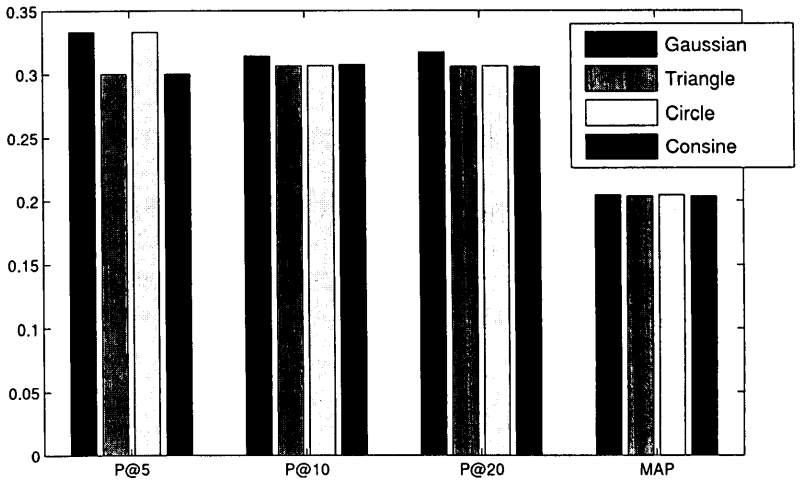

Figure 4.6: Performance of Proximity with Four Kernel Functions: Gaussian achieves the best.

Table 4.2: Results of TF-IDF, Proximity without Topic Model, Proximity with Topic Model

Models	Kernels	Topics #	P@5	P@10	P@20	MAP
TF-IDF	N/A	N/A	0.3000	0.3000	0.2815	0.1958
Proximity without Topic Model	Gaussian	N/A	0.3000	**0.3141**	**0.3173**	**0.2042**
			(0.00%)	(4.70%)	(12.72%)	(4.29%)
	Triangle		0.3000	0.3064	0.3058	0.2032
			(0.00%)	(2.13%)	(8.63%)	(3.78%)
	Circle		**0.3333**	0.3067	0.3062	0.2041
			(11.10%)	(2.23%)	(8.77%)	(4.24%)
	Cosine		0.3000	0.3073	0.3058	0.2031
			(0.00%)	(2.43%)	(8.63%)	(3.73%)
Proximity with Topic Model	Gaussian	topics=10	**0.4000**	0.3167	**0.3583**	**0.2239**
			(33.33%)	(5.57%)	(27.28%)	(14.35%)
	Triangle		0.3000	0.3167	0.3083	0.2225
			(0.00%)	(5.57%)	(9.52%)	(13.64%)
	Circle		0.3000	**0.3167**	0.3333	0.2235
			(0.00%)	(5.57%)	(18.40%)	(14.15%)
	Cosine		0.3333	0.3167	0.3083	0.2224
			(11.10%)	(5.57%)	(9.52%)	(13.59%)
Proximity with Topic Model	Gausiian	topics=50	0.4333	**0.4167**	**0.4083**	**0.2249**
			(44.43%)	(38.90%)	(45.04%)	(14.86%)
	Triangle		0.4000	0.4000	0.3417	0.2237
			(33.33%)	(33.33%)	(21.39%)	(14.25%)
	Circle		**0.4333**	**0.4167**	0.3833	0.2248
			(44.43%)	(38.90%)	(36.16%)	(14.81%)
	Cosine		0.4000	**0.4167**	0.3250	0.2247
			(33.33%)	(38.90%)	(15.45%)	(14.76%)

Table 4.3: Case study: how proximity with topic model outperforms proximity without topic model and TF-IDF

(1) the Gaussian kernel is applied; (2) "Y" stands for relevant to the query and "N" for non-relevant; (3) top 5 retrieved entities are presented.

⟨entity_name⟩Jefferson Airplane⟨/entity_name⟩

⟨target_entity⟩person⟨/target_entity⟩

⟨narrative⟩Members of the band Jefferson Airplane.⟨/narrative⟩

TF-IDF		Proximity		Prox.+Topic M.	
bob dylan	N	david bowie	N	kantner	Y
janis joplin	N	neil young	N	slick	Y
grace slick	Y	grace slick	Y	bob dylan	N
jorma kaukonen	Y	jorma kaukonen	Y	kaukonen	Y
jimi hendrix	N	kaukonen	Y	balin	Y

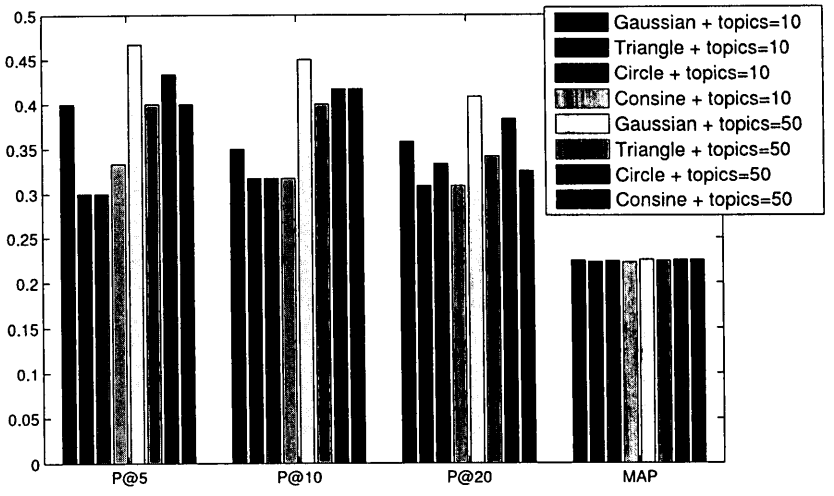

Figure 4.7: Performance of Proximity with Four Kernel Functions and Topic Model

(1) Gaussian achieves the best; (2) topic numbers are set to be 10 and 50 separately, topics=50 outperforms topics=10.

5 Multi-source Fusion

First, a baseline combination problem is formally presented. Then, three modified methods of reciprocal, CombMNZ and CombSUM are introduced respectively. After that, a brief review is given for three IR models of DFR, BM25 and language model. Finally, the experimental results and discussions are shown in the results and discussion section, where the IR environment is introduced with the descriptions of the data sets, queries and evaluation measures. The comprehensive empirical study includes the analysis for the baselines, the proposed approach, the comparisons of CombMNZ and CombSUM to reciprocal, and the influence of the proposed approach on the single model BM25.

5.1 Problem Definition

A multi-source fusion approach is explored for a metasearch system, where the metasearch approach has access to multiple IR systems that retrieve and rank documents/passages with their own models. Therefore, there is an interesting scenario in which the proposed approach only concerns the baselines retrieved by the IR models and then re-rank the

results as the output for evaluation.

For simplicity, throughout this work, the proposed approach works on three kind of baselines: (1) a DFR baseline, B_1; (2) a BM25 baseline, B_2 and (3) a language model baseline, B_3. Furthermore, these baselines are selected from the official submissions of the TREC 2007 Genomics Track. In addition, considered the performance range and effectiveness of the baselines, more than a base run with the higher/lower performance is chosen. Since DFR is often used in fusion as one of the components, there is only a run named "UniNE1" from University of Neuchatel (Fautsch and Savoy 2007) which used DFR as a single model but did not combine many other models. Hence, "UniNE1" is as a seed B_1 of DFR in the proposed metasearch system. For BM25, two baselines are chosen as "MuMshFd", B_{21} from University of Melbourne (Stokes et al. 2007) and "york07ga2", B_{22} from York University (Huang et al. 2007). Two language model baselines are "UBexp1", B_{31} from University Buffalo (Ruiz et al. 2007) and "kyoto1", B_{32} from Kyoto University (Wan et al. 2007).

Hence, given a query q, I put all retrieval documents by three baselines B_1, B_{2i} and B_{3j} (where $i, j = 1, 2$) as D, the corresponding weights of the documents as R. Based on the combination methods, reciprocal, CombMNZ and CombSUM, the proposed approach re-ranks the documents/passages as the new output.

5.2 Reciprocal

The intuition in choosing reciprocal as the formula in Equation 5.1, derives from the fact of an exponential function, while highly ranked documents are more important than the lower ranked documents.

Reciprocal simply sorts the documents according to a naive scoring formula. Given a set D of documents to be ranked and a set of rankings R, for each permutation on $1..|D|$, there is

$$Reciprocal_{score}(d \in D) = \sum_{r \in R} \frac{1}{k + r(d)} \tag{5.1}$$

where $r(d)$ stands for the weight of the document, and the constant k mitigates the impact of high weights. $k = 60$ (Itakura and Clarke 2009) is fixed during a pilot investigation and not altered during subsequent validation, which will not be discussed because of the limit space.

5.3 CombMNZ

Fox and Shaw (Fox and Shaw 1994) introduced several combination methods such as CombMax, CombMin, CombSUM, CombANZ, CombMNX and CombMed, and they found CombSUM to be the best performing combination method. Lee (Lee 1995) conducted extensive experiments with Fox and Shaw combination method based on the TREC data, and he found CombMNZ emerges as the best combination method. Here

CombMNZ is applied in the proposed approach as part of the proposed fusion framework.

CombMNZ requires for each r a corresponding scoring function $s_r : D \rightarrow R$ and a cutoff rank c which all contribute to the CombMNZ score:

$$\text{CMNZ}_{score}(d \in D) = |\{r \in R | r(d) \leq c\}| * \sum_{\{r|r(d)\leq c\}} s_r(d) \tag{5.2}$$

5.4 CombSUM

As one of the famous combination methods proposed by Fox and Shaw (Fox and Shaw 1994), CombSUM is defined as the summation of the set of similarity values, or, equivalently, the numerical mean of the set of the set of similarity values. In (Fox and Shaw 1994), the CombSUM method made the significant improvements over all the baselines such that CombSUM is claimed to perform better than the rest of other methods such as CombMIN, CombANZ on the TREC-2 data set. In the image retrieval domain, Chatzichristofis et al. (Chatzichristofis and Arampatzis 2010) also proved that the CombSUM method was beneficial to improve image information retrieval performance. The CombSUM method is employed to evaluate its effectiveness on the genomics domain.

5.5 IR Systems

A brief review for three well-known weighting models has been given as the Okapi BM25 (Beaulieu et al. 1997), language model (Ponte and Croft 1998, Zhang et al. 2009), and DFR (Amati 2003).

5.5.1 Divergence From Randomness

$$w(d,t) = qtw(t) \cdot IG \cdot (-log_2 Prob(tf)) \tag{5.3}$$

where *IG* is the information gain, which is given by a conditional probability of success of encountering a further token of a given word in a given document on the basis of the statistics on the retrieved set. *Prob(tf)* is the probability of observing the document d given tf occurrences of the query term t. $-log_2 Prob(tf)$ measures the amount of information that term t carries in d. *qtw* is the query term weight component. Similarly to the query model in language modeling (Ponte and Croft 1998), *qtw* measures the importance of individual query terms. In the DFR framework, the query term weight is given by:

$$qtw(t) = \frac{qtf(t)}{qtf_{max}} \tag{5.4}$$

where *qtf(t)* is the query term frequency of t, namely the number of occurrences of t in the query. qtf_{max} is the maximum query term frequency in the query.

The other two components, namely information gain (IG) and information amount

$(-log_2 Prob(tf))$, can be approximated by different statistics so that various instantiations of DFR are implemented.

5.5.2 Okapi BM25

$$\omega = \frac{(k_1 + 1) * tf}{k_1 * ((1 - b) + b * dl/avdl) + tf} * \log \frac{N - n + 0.5}{n + 0.5} * \frac{(k_3 + 1) * qtf}{k_3 + qtf} \tag{5.5}$$

where w is the weight of a query term, N is the number of indexed documents in the collection, n is the number of documents containing the term, R is the number of documents known to be relevant to a specific topic, r is the number of relevant documents containing the term, tf is within-document term frequency, qtf is within-query term frequency, dl is the length of the document, $avdl$ is the average document length, nq is the number of query terms, the k_is are tuning constants (which depend on the database and possibly on the nature of the queries and are empirically determined), K equals to $k_1 * ((1 - b) + b * dl/avdl)$.

5.5.3 Language Model

$$\omega = (1 + \frac{\mu}{1 - \mu} * \frac{tf * FreqTotColl}{l * F_t}) \tag{5.6}$$

where w is the weight of a query term, tf is within-document term frequency, $FreqTotColl$ is within-collection term frequency, l is document length, F_t is length of the whole collection, the mu is tuning constants.

5.6 IR Environment

5.6.1 Data Sets and Queries

The model and algorithms have been evaluated on the 2007 and 2006 TREC data sets. The TREC 2007 and 2006 Genomics data sets provide a test collection of 162,259 full-text documents assembled with 36 queries in 2007 and 28 queries in 2006. The TREC 2007 queries are in the form of questions asking for lists of specific entities. The definitions for these entity types are based on controlled terminologies from different sources, with the source of the terms depending on the entity type (Hersh et al. 2007). The TREC 2006 queries are derived from the set of biologically relevant questions based on the Generic Topic Types (GTTs) (Hersh et al. 2005). There is a sample query as Query 200 as "What serum [PROTEINS] change expression in association with high disease activity in lupus?". More information is available on the official genomics website at: http://ir.ohsu.edu/genomics.

5.6.2 Evaluation Measures

The TREC Genomics Track has three evaluation measures that are the document-level, the aspect-level and the passage2-level (a new measure for the TREC 2007 queries) (Hersh et al. 2007). Each of these provides insight into the overall performance for a user trying to answer the given queries and measured by some variant of mean average

precision (MAP), which are briefly described as follows.

Document-level This is a standard IR measure. The precision is measured at every point where a relevant document is obtained and then averaged over all relevant documents to obtain the average precision for a given query. For a set of queries, the mean of the average precision for all queries is the mean average passage precision of that IR system.

Aspect-level A question could be addressed from different aspects. For example, the question "what is the role of gene PRNP in the Mad cow disease?" could be answered from aspects like "Diagnosis", "Neurologic manifestations", or "Prions/Genetics". This measure indicates how comprehensive the question is answered (Hersh et al. 2006).

Passage2-level This is a new character-based MAP measure which is added to compare the accuracy of the extracted answers and modified from the original measure Passage MAP. Passage2 treats each individually retrieved character in published order as relevant or not, in a sort of "every character is a mini relevance-judged document" approach (Hersh et al. 2007). This is done to increase the stability of the passage MAP measure against arbitrary passage splitting techniques.

5.7 Results and Discussion

5.7.1 Performance of Official Baselines

Table 5.1 presents the performance of five selected baselines which are the official submissions in the TREC 2007 Genomics Track. The models applied in each baseline are specified in the parentheses as "DFR", "BM25" and "LM". Here "LM" stands for "language model". I can see that "MuMshFd" and "UBexp1" have better performance than "york07ga2" and "kyoto1". These baselines are chosen in a performance range in order to check what kind of combination will be most effective. More details will be discussed in the following sections.

5.7.2 Influence of Reciprocal

Corresponding to the baselines, the combinations applying the reciprocal method are evaluated. Due to three kind of IR models, there are four combinations as listed in Table 5.2. Each combination contains a DFR baseline, a BM25 baseline and a LM baseline. The values in the parentheses are the relative rates of improvement over the best results of the baselines.

First, the reciprocal method works very well on the passage2-level and the aspect-level, while it does not contribute a lot on the document-level. Second, "UniNE1+MuMshFd+UBexp1" achieves the best performance, especially in terms of the passage2-level. As noted in

Table 5.1: Baseline Performance

(1) The baselines are the official submissions in the TREC 2007 Genomics Track; (2) the model applied in each baseline is specified in the parentheses as "DFR", "BM25" and "LM". Here "LM" stands for "language model".

baseline	document	aspect	passage2
UniNE1 (DFR)	0.2777	**0.2189**	**0.0988**
MuMshFd (BM25)	**0.2906**	0.2068	0.0895
york07ga2 (BM25)	0.2150	0.1306	0.0472
kyoto1 (LM)	0.1892	0.1208	0.0209

Table 5.1, "MuMshFd" and "UBexp1" have better performance than "york07ga2" and "kyoto1". It can be observed that the alliance of giants is the winner on all the measures. In addition, for the overall performance on the passage2-level, the performance generated by the alliance of giants "UniNE1+MuMshFd+UBexp1", almost catches up with the top official automatic run, "NLMfusion" (Demner-Fushman et al. 2007). Note that "NLMFusion" is an automatic run obtained by five baselines, instead of three in the experiments.

In Table 5.2, both "UniNE1+MuMshFd+UBexp1" and "UniNE1+York07ga2+UBexp1" make improvements in terms of the passage2-level and the aspect-level. Focusing on the passage2-level, the different components of these two combinations are the BM25 baselines, "york07ga2" and "MuMshFd". Then it is argued that the language model "UBexp1" contributes more than the BM25 model "MuMshFd" in the proposed approach. This conclusion can also be confirmed by comparing "UniNE1+York07ga2+UBexp1" with "UniNE1+MuMshFd+kyoto1", in which the latter one has better performance than the preceding one.

Furthermore, a common conclusion can also be drawn that the baselines who have better performance effect the combination results more significantly. For example, the alliance of giants "UniNE1+MuMshFd+UBexp1", which has the best DFR run, the best BM25 run and the best language model run, achieves the best fusion result. "UniNE1+MuMshFd+kyoto1" is better than "UniNE1+york07ga2+kyoto1", because

Table 5.2: Reciprocal Performance

(1) Due to three kind of IR models, there are four combinations as listed; (2) each combination contains a DFR baseline, a BM25 baseline and a LM baseline; (3) the values in the parentheses are the relative rates of improvement over the best results of the baselines; (4) One of the conclusions is that the alliance of giants with boldface is the winner on all the measures.

Component	document	aspect	passage2
Best of baselines	0.2906	0.2189	0.0988
UniNE1+York07ga2+kyoto1	0.2743	0.2065	0.0978
	(-5.60%)	(-5.63%)	(-1.01%)
UniNE1+York07ga2+UBexp1	0.2802	0.2219	0.1047
	(-3.56%)	(1.38%)	(5.96%)
UniNE1+MuMshFd+kyoto1	0.2828	0.2221	0.0997
	(-2.66%)	(1.46%)	(0.86%)
UniNE1+MuMshFd+UBexp1	**0.2906**	**0.2380**	**0.1059**
	(0.00%)	**(8.75%)**	**(7.19%)**

"MuMshFd" is better than "york07ga2".

5.7.3 Comparison to CombMNZ

Table 5.3 presents the performance of applying the CombMNZ method. In order to deeply evaluate the benefits of CombMNZ, three versions are introduced as CombMNZ-with-normalization, CombMNZ-with-assigned-weight and CombMNZ-with-multiple respectively. The values in the parentheses are the relative rates of improvement over the best results of the baselines.

In CombMNZ-with-normalization, the standard zero-one normalization method is employed, in which all base weights are scaled between zero being the lowest value and one being the absolute highest value. CombMNZ-with-normalization is the most popular version such that another two versions of CombMNZ are generated to check its effectiveness.

In CombMNZ-with-assigned-weight, the baselines earn their weights depending on their models. For N baselines, different weights are assigned to them linearly, in which the sum of the weights equals to one always. Here the experiments are conducted with tuning the assigned weights. Only the optimal results are presented in Table 5.3.

In CombMNZ-with-multiple, the CombMNZ method is applied for multiple times. In the experiments, m times (where m is set to be one of $\{1, 2, 3, 5\}$) are tested on the baselines. No normalization and additional weights has been given to the baselines. Only

Table 5.3: Performance of CombMNZ

(1) There are three versions, CombMNZ-with-normalization, CombMNZ-with-assigned-weight and CombMNZ-with-multiple; (2) the standard zero-one normalization method is employed in CombMNZ-with-normalization; (3) m is set to be one of $\{1, 2, 3, 5\}$ in CombMNZ-with-multiple; (4) the values in the parentheses are the relative rates of improvement over the best results of the baselines.

Components	w/ Normalization			w/ Assigned Weights			w/ Multiple		
	document	aspect	passage2	document	aspect	passage2	document	aspect	passage2
Best of baselines	0.2906	0.2189	0.0988	0.2906	0.2189	0.0988	0.2906	0.2189	0.0988
UniNE1+York07ga2+kyoto1	0.2671	0.1535	0.0937	0.2729	0.1854	0.0957	0.2571	0.1547	0.0924
	(-8.08%)	(-29.86%)	(-5.13%)	(-6.09%)	(-15.27%)	(-3.19%)	(-11.53%)	(-29.33%)	(-6.49%)
UniNE1+York07ga2+UBexp1	0.2656	0.1772	0.0879	0.2591	0.1878	0.0867	0.2639	0.1753	0.0885
	(-8.61%)	(-19.03%)	(-10.99%)	(-10.82%)	(-14.18%)	(-12.30%)	(-9.16%)	(-19.92%)	(-10.43%)
UniNE1+MuMshFd+kyoto1	0.2559	0.1801	0.0985	0.2503	0.1837	0.0908	0.2401	0.1599	0.0958
	(-11.95%)	(-17.70%)	(-0.30%)	(-13.85%)	(-16.09%)	(-8.06%)	(-17.38%)	(-26.96%)	(-3.04%)
UniNE1+MuMshFd+UBexp1	0.2416	0.1720	0.0871	0.2466	0.1787	0.0839	0.2419	0.1716	0.0872
	(-16.85%)	(-21.43%)	(-11.86%)	(-15.11%)	(-18.36%)	(-15.09%)	(-16.74%)	(-21.61%)	(-11.72%)

the optimal results are presented in Table 5.3 as well.

Although CombMNZ has been confirmed by Lee (Lee 1995), Fox and Shaw (Fox and Shaw 1994) as an effective method. However, in the experiments in the biomedicine domain, CombMNZ does not show any advantage at all, although three different versions have been generated. In Table 5.3, all the combinations get worse compared with the best results of the baselines, especially in terms of the passage2-level and the aspect-level. On the genomics data, reciprocal outperforms CombMNZ thoroughly.

5.7.4 Comparison to CombSUM

Fox and Shaw (Fox and Shaw 1994) proved that the CombSUM method can achieve good performance on the TREC-2 data set. CombSUM is applied as a second comparison to reciprocal, since CombMNZ doesn't work on the genomics data set.

In Table 5.4, CombSUM does not work very well on the baselines. However, the alliance of giants "UniNE1+MuMshFd+UBexp1" outperforms the best baseline on the passage2-level. The CombSUM method has great potential to improve the retrieval performance on multi-source baselines in the genomics domain. Compared to reciprocal, reciprocal outperforms CombSUM on all the measures as well. Although both CombSUM and CombMNZ do not work as well as reciprocal, CombSUM provides its effectiveness better than CombMNZ with the evidence of the improved passage2-level performance.

Furthermore, the application of CombSUM repeatedly confirms that the alliance of

Table 5.4: Performance of CombSUM

The values in the parentheses are the relative rates of improvement over the best results of the baselines.

Component	document	aspect	passage2
Best of baselines	0.2906	0.2189	0.0988
UniNE1+York07ga2+kyoto1	0.2692	0.1552	0.0939
	(-7.36%)	(-29.07%)	(-4.94%)
UniNE1+York07ga2+UBexp1	0.2690	0.1840	0.0944
	(-7.41%)	(-15.94%)	(-4.49%)
UniNE1+MuMshFd+kyoto1	0.2567	0.1809	0.0985
	(-11.66%)	(-17.35%)	(-0.30%)
UniNE1+MuMshFd+UBexp1	0.2630	0.1919	0.0991
	(-9.49%)	(-12.32%)	(0.30%)

giants achieves the best results over the other combinations. In addition, comparing "UniNE1+MuMshFd+kyoto1" with "UniNE1+MuMshFd+UBexp1", the evidences can be observed that there is no big performance gap on all the measures and only a different component between them. Then a conclusion can be drawn that "UBexp1" doesn't contribute much more than "kyoto1", although "UBexp1" outperforms "kyoto1" much. On the other hand, compared "UniNE1+York07ga2+UBexp1" with "UniNE1+MuMshFd+UBexp1", the evidences are also obtained that there exists a big performance gap especially on the passage2-level and only a different component between them. Another conclusion is drawn that "MushMshFd" contributes much more than "York07ga2", since "MushMshFd" has much better performance than "York07ga2".

5.7.5 Influence of the Proposed Approach on the Single Source

In the previous sections, the proposed approach is evaluated on the official multi-source submissions of the TREC 2007 Genomics track. Based on three different models, reciprocal obtains nice performance as a good combination method. The proposed approach is examined how it works based on the single source of Okapi BM25.

First of all, the baselines are from three different indices under the same IR model, BM25, instead of those from three kind of IR models. Second, three indices are built on the 2007 and 2006 genomics data sets according to three passage extraction methods (Hu and Huang 2008, 2010, Huang and Hu 2009). Here "word" stands for "word-base",

Table 5.5: Performance of the Fusion Approach on Okapi 2007 and 2006

(1) The baselines are from three different indices under the same IR model, BM25, instead of those from three kind of IR models; (2) "word" stands for "word-base", "sentence" for "sentence-base" and "paragraph" for "paragraph-base"; (3) the Okapi tuning parameters of the selected runs are $(k_1, b) = (0.5, 1.3)$; (4) the values in the parentheses are the relative rates of improvement over the best results of the baselines.

Components	Okapi 2007			Okapi 2006	
	document	aspect	passage2	document	aspect
word	0.2108	0.1080	0.0364	0.3140	0.1237
sentence	0.1805	0.0970	0.0350	0.3030	0.1206
paragraph	0.1588	0.0616	0.0333	0.3109	0.1410
reciprocal	**0.2219**	**0.1237**	**0.0478**	**0.3168**	**0.1449**
	(5.29%)	**(14.51%)**	**(31.40%)**	**(1.07%)**	**(12.25%)**
CombMNZ-with-normalization	0.1703	0.0643	0.0270	0.2352	0.0498
	(-19.20%)	(-40.43%)	(-25.92%)	(-26.55%)	(-61.46%)
CombMNZ-with-assigned-weights	0.1777	0.0701	0.0273	0.2441	0.0524
	(-15.72%)	(-35.12%)	(-24.88%)	(-23.78%)	(-59.43%)
CombMNZ-with-multiple	0.1730	0.0651	0.0277	0.2375	0.0508
	(-17.93%)	(-39.73%)	(-24.01%)	(-25.85%)	(-60.62%)
CombSUM	0.1818	0.0718	0.0297	0.2559	0.0719
	(-13.76%)	(-33.56%)	(-18.43%)	(-20.10%)	(-44.32%)

"sentence" for "sentence-base" and "paragraph" for "paragraph-base". Third, the Okapi tuning parameters of the selected runs are $(k_1, b) = (0.5, 1.3)$. Similarly, reciprocal, CombMNZ and CombSUM are applied as the same way in the previous experiments. Table 5.5 shows the performance of baselines and combinations in 2007 and 2006 respectively.

In the TREC 2007 Genomics Track overview (Tari et al. 2007), the measure correlation of the four measures shows that the passage2-level is highly correlated with the aspect-level. Therefore, on the 2006 data set, the aspect-level is chosen as the main measure, since there is no passag2-level in 2006. Focused on the passage2-level and the aspect-level, the reciprocal method outperforms CombMNZ and CombSUM obviously in Table 6. The reciprocal method achieves great improvements on the passage2-level, the aspect-level and the document-level on both 2007 and 2006 genomics data sets. The standard normalization method, tuning the assigned weights and using multiple times CombMNZ can not help CombMNZ to make progress on the 2007 and 2006 data sets respectively. CombSUM does not work well on both 2007 and 2006 data sets. However, the consistent conclusion can be drawn that the CombSUM method works slightly well than the CombMNZ method, although both of them are not as good as reciprocal.

6 Conclusions

This thesis work focuses on search beyond probabilistic information retrieval, in which approaches to five open research problems of the traditional probabilistic modelling are proposed. First, term association is deeply examined, whereas most probabilistic models are based on the assumption that query terms are independent of each other and a document is represented as a bag of words. Second, associations are considered at the higher level, including the document/passage level and the entity level. Third, the relevance are not restricted to rather simple forms of inference. In the probabilistic models, only terms among queries or some relevance feedback are considered. In the proposed approaches, information over the documents, within the collection or thesaurus is included. Fourth, the entity-based approach provides entity information which is summarized from multiple documents, instead of calculating probabilities for each documents in the probabilistic models. Fifth, the fusion approach treats the data from different sources, i.e. multiple IR models/functions/formulas.

6.1 Term Association

Term association considering co-occurrence and dependency among the keywords produces better results than the baselines treating the keywords independently. In the other hand, the unigrams, bigrams and trigrams are terms independently computed by the factor analysis based model, which means that the trigrams are not dependent on the bigrams' importance, and the bigrams are not dependent on the unigrams' importance. Their importance is decided by the model and the appearances in the passages. This is also confirmed by the GSP algorithm.

In the term association approach, keywords and the retrieved passages are the observable data, and the factor analysis based model is built up to discover the unobservable latent factors. Factor loadings are computed to indicate the weights of the common factors. Communalities are calculated based on factor loadings to represent the importance and reliance of the corresponding terms associations. Finally, a ranking term association list is given by the model. Then we recursively re-rank the baselines and report the experimental results.

The experimental results show that term association outperforms the baselines and the GSP results on all the evaluation measures, which provides a promising avenue for improving the information retrieval performance.

6.2 Entity Ranking

The conclusion of this part of work is four-fold. First, a novel approach is proposed on searching and ranking entity in EntityCube. Second, entity ranking is investigated by integrating entity relevance and popularity. The entity relevance is modelled by entity proximity as the local relevance and entity topic model as the global enhancer. Third, the proposed approach is well evaluated in EntityCube and is duplicable through sending any query to EntityCube. Fourth, some effective findings are drawn from the experimental results, such as proximity with topic model outperforms TF-IDF and proximity without topic model, the Gaussian kernel in proximity works better than the triangle kernel, the circle kernel and the consine kernel.

6.3 Multi-source Fusion

Empirical study on three different IR models demonstrates the utility of the proposed approach. Compared to the CombMNZ and CombSUM methods, the reciprocal method provides notable improvements using the baselines from a DFR model, a BM25 model and a language model respectively. The improvements are significant for both TREC 2007 and 2006 genomics data set, in which the improved result in terms of the passage2-level in 2007 almost catches up with the highest official result "NLMFusion" Demner-Fushman et al. (2007). While CombMNZ does not achieve good performance, we

conduct three versions as CombMNZ-with-normalization, CombMNZ-with-assigned-weight and CombMNZ-with-multiple to further improve and evaluate the CombMNZ method. Although the CombSUM method does not work as well as reciprocal, Comb-SUM makes progress on the passage2-level, also works better than CombMNZ on all the three versions.

Five baselines are selected from three kind of IR models as DFR, BM25 and language model. The experimental results implement the following conclusions: 1) the alliance of giants achieves the best result; 2) under the same combination, the better the baseline performance is, the more contribution the baseline provides. Furthermore, the proposed robust approach makes improvements not only for combining the baselines from different sources, but also for combining the baselines from the single source such as Okapi BM25.

Bibliography

J. Allan. HARD Track Overview in TREC 2004. In *Proceedings of 13th Text REtrieval Conference*. NIST Special Publication, 2004.

G. Amati. Probabilistic Models for Information Retrieval Based on Divergence From Randomness. *PhD thesis, Department of Computing Science, University of Glasgow*, 2003.

Christophe Andrieu, Nando de Freitas, Arnaud Doucet, and Michael I. Jordan. An Introduction to MCMC for Machine Learning. *Machine Learning*, 50(1-2):5–43, 2003.

Thi Truong Avrahami, Lawrence Yau, Luo Si, and James P. Callan. The FedLemur project: Federated Search in the Real World. *JASIST*, 57(3):347–358, 2006.

Krisztian Balog, Leif Azzopardi, and Maarten de Rijke. Formal Models for Expert Finding in Enterprise Corpora. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '06, pages 43–50, 2006. ISBN 1-59593-369-7.

Krisztian Balog, Pavel Serdyukov, and Arjen P. de Vries. Overview of the TREC 2010 Entity Track. In *Proceedings of 19th Text REtrieval Conference*, 2010.

M. Beaulieu, M. Gatford, Xiangji Huang, Stephen E. Robertson, S. Walker, and P. Williams. Okapi at TREC-5. In *Proceedings of 5th Text REtrieval Conference*, pages 143–166. NIST Special Publication, 1997.

Michael Bendersky and W. Bruce Croft. Modeling Higher-order Term Dependencies in Information Retrieval Using Query Hypergraphs. In *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '12, pages 941–950, 2012. ISBN 978-1-4503-1472-5.

Peter Biebricher, Norbert Fuhr, Gerhard Lustig, Michael Schwantner, and Gerhard Knorz. The Automatic Indexing System AIR/PHYS – From Research to Application. In *Proceedings of the 11st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '88, pages 333–342, 1988.

Roi Blanco and Hugo Zaragoza. Finding Support Sentences for Entities. In *Proceeding of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '10, pages 339–346, 2010. ISBN 978-1-4503-0153-4.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

A. Bookstein and D. R. Swanson. Probabilistic Models for Automatic Indexing. *Journal of the American Society for Information Science*, 25:312–318, 1974.

Abraham Bookstein. Implications of Boolean Structures for Probabilistic Retrieval. In *Proceedings of the 8th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '85, pages 11–17, 1985.

Abraham Bookstein and Gary Fouty. A Mathematical Model for Estimating the Effectiveness of Bigram Coding. *Inf. Process. Manage.*, 12(2):111–116, 1976.

Marc Bron, Krisztian Balog, and Maarten de Rijke. Ranking Related Entities: Components and Analyses. In *Proceedings of the 19th ACM International Conference on Information and knowledge management*, CIKM '10, pages 1079–1088, 2010. ISBN 978-1-4503-0099-5.

Andreas Broschart and Ralf Schenkel. Proximity-aware Scoring for XML Retrieval. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '08, pages 845–846, 2008. ISBN 978-1-60558-164-4.

Stefan Büttcher, Charles L. A. Clarke, and Brad Lushman. Term Proximity Scoring for Ad-hoc Retrieval on Very Large Text Collections. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '06, pages 621–622, 2006. ISBN 1-59593-369-7.

Deng Cai, Shipeng Yu, Ji rong Wen, Wei ying Ma, Deng Cai, Shipeng Yu, Ji rong Wen, and Wei ying Ma. 1 VIPS: a Vision-based Page Segmentation Algorithm, 2003.

James P. Callan, Zhihong Lu, and W. Bruce Croft. Searching Distributed Collections with Inference Networks. In *Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '95, pages 21–28, 1995.

Youngchul Cha and Junghoo Cho. Social-Network Analysis Using Topic Models. In *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '12, pages 565–574, 2012. ISBN 978-1-4503-1472-5.

Rita Chattopadhyay, Jieping Ye, Sethuraman Panchanathan, Wei Fan, and Ian Davidson. Multi-source Domain Adaptation and its Application to Early Detection of Fatigue. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge discovery and Data mining*, KDD '11, pages 717–725, 2011. ISBN 978-1-4503-0813-7.

Savvas A. Chatzichristofis and Avi Arampatzis. Late Fusion of Compact Composite Descriptors for Retrieval from Heterogeneous Image Databases. In *Proceeding of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '10, pages 825–826, 2010. ISBN 978-1-4503-0153-4.

Tao Cheng, Xifeng Yan, and Kevin Chen-Chuan Chang. EntityRank: Searching Entities Directly and Holistically. In *Proceedings of the 33rd International Conference on Very large Data bases*, VLDB '07, pages 387–398, 2007. ISBN 978-1-59593-649-3.

William S. Cooper. A Definition of Relevance for Information Retrieval, journal = Information Storage and Retrieval, volume = 7, number = 1, year = 1971, pages = 19-37,.

William S. Cooper. Some Inconsistencies and Misidentified Modeling Assumptions in Probabilistic Information Retrieval. *ACM Trans. Inf. Syst.*, 13(1):100–111, 1995.

Fabio Crestani, Mounia Lalmas, C. J. van Rijsbergen, and Iain Campbell. "Is This Document Relevant? ... Probably": A Survey of Probabilistic Models in Information Retrieval. *ACM Comput. Surv.*, 30(4):528–552, 1998.

W. Bruce Croft, R. Wolf, and R. Thompson. A Network Organization Used for Document Retrieval. In *Proceedings of the 6th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '83, pages 178–188, 1983.

W. Bruce Croft, James P. Callan, and John Broglio. TREC-2 Routing and Ad-Hoc Retrieval Evaluation Using the INQUERY System. In *Proceedings of The 2nd Text REtrieval Conference*, pages 75–84, 1993.

Scott Deerwester, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, and Richard Harshman. Indexing by Latent Semantic Analysis.

Dina Demner-Fushman, Susanne M. Humphrey, Nicholas C. Ide, Russell F. Loane, James G. Mork, Patrick Ruch, Miguel E. Ruiz, Lawrence H. Smith, W. John Wilbur, and Alan R. Aronson. Combining Resources to Find Answers to Biomedical Questions. In *Proceedings of 16th Text REtrieval Conference*, 2007.

Claire Fautsch and Jacques Savoy. IR-Specific Searches at TREC 2007: Genomics & Blog Experiments. In *Proceedings of 16th Text REtrieval Conference*, 2007.

E. A. Fox and J. A. Shaw. Combination of Multiple Searches. In *TREC-2*, 1994.

Norbert Fuhr. Models for Retrieval with Probabilistic Indexing. *Inf. Process. Manage.*, 25(1):55–72, 1989.

Norbert Fuhr. Probabilistic Models in Information Retrieval. *Comput. J.*, 35(3):243–255, 1992.

Norbert Fuhr and Chris Buckley. A Probabilistic Learning Approach for Document Indexing. *ACM Trans. Inf. Syst.*, 9(3):223–248, 1991.

Norbert Fuhr, Ulrich Pfeifer, C. Bremkamp, and Michael Pollmann. Probalistic Learning Approaches for Indexing and Retrieval with the TREC-2 Collection. In *Proceedings of The 2nd Text REtrieval Conference*, pages 67–74, 1993.

Gregory Grefenstette. Use of Syntactic Context to Produce Term Association Lists for Text Retrieval. In *Proceedings of the 15th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 89–97, 1992. ISBN 0-89791-523-2.

Maya R. Gupta and Yihua Chen. Theory and Use of the EM Algorithm. *Found. Trends Signal Process.*, 4:223–296, March 2011. ISSN 1932-8346.

Zoltán Gyöngyi, Hector Garcia-Molina, and Jan Pedersen. Combating Web Spam with Trustrank. In *Proceedings of the Thirtieth International Conference on Very large Data bases - Volume 30*, VLDB '04, pages 576–587, 2004. ISBN 0-12-088469-0.

Zhen Hai, Kuiyu Chang, and Gao Cong. One Seed to Find Them All: Mining Opinion Features via Association. In *Proceedings of the 21st ACM International Conference on Information and knowledge management*, CIKM '12, pages 255–264, 2012. ISBN 978-1-4503-1156-4.

David Hawking and Paul B. Thistlewaite. Proximity Operators - So Near And Yet So Far. In *Proceedings of 4th Text REtrieval Conference*, 1995.

William Hersh, Aaron Cohen, and Jianji Yang. TREC 2005 Genomics Track Overview. In *Proceedings of 14th Text REtrieval Conference*. NIST Special Publication, 2005.

William Hersh, Aaron M. Cohen, and Phoebe Roberts. TREC 2006 Genomics Track Overview. In *Proceedings of 15th Text REtrieval Conference*. NIST Special Publication, 2006.

William Hersh, Aaron M. Cohen, and Phoebe Roberts. TREC 2007 Genomics Track Overview. In *Proceedings of 16th Text REtrieval Conference*. NIST Special Publication, 2007.

Birger Hjørland and Frank Sejer Christensen. Work Tasks and Socio-cognitive Relevance: A Specific Example. *J. Am. Soc. Inf. Sci. Technol.*, 53(11,).

Thomas Hofmann. 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, 1999. ISBN 1-58113-096-1.

Qinmin Hu and Jimmy X. Huang. A Dynamic Window Based Passage Extraction Algorithm for Genomics Information Retrieval. In *ISMIS 2008, Foundations of Intelligent Systems, 17th International Symposium, May 20-23, 2008, Toronto, Canada*, pages 434–444, 2008.

Qinmin Hu and Jimmy X. Huang. Passage Extraction and Result Combination for Genomics Information Retrieval. *J. Intell. Inf. Syst.*, 34(3):249–274, 2010.

Qinmin Hu, Jimmy X. Huang, and Jun Miao. Exploring a Multi-source Fusion Approach for Genomics Information Retrieval. In *BIBM*, pages 669–672, 2010.

Qinmin Hu, Jimmy X. Huang, and Xiaohua Hu. A Term Association Approach for Genomics Information Retrieval. In *BIBM*, pages 532–537, 2011a.

Qinmin Hu, Jimmy X. Huang, and Jun Miao. A Robust Approach to Optimizing Multi-source Information for Enhancing Genomics Retrieval Performance. *BMC Bioinformatics*, 12(S6):18, 2011b.

Qinmin Hu, Jimmy Xiangji Huang, and Xiaohua Hu. Modeling and Mining Term Association for Improving Biomedical Information Retrieval Performance. *BMC Bioinformatics*, 13(Suppl 9):S2:35, 2012a.

Qinmin Hu, Zaiqing Nie, Jimmy X. Huang, and Nick Cercone. Integrating Proximity Information into Topic Models for Entity Search. *Submitted to Journal of ACM Transactions on Knowledge Discovery from Data (TKDD)*, 2012b.

X. Huang, Y.R. Huang, and M. Wen. A Dual Index Model for Contextual Information Retrieval. In *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, August 15-19, 2005, Salvador, Brazil*, pages 613–614, 2005a. ISBN 1-59593-034-5.

X. Huang, M. Wen, A. An, and Y.R. Huang. A Platform for Okapi-based Contextual Information Retrieval. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, August 6-11, 2006, Seattle, Washington, USA*, pages 728–728, 2006. ISBN 1-59593-369-7.

Xiangji Huang and Qinmin Hu. A Bayesian Learning Approach to Promoting Diversity in Ranking for Biomedical Information Retrieval. In *Proceedings of the 32nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2009, Boston, MA, USA, July 19-23, 2009*, SIGIR '09, pages 307–314, 2009. ISBN 978-1-60558-483-6.

Xiangji Huang, F. Peng, D. Schuurmans, Nick Cercone, and Stephen E. Robertson. Applying Machine Learning to Text Segmentation for Information Retrieval. *Information Retrieval Journal*, 6(4):333–362, 2003.

Xiangji Huang, Ming Zhong, and Luo Si. York University at TREC 2005: Genomics Track. In *Proceedings of the 14th Text Retrieval Conference*, 2005b.

Xiangji Huang, Damon Sotoudeh-Hosseinii, Hashmat Rohian, and Xiangdong An. York University at TREC 2007: Genomics Track. In *Proceedings of 16th Text REtrieval Conference*, 2007.

Kelly Y. Itakura and Charles L. A. Clarke. Using Dynamic Markov Compression to Detect Vandalism in the Wikipedia. In *Proceedings of the 32nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2009, Boston, MA, USA, July 19-23, 2009*, pages 822–823, 2009.

Antonio Jimeno, Piotr Pezik, and Dietrich Rebholz-Schuhmann. Information Retrieval and Information Extraction in TREC Genomics 2007. In *Proceedings of 16th Text REtrieval Conference*, 2007.

Hiroyuki Kaji, Yasutsugu Morimoto, Toshiko Aizono, and Noriyuki Yamasaki. Corpus-dependent Association Thesauri for Information Retrieval. In *Proceedings of the 18th Conference on Computational linguistics*, pages 404–410, 2000. ISBN 1-55860-717-X.

Rianne Kaptein, Pavel Serdyukov, Arjen De Vries, and Jaap Kamps. Entity Ranking Using Wikipedia as a Pivot. In *Proceedings of the 19th ACM International Conference on Information and knowledge management*, CIKM '10, pages 69–78, 2010. ISBN 978-1-4503-0099-5.

Saurabh S. Kataria, Krishnan S. Kumar, Rajeev R. Rastogi, Prithviraj Sen, and Srinivasan H. Sengamedu. Entity Disambiguation with Hierarchical Topic Models. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge discovery and Data mining*, KDD '11, pages 1037–1045, 2011. ISBN 978-1-4503-0813-7.

E. Michael Keen. The use of Term Position Devices in Ranked Output Experiments. *J. Doc.*, 47:1–22, March 1991. ISSN 0022-0418.

E. Michael Keen. Some Aspects of Proximity Searching in Text Retrieval Systems. *J. Inf. Sci.*, 18:89–98, Feb. 1992. ISSN 0165-5515.

Jon M. Kleinberg. Hubs, Authorities, and Communities. *ACM Comput. Surv.*, 31, December 1999. ISSN 0360-0300.

Craig A. Knoblock, Steven Minton, José Luis Ambite, Maria Muslea, Jean Oh, and Martin Frank. Mixed-initiative, Multi-source Information Assistants. In *Proceedings of the 10th International Conference on World Wide Web*, WWW '01, pages 697–707, 2001. ISBN 1-58113-348-0.

Joon Ho Lee. Combining Multiple Evidence from Different Properties of Weighting Schemes. In *Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 180–188, 1995. ISBN 0-89791-714-6.

Jimmy J. Lin and W. John Wilbur. PubMed Related Articles: a Probabilistic Topic-based Model for Content Similarity. *BMC Bioinformatics*, 8, 2007.

Christina Lioma, Ben He, Vassilis Plachouras, and Iadh Ounis. The University of Glasgow at CLEF 2004: French Monolingual Information Retrieval with Terrier. In *CLEF*, pages 253–259, 2004.

Xiaojiang Liu, Zaiqing Nie, Nenghai Yu, and Ji-Rong Wen. BioSnowball: Automated Population of Wikis. In *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge discovery and Data mining*, KDD '10, pages 969–978, 2010a. ISBN 978-1-4503-0055-1.

Yang Liu, Xiaohui Yu, Xiangji Huang, and Aijun An. S-PLASA+: Adaptive Sentiment Analysis with Application to Sales Performance Prediction. In *Proceeding of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2010, Geneva, Switzerland, July 19-23, 2010*, pages 873–874, 2010b.

Yue Lu, Qiaozhu Mei, and ChengXiang Zhai. Investigating Task Performance of Probabilistic Topic Models: an Empirical Study of PLSA and LDA. *Information Retrieval*, 14:178–203, 2011.

Yuanhua Lv and ChengXiang Zhai. Positional language Models for Information Retrieval. In *Proceedings of the 32nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2009, Boston, MA, USA, July 19-23, 2009*, pages 299–306, 2009. ISBN 978-1-60558-483-6.

Alexei Manso Correa Machado, Celina N. J. Marinho, and Mrio Fernando Montenegro Campos. An Image Retrieval Method Based on Factor Analysis. *Computer Graphics and Image Processing, Brazilian Symposium on*, 0:191, 2003. ISSN 1530-1834.

Thomas Mandl. Efficient Preprocessing for Information Retrieval with Neural Networks. In *Proceedings of the 7th European Congress on Intelligent Techniques and Soft Computing*, 1999.

Sukanya Manna and Tom Gedeon. A Term Association Inference Model for Single Documents: A Stepping Stone for Investigation through Information Extraction. In *Proceedings of the IEEE ISI 2008 PAISI, PACCF, and SOCO International workshops on Intelligence and Security Informatics*, PAISI, PACCF and SOCO '08, pages 14–20, 2008. ISBN 978-3-540-69136-5.

C.D. Manning and H. Schlztze. *Foundations of Statistical Natural Language Processing*. MIT Press, 1999.

M. E. Maron and J. L. Kuhns. On Relevance, Probabilistic Indexing and Information Retrieval. *J. ACM*, 7(3):216–244, 1960. ISSN 0004-5411.

Andrew McCallum, Xuerui Wang, and Andrés Corrada-Emmanuel. Topic and Role Discovery in Social Networks with Experiments on Enron and Academic Email. *J. Artif. Int. Res.*, 30:249–272, October 2007. ISSN 1076-9757.

Richard M. C. McCreadie, Craig Macdonald, and Iadh Ounis. CrowdTerrier: Automatic Crowdsourced Relevance Assessments with Terrier. In *Proceedings of the 35th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '12, page 1005, 2012.

B. Mehta, T. Hofmann, and P. Frankhaser. Cross System Personalization by Factor Analysis. In *Proceedings of the 21st National Conference on AAAI*, 2006.

Donald Metzler and W. Bruce Croft. Combining the Language Model and Inference Network Approaches to Retrieval. *Inf. Process. Manage.*, 40(5):735–750, 2004.

Donald Metzler and W. Bruce Croft. A Markov Random Field Model for Term Dependencies. In *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '05, pages 472–479, 2005. ISBN 1-59593-034-5.

Donald Metzler, Trevor Strohman, Howard R. Turtle, and W. Bruce Croft. Indri at TREC 2004: Terabyte Track. In *Proceedings of The 13th Text REtrieval Conference*, 2004.

William L. Miller. A Probabilistic Search Strategy for MEDLARS. *Journal of Documentation*, 27:254–266, 1971.

Seung-Hoon Na and Hwee Tou Ng. A 2-poisson Model for Probabilistic Coreference of Named Entities for Improved Text Retrieval. In *Proceedings of the 32nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2009, Boston, MA, USA, July 19-23, 2009*, pages 275–282, 2009. ISBN 978-1-60558-483-6.

Zaiqing Nie, Yuanzhi Zhang, Ji-Rong Wen, and Wei-Ying Ma. Object-level Ranking: Bringing Order to Web Objects. In *Proceedings of the 14th International Conference on World Wide Web*, WWW '05, pages 567–574, 2005. ISBN 1-59593-046-9.

Paul Ogilvie and James P. Callan. Experiments Using the Lemur Toolkit. In *Proceedings of The 10th Text REtrieval Conference*, 2001.

Emilio Soria Olivas, Jos David Martn Guerrero, Marcelino Martinez-Sober, Jose Rafael Magdalena-Benedito, and Antonio Jos Serrano Lpez. *Handbook of Research on Machine Learning Applications and Trends: Algorithms, Methods, and Techniques*. IGI Global, 2009.

Iadh Ounis, Gianni Amati, Vassilis Plachouras, Ben He, Craig Macdonald, and Douglas Johnson. Terrier Information Retrieval Platform. In *ECIR*, pages 517–519, 2005.

Desislava Petkova and W. Bruce Croft. Proximity-based Document Representation for Named Entity Retrieval. In *Proceedings of the sixteenth ACM Conference on Conference on Information and knowledge management*, CIKM '07, pages 731–740, 2007. ISBN 978-1-59593-803-9.

Jay M. Ponte and W. Bruce Croft. A Language Modeling Approach to Information Retrieval. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '98, pages 275–281, 1998.

Yves Rasolofo and Jacques Savoy. Term Proximity Scoring for Keyword-Based Retrieval Systems. In *ECIR*, pages 207–218, 2003.

R. Reyment and G. Joreskog. *Applied Factor Analysis in the Natural Sciences (2rd Edition)*. Cambridge University Press, 1996.

Ricardo Ribeiro and David Martins de Matos. Mixed-source Multi-document Speech-to-text Summarization. In *Proceedings of the Workshop on Multi-source Multilingual Information Extraction and Summarization*, MMIES '08, pages 33–40, 2008. ISBN 978-1-905593-51-4.

Stephen E. Robertson. The Probabilistic Character of Relevance. *Inf. Process. Manage.*, 13(4):247–251, 1977.

Stephen E. Robertson and Karen Spark Jones. Relevance Weighting of Search Terms. *Journal of the American Society for Information Science*, 27:129–146, 1976.

Stephen E. Robertson and Steve Walker. Some Simple Effective Approximations to the 2-Poisson Model for Probabilistic Weighted Retrieval. In *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 3-6 July 1994, Dublin, Ireland*, pages 232–241, 1994.

Stephen E. Robertson, M. E. Maron, and William S. Cooper. The Unified Probabilistic Model for IR. In *Proceedings of the 5th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '82, pages 108–117, 1982.

Henning Rode, Pavel Serdyukov, and Djoerd Hiemstra. Combining Document- and Paragraph-based Entity Ranking. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '08, 2008. ISBN 978-1-60558-164-4.

Miguel E. Ruiz, Ying Sun, Jianqiang Wang, and Hongfang Liu. Exploring Traits of Adjectives to Predict Polarity Opinion in Blogs and Semantic Filters in Genomics. In *Proceedings of 16th Text REtrieval Conference*, 2007.

G. Salton. *Automatic Information Organization and Retrieval.* 1968.

Gerard Salton and Michael E. Lesk. The SMART Automatic Document Retrieval Systems - an Illustration. *Commun. ACM*, 8(6):391–398, 1965.

Gerard Salton, Edward A. Fox, and Harry Wu. Extended Boolean Information Retrieval. *Commun. ACM*, 26(11):1022–1036, 1983. ISSN 0001-0782.

Rodrygo L. T. Santos, Richard M. C. McCreadie, Craig Macdonald, and Iadh Ounis. University of Glasgow at TREC 2010: Experiments with Terrier in Blog and Web Tracks. In *Proceedings of The 19th Text REtrieval Conference*, 2010.

Luis Sarmento, Valentin Jijkuon, Maarten de Rijke, and Eugenio Oliveira. "More Like These": Growing Entity Classes from Seeds. In *Proceedings of the sixteenth ACM Conference on Conference on Information and knowledge management*, CIKM '07, pages 959–962, 2007. ISBN 978-1-59593-803-9.

Prithviraj Sen. Collective context-aware Topic Models for Entity disambiguation. In *Proceedings of the 21st International Conference on World Wide Web*, WWW '12, pages 729–738, 2012. ISBN 978-1-4503-1229-5.

Lixin Shi and Jian-Yun Nie. Using Various Term Dependencies According to Their Utilities. In *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*, CIKM '10, pages 1493–1496, 2010. ISBN 978-1-4503-0099-5.

Fei Song and W. Bruce Croft. A General Language Model for Information Retrieval. In *CIKM*, pages 316–321, 1999.

Ruihua Song, Michael J. Taylor, Ji-Rong Wen, Hsiao-Wuen Hon, and Yong Yu. Viewing Term Proximity from a Different Perspective. In *ECIR*, pages 346–357, 2008.

Ramakrishnan Srikant and Rakesh Agrawal. Mining Sequential Patterns: Generalizations and Performance Improvements. In *Proceedings of the 5th International Conference on Extending Database Technology: Advances in Database Technology*, EDBT '96, pages 3–17, 1996. ISBN 3-540-61057-X.

Nicola Stokes, Yi Li, Lawrence Cavedon, Eric Huang, Jiawen Rong, and Justin Zobel. Entity-Based Relevance Feedback for Genomic List Answer Retrieval. In *Proceedings of 16th Text REtrieval Conference*, 2007.

C. Subbaraoand, N.V. Subbarao, and S.N. Chandu. Characterisation of Groundwater Contamination Using Factor Analysis. *Environmental Geology*, 28:175–180, 1995.

Tao Tao and ChengXiang Zhai. An Exploration of Proximity Measures in Information Retrieval. In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '07, pages 295–302, 2007. ISBN 978-1-59593-597-7.

Luis Tari, Phan Huy Tu, Barry Lumpkin, Robert Leaman, Graciela Gonzalez, and Chitta Baral. Passage Relevancy Through Semantic Relatedness. In *Proceedings of 16th Text REtrieval Conference*, 2007.

Stefanie Tellex, Boris Katz, Jimmy Lin, Aaron Fernandes, and Gregory Marton. Quantitative Evaluation of Passage Retrieval Algorithms for Question Answering. In *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in informaion Retrieval*, SIGIR '03, pages 41–47, 2003. ISBN 1-58113-646-3.

John A. Tomlin. A New Paradigm for Ranking Pages on the World Wide Web. In *Proceedings of the 12th International Conference on World Wide Web*, WWW '03, pages 350–355, 2003. ISBN 1-58113-680-3.

H. Turtle and W. Bruce Croft. Inference Networks for Document Retrieval. In *Proceedings of the 13th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '90, pages 1–24, 1990. ISBN 0-89791-408-2.

C. J. van Rijsbergen. A New Theoretical Framework for Information Retrieval. In *Proceedings of the 9th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '86, pages 194–200, 1986.

C. J. van Rijsbergen. Probabilistic Retrieval Revisited. *Comput. J.*, 35(3):291–298, 1992.

C. J. van Rijsbergen and Karen Spark Jones. A Test for the Separation of Relevant and Nonrelevant Documents in Experimental Retrieval Collections. *Journal of Documentation*, 29:251–257, 1973.

Raymond Wan, Vo Ngoc Anh, and Hiroshi Mamitsuka. Passage Retrieval with Vector Space and Query-Level Aspect Models. In *Proceedings of 16th Text REtrieval Conference*, 2007.

Steven Wartik. Information Retrieval. chapter Boolean Operations, pages 264–292. 1992. ISBN 0-13-463837-9.

Xing Wei and W. Bruce Croft. 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.*

J. Yamron. Topic Detection and Tracking Segmentation Task. In *Proceedings of The Topic Detection and Tracking Workshop*, 1997.

Zhihao Yang, Hongfei Lin, Baojin Cui, Yanpeng Li, and Xiao Zhang. DUTIR at TREC 2007 Genomics Track. In *Proceedings of 16th Text REtrieval Conference*, 2007.

Xiaoshi Yin, Xiangji Huang, and Zhoujun Li. Promoting Ranking Diversity for Biomedical Information Retrieval Using Wikipedia. In *ECIR*, pages 495–507, 2010.

Lei Yuan, Yalin Wang, Paul M. Thompson, Vaibhav A. Narayan, and Jieping Ye. Multi-source Learning for Joint Analysis of Incomplete Multi-modality Neuroimaging Data. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge discovery and Data mining*, KDD '12, pages 1149–1157, 2012. ISBN 978-1-4503-1462-6.

Hugo Zaragoza, Henning Rode, Peter Mika, Jordi Atserias, Massimiliano Ciaramita, and Giuseppe Attardi. Ranking Very Many Typed Entities on Wikipedia. In *Proceedings of the sixteenth ACM Conference on Conference on Information and knowledge management*, CIKM '07, pages 1015–1018, 2007. ISBN 978-1-59593-803-9.

ChengXiang Zhai and John D. Lafferty. A Study of Smoothing Methods for Language Models Applied to Ad Hoc Information Retrieval. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, September 9-13, 2001, New Orleans, Louisiana, USA*, pages 334–342, 2001.

R. Zhang, C. Yi, Z. Zheng, D. Metzler, and J. Nie. Search Result Re-Ranking by Feedback Control Adjustment for Time-Sensitive Query. In *Proceedings of North American Chapter of the Association for Computational Linguistics - Human Language Technologies (NAACL HLT)*, 2009.

Jiashu Zhao, Jimmy X. Huang, and Ben He. CRTER: Using Cross Terms to Enhance Probabilistic Information Retrieval. In *Proceeding of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2011, Beijing, China, July 25-29, 2011*, pages 155–164, 2011.

M. Zhong and X. Huang. Concept-based Biomedical Text Retrieval. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, August 6-11, 2006, Seattle, Washington, USA*, pages 723–724, 2006. ISBN 1-59593-369-7.

Wei Zhou and Clement T. Yu. TREC Genomics Track at UIC. In *Proceedings of 16th Text REtrieval Conference*, 2007.

Xiaofeng Zhou, Jimmy Xiangji Huang, and Ben He. Enhancing Ad-hoc Relevance Weighting Using Probability Density Estimation. In *Proceeding of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2011, Beijing, China, July 25-29, 2011*, pages 175–184, 2011.

Jun Zhu, Zaiqing Nie, Xiaojiang Liu, Bo Zhang, and Ji-Rong Wen. StatSnowball: a Statistical Approach to Extracting Entity Relationships. In *Proceedings of the 18th*

International Conference on World wide web, WWW '09, pages 101–110, 2009. ISBN 978-1-60558-487-4.

A Scripts for Duplicating the Okapi Experiments

Our experiments were conducted on a double-processor server which has 2 Intel Xeon 2.40GHz CPU and 2G memory. The version of Linux kernel we used is version 2.4.26, the Okapi system is v2.4.1.

As we describe in Chapter 6, the whole process is consistent of: environment preparation, data preprocessing, generating exchange file, building index, query processing, result searching, re-ranking. In this chapter, we present the scripts for the 2007-2004 data sets respectively.

A.1 Environment Preparation

There is a system file named environmentSettings.cshrc which contains the full paths and must be customized for the individual's directory correctly. The commands and the environmentSettings.cshrc file are presented below.

```
[qinmin@teln711-027 javokapi]$ tcsh
[qinmin@teln711-027 javokapi]$source environmentSettings.cshrc
```

```
! /bin/csh
unsetenv TMPDIR
unsetenv OKAPI_ROOT
unsetenv OKAPI_LIBDIR
unsetenv OKAPI_BINDIR
unsetenv GUI_CONFIG_FILES
unsetenv OKAPI_LOGS_DIR
unsetenv INDEXER_LOGS
unsetenv BSS_TEMPPATH
unsetenv BSS_PARMPATH
unsetenv BSS_PASSAGE_AVEDOCLEN

## TMPDIR: Possibly still used by parts of the system.
setenv TMPDIR /tmp

## OKAPI_ROOT: The full pathname of the package installation
directory setenv OKAPI_ROOT /home/qinmin/javokapi

## OKAPI_LIBDIR: The pathname of the BSS library, libi0+.a
setenv OKAPI_LIBDIR $OKAPI_ROOT/lib

## OKAPI_BINDIR: Okapi binaries. You might like to add this to your
search PATH setenv OKAPI_BINDIR $OKAPI_ROOT/bin

## BSS_TEMPPATH: Directory for temporary files produced by all parts
of the system. setenv BSS_TEMPPATH /tmp

## BSS_PARMPATH: The full pathname of the database parameter files.
setenv BSS_PARMPATH $OKAPI_ROOT/databases
```

```
## GUI_CONFIG_FILES: The full pathname of the location of the GUI
configuration files. setenv GUI_CONFIG_FILES $OKAPI_ROOT

## INDEXER_LOGS: The full pathname of the indexer log files. setenv
INDEXER_LOGS $OKAPI_ROOT/IndexerLogs

## OKAPI_LOGS_DIR: The full pathname of the interface log files.
setenv OKAPI_LOGS_DIR $OKAPI_ROOT/OkapiLogs

## BASIC SEARCH SYSTEM PASSAGE AVERAGE DOCUMENT LENGTH setenv
BSS_PASSAGE_AVEDOCLEN 2500

## OKAPI_SOURCE: The full pathname of the Okapi 2.41 source files.
setenv OKAPI_SOURCE $OKAPI_ROOT/source

## OKAPI_SOURCE: The full pathname of the Okapi parse files. setenv
OKAPI_PARSE $OKAPI_ROOT/parse


setenv MAIN_DEBUG 0
setenv MAKE_RS_SETS_DEBUG 0
setenv READ_PARAMETER_FILES_DEBUG 0
setenv CHECK_FOR_PARAGRAPH_FILE_DEBUG 0
setenv SET_ENV_DEBUG 0
setenv READ_ENV_DEBUG 0

## Set BSS variables

setenv ADD_TO_SEEN_SET_DEBUG 0
setenv ADD_TO_BIGR_SET_DEBUG 0
```

```
setenv ADJUST_RSV_FACTOR_DEBUG 0
setenv BIGRLOAD 0
setenv BM_TARGET 0
setenv BOTH_PHRASE_OPS 0
setenv BSS_SEARCH_DEBUG 0
setenv BUILD_HITLIST_DEBUG 0
setenv CALC_RSV_DEBUG 0
setenv CALC_WGT_DEBUG 0
setenv CHARS_PER_PAGE 0
setenv CHECK_FOR_PARAGRAPH_FILE_DEBUG 0
setenv CHECK_USER_RELS_DEBUG 0
setenv CLEAR_RF_DEBUG 0
setenv CONSTRUCT_DOCLENGTH_FIELD_DEBUG 0
setenv CONSTRUCT_TITLE_DEBUG 0
setenv DB_SEARCH_DEBUG 0
setenv DETERMINE_DOC_LENGTH_DEBUG 0
setenv DIRECTORY_REC_LEN 0
setenv DISPLAY_DEBUG 0
setenv DOC_THRESHOLD 0
setenv EXTRACT_TERMS_DEBUG 0
setenv FIND_DOCSET_DEBUG 0
setenv HEADER_SHOW_FORMAT 0
setenv HIGHLIGHT_REC_LEN 0
setenv HYPHEN_POS 0
setenv MAIN_DEBUG 0
setenv MAKERJ_DEBUG 0
setenv MAKE_REL_DEBUG 0
setenv MAX_RECS_TO_SHOW 0
setenv MAX_RELS 0
setenv MAX_TERMSET_SIZE 0
```

```
setenv MAX_TERMS_PER_DOC 0
setenv MAX_TITLE_CHARS 0
setenv OFFSET_X 0
setenv OFFSET_Y 0
setenv PARSE_HEADER_DEBUG 0
setenv PARSE_SHOW_FILE_DEBUG 0
setenv PASSAGE_REC_LEN 0
setenv PASSAGE_SHOW_FORMAT 0
setenv P_STEP 0
setenv P_UNIT 0
setenv QUERY_WINDOW_WIDTH 0
setenv READ_ENV_DEBUG 0
setenv READ_PARAMETER_FILES_DEBUG 0
setenv REMOVE_FROM_BIGR_SET_DEBUG 0
setenv RLOAD 0
setenv SET_ENV_DEBUG 0
setenv SET_LR_THRESHOLD_DEBUG 0
setenv SET_RSV_FACTOR_DEBUG 0
setenv SHOW_DEBUG 0
setenv SPLIT_UP_DEBUG 0
setenv TERM_INPUT_DEBUG 1
setenv TERM_OCCURRENCE_DEBUG 1
setenv UPDATE_USER_RELS_DEBUG 0
setenv WEIGHT_FUNCTION 0
setenv WRITE_NEW_FILE_DEBUG 0
setenv WRITE_USER_TERMS_DEBUG 0
setenv TERM_ENTRY_DEBUG 1

##  Necessary for the okapi interface application.
```

```
limit stacksize unlimited
```

A.2 Data Preprocessing

Raw data, which are presented in Appendix C, are HTML/XML files. Then we parse these files into text files by getting rid of the HTML/XML tags like $\langle p \rangle$ and $\langle /p \rangle$. There are many kinds of tools for this part.

For the three passage extraction algorithms proposed in this thesis, we process the raw data into three formats according to the passage extraction algorithms. The paragraph-Parsed, sentenceParsed and wordSentenceParsed algorithms are applied on the data set and ready to be generated as the exchange files. Furthermore, the porter stemming and a stop-word set are utilized.

A.3 Generating Exchange Files

After parsing the raw data, we generate the pure text files into the exchange files which have the format for building the index. Then three kinds of exchange files are generated for the purpose of building three indices.

In the exchange files, the fields include: Document ID, Title, Abstract, Document Body, Chemical List and MeSH Terms. Document ID is the identification for an individual document. Document body is the main content of a document. As the definitions

in Chapter 3, three passage extraction algorithms are mainly applied on the document bodies. Chemical List describes what the related chemical substance is involved. The Medical Subject Headings (MeSH) terms comprise NLM's controlled vocabulary used for indexing articles, for cataloging books and other holdings, and for searching MeSH-indexed databases.

A.4 Building Index

For the Okapi system, we use commands to build index. Taking the wordSentenceParsed index as an example, we give the scripts for how to build the index for the 2007 Genomics data set as follows. Note that the files named wordSentenceParsed-1.exch is one of the exchange files generated according to the wordSentenceParsed algorithm.

```
Editing the following three files under the folder
"/home/qinmin/javokapi/databases"
2007gendoc_word
2007gendoc_word.search_groups
db_avail

Run the commands under the corresponding directory
[qinmin@teln711-027 javokapi]$ cd bin
[qinmin@teln711-027 bin]$
../bin/convert_runtime -c $BSS_PARMPATH gen07  <
/TB/genomics-data/2006/genomic06-data/docExch

../bin/ix1 -m 500 -c $BSS_PARMPATH -doclens -deltmp -delfinal gen07
```

```
0 | ../bin/ixf -c $BSS_PARMPATH gen07  0


../bin/ix1 -m 500 -c $BSS_PARMPATH -doclens -deltmp -delfinal gen07
1 | ../bin/ixf -c $BSS_PARMPATH gen07  1
```

Therefore, we present the scripts for "2007gendoc_word", "2007gendoc_word.search_groups" and "db_avail" respectively.

2007gendoc_word

```
name=2007gendoc_word
lastbibvol=5
bib_basename=trec.2007gendoc_word0.bib
bib_basename=trec.2007gendoc_word1.bib
bib_basename=trec.2007gendoc_word2.bib
bib_basename=trec.2007gendoc_word3.bib
bib_basename=trec.2007gendoc_word4.bib
bib_basename=trec.2007gendoc_word5.bib
bib_dir=/TB/genomics-data/2006/genomic06-data/bibfiles/
bib_dir=/TB/genomics-data/2006/genomic06-data/bibfiles/
bib_dir=/TB/genomics-data/2006/genomic06-data/bibfiles/
bib_dir=/TB/genomics-data/2006/genomic06-data/bibfiles/
bib_dir=/TB/genomics-data/2006/genomic06-data/bibfiles/
bib_dir=/TB/genomics-data/2006/genomic06-data/bibfiles/
bibsize=2048
bibsize=2048
bibsize=2048
bibsize=2048
bibsize=2048
bibsize=2048
```

```
real_bibsize=536870984

real_bibsize=536871015

real_bibsize=536870924

real_bibsize=536870964

real_bibsize=536871013

real_bibsize=383475394

display_name=trec.2007gendoc_word

nr=34096344

nf=2

f_abbrev=DN

f_abbrev=TX

rec_mult=4

db_type=ai

has_lims=0

maxreclen=18660

ni=2

last_ixvol=0

ix_stem=/TB/genomics-data/2006/genomic06-data/index//2007gendoc_word

ix_volsize=3072

ix_type=13

last_ixvol=0

ix_stem=/TB/genomics-data/2006/genomic06-data/index//2007gendoc_word

ix_volsize=15360

ix_type=13

no_drl=0
```

2007gendoc_word.search_groups :

```
dn 1 0 literal nostem gsl.lemur1 1 0 -1

tx 1 1 words3 sstem gsl.lemur1 2 0 -1
```

db_avail :

```
## <DB_ROOT>/databases/db_avail
##
##  Where <DB_ROOT> is the pathname to the "databases" directory
##  where all parameter files are stored.
##
##  Each line contains:
##
##    <database_name> <user_list>
##
##  <user_list> defines allowed / disallowed users for <database_name>
##  and is made up of combinations of all or some of:
##
##              *      available to anyone
##           gid       available to group
##          -gid       not available to group
##  <login name>       available to user with given login name
##
##  NOTE: gid must be numeric
##
##  The <user_list> is read until a parameter that applies to
##  the current user is found or to the end of the list if none
##  is found. Individual users are automatically excluded if their
##  <login_name> is not included in <user_list> and <user_list>
##  does not contain a <gid> that the user is in, or a *
##
##  Example entries:
##
##     inspec -34 *              database "inspec" is available to
##                               everyone except those in group 34
```

```
##
##     d25_96 okapi sw mg ntd   database "d25_96" is available only to
##                              users with login names  sw, mg, ntd
##
##     bible *                  database "bible" is available to
##                              everyone.

2004gendoc *
2007gendoc *
2007gendoc_word *
2007gendoc_sentence *
2007gendoc_sentenceN *
```

A.5 Query Processing

In this thesis, we use a query expansion program which is described in detail by Zhong and Huang Huang et al. (2005b), Zhong and Huang (2006). The command is:

```
[qinmin@teln711-027 bin]$ javac expandedTermIndexFinder.java
[qinmin@teln711-027 bin]$ java -Djava.library.path=.
expandedTermIndexFinder 2007topics.txt 2007gendoc_word tx
```

The query expansion program is presented as follows.

```
import java.io.*;
import java.util.*;

public class expandedTermIndexFinder
{
        public static relex okapi_interface;
```

```
        public static void main(String[] args) throws Exception
        {
                String inFilename = args[0];
                String outFilename = "expansion-index-of-" + inFilename;

                okapi_interface = new relex();
        okapi_interface.javainit();
        okapi_interface.comm("ch " + args[1]);
        okapi_interface.comm("set a=" + args[2]);

                BufferedReader reader = new BufferedReader(new FileReader(inFile
name));
                BufferedWriter writer = new BufferedWriter(new FileWriter(new Fi
le(outFilename)));
                String line = "";
                String counter = "";
                while ((line = reader.readLine()) != null)
                {
                        line = line.replaceFirst("<", "");
                        counter = line.substring(0, line.indexOf('>'));
                        String output = "Topic #" + counter + " : ";
                        line = line.substring(line.indexOf('>') + 1, line.length());
                        String okapiOutput = okapi_interface.comm("p t=" + line);
                        StringTokenizer tokens = new StringTokenizer(okapiOutput);
                        String[] terms = new String[Integer.parseInt(tokens.nextToken())];
                        for (int i = 0; i < terms.length; i++)
                        {
                                String s = tokens.nextToken();
                                terms[i] = s.substring(s.indexOf('=') + 1);
                        }
```

```
                output = output + countNonDuplicateTerms(terms);
                writer.write(output);
            writer.newLine();
            writer.flush();
            }
            reader.close();
            writer.close();
        }

        private static int countNonDuplicateTerms(String[] x)
        {
            int count = 0;
            for (int i = 0; i < x.length; i++)
            {
                boolean duplicateFound = false;
                String str = x[i];
                for (int j = 0; j < i; j++)
                    if (x[j].equals(str))
                    {
                        duplicateFound = true;
                        break;
                    }
                if (!duplicateFound)
                    count++;
            }
            return count;
        }
    }
```

A.6 Parameter Settings

The parameters (k_1, b) in the BM25 weighting function are conducted by different values. We change the parameter settings in the *squery.java* program.

```
String  result = x.comm("f " + sbf.toString() +
    " k1=0.5 bm25_b=1.3 op=bm25 ");
System.out.println("SQUERYSQUERYSQUERY \"qoutput\"
    sbf.toString():     " + sbf.toString());
System.out.println("RECENT>>> " + result);
String[] parsed = result.split(" maxwt=");
String[] parsed2 = parsed[0].split(" np="); //number of papers
int totaldoc = new Integer(parsed2[1]).intValue();
StringBuffer final_result_doc = new StringBuffer();
```

A.7 Result Searching

The results related to the topics are searched by the Okapi system based on the BM25 weighting function. A shell script named run-okapi-2007.sh is provided to simply run the command. Note that in the command below, The first top 10 documents from initial retrieval are selected as the pseudo relevance feedback.

```
[qinmin@teln711-027 bin]$ ./run-okapi-2007.sh 2007gendoc_word tx dn
2006topics.txt 0.25 1000 10 1 expansion-index-of-2006topics.txt
yorku07gal &
```

The script for the run-okapi-2007.sh shell is:

```
#!/bin/tcsh
# runs the ssearch program and creates varius outputs
# echo USAGE: ./run-okapi-2007.sh followed by arguments as follows
#1 database-name        : the name of database okapi opens
#2 passage-index-name   : the index used for retrieval of terms
#3 passageID-index-name : the index used for retrieval of passages
                           given passage id
#4 topics-file          : file that contains the topics
#5 feedback-weight      : weight given to feedback terms
#6 number-of-outputs    : number of passages retrieved for each topic
#7 #-of-top-passages    : number of top passages returned in a file
                           for manual view
#8 expanded weight      : this factor will be multiplied by the weight
                           of each expanded terms
#9 expanded file         : this file contains the starting index of
                           expanded terms
# run-name              : run tag of each output for trec purposes

java -Djava.library.path=. ssearch $1 $2 $3 $4 $5 $6 $7 $8 $9 >
ssearch-output
cd ./$1-$2-$3-$4-$5-$6-$7-$8/
cat doc* > combined-doc
cat txt* > combined-txt
rm doc*
rm txt*
cd ..
javac TopicTermCounter.java

java -Djava.library.path=. TopicTermCounter $1 $2 $3 $4
./$1-$2-$3-$4-$5-$6-$7-$8/combined-doc $6 >
```

```
./$1-$2-$3-$4-$5-$6-$7-$8/term-count-of-docs

java -Djava.library.path=. TopicTermCounter $1 $2 $3 $4
./$1-$2-$3-$4-$5-$6-$7-$8/combined-txt $6 >
./$1-$2-$3-$4-$5-$6-$7-$8/term-count-of-txts

javac FeedbackTermCounter.java java -Djava.library.path=.
FeedbackTermCounter $1 $2 $3
./$1-$2-$3-$4-$5-$6-$7-$8/feedback-terms
./$1-$2-$3-$4-$5-$6-$7-$8/combined-txt $6 >
./$1-$2-$3-$4-$5-$6-$7-$8/term-count-of-feedbacks
cd ./$1-$2-$3-$4-$5-$6-$7-$8/
cp ../OutputFormatter.java .
javac OutputFormatter.java
java OutputFormatter combined-doc $10
java OutputFormatter combined-txt $10
echo doc term statistics > statistics-of-term-counts
cat term-count-of-docs | grep '##' >> statistics-of-term-counts
echo >> statistics-of-term-counts
echo txt term statistics >> statistics-of-term-counts
cat term-count-of-txts | grep '##' >> statistics-of-term-counts
echo >> statistics-of-term-counts
echo feedback term statistics >> statistics-of-term-counts
cat term-count-of-feedbacks | grep '##' >> statistics-of-term-counts
echo >> statistics-of-term-counts
rm *.java
rm *.class
#rm combined*
mv ../ssearch-output .
cd ..
```

```
javac retrievePassages.java
java -Djava.library.path=. retrievePassages $1 $3
./$1-$2-$3-$4-$5-$6-$7-$8/top-passageIDs

mv top-passages ./$1-$2-$3-$4-$5-$6-$7-$8/
cd ./$1-$2-$3-$4-$5-$6-$7-$8/
chmod 644 *
```

Ssearch.java is the main program for the result searching process. There are two versions for four data sets. We present the programs in detail respectively.

Ssearch.java in 2007 and 2006

```
import java.io.*;
import java.util.LinkedList;

public class ssearch
{
        public static relex okapi_interface;
        // THIS IS THE STARTING INDEX FOR EXPANDED TERMS OF EACH OF THE TOPICS IN ORDER.
        // THIS HELPS DETERMINE WHERE IN THE LL1 LINKED LIST OF SQUERY THE EXPANDED
        // TERMS ARE LOCATED.
        private static int[] expandedTermsIndecies;
        public static void main (String[]args) throws Exception
        {
        okapi_interface = new relex();
        okapi_interface.javainit();
        Infokit ink = new Infokit();

                // TO FURTHER ENHANCE THE WORKING OF PROGRAM. MANY OF THE HARDCODED
```

```
        // INFORMATIONS WILL BECOME COMMAND LINE ARGUMENTS SO ELIMINATE THE NEED
        // TO RECOMPILE THE PROGRAM.

        if (args.length != 9)
        {
                System.out.println("wrong number of arguments passed.");
                System.out.println("java ssearch <database name> <main index>
                <passage id index> <topics file> <feedback weight> <# output passages>
                <# top passages> <expanded terms weight> <expanded terms file>");
                System.exit(0);
        }
        String databaseName = args[0];
        String passageIndex = args[1];
        String passageIDIndex = args[2];
        String topicsFile = args[3];
float feedbackWeight = new Float(args[4]).floatValue();
int outputNumber = Integer.parseInt(args[5]);
        int numberOfTopPassages = Integer.parseInt(args[6]);

        float expandedTermsWeight = Float.parseFloat(args[7]);
        String expandedTermsIndexFile = args[8];

        // SOME BASIC ERROR CHECKING IN CASE ANY SIMPLE MISTAKE IS MADE
        if (feedbackWeight < 0)
        {
                System.out.println("feedbackWeight cannot be zero.");
                System.exit(0);
        }
        if (outputNumber < 0)
        {
```

```
        System.out.println("outputNumber cannot be zero.");
        System.exit(0);
}
if (numberOfTopPassages < 0)
{
        System.out.println("numberOfTopPassages cannot be zero.");
        System.exit(0);
}
if (outputNumber < numberOfTopPassages)
{
        System.out.println("outputNumber cannot be less than
        numberOfTopPassages.
        If you want only " + outputNumber);
        System.out.println("to be retreived, then there are " +
        outputNumber + "top passages. Cannot ask for");
        System.out.println(numberOfTopPassages + "top passages.");
        System.exit(0);
}

// CHOOSES THE DATABASE FILE AND THE SPECIFIED INDEX THAT MUST BE USED
// FOR PARSING AND FINDINF OF RESULTS. IF WRONG DATABASE OR INDEX NAME
// IS PASSED, AT THE MOMENT, WE DO NOT HAVE A WAY OF CATCHING THE ERROR.
okapi_interface.comm ("ch " + databaseName);
okapi_interface.comm ("set attribute=" + passageIndex);

// THE PROGRAM CREATES MANY FILES FOR EACH OF THE TOPICS IT PROCESSES ALONG
// WITH MANY OTHER FILES SUCH AS TOP PASSAGES. WE CREATE A DIRECTORY FOR
// THIS PURPOSE SO ALL THE GENERATED FILES ARE PLACED INSIDE IT.
String outputDirectory = args[0] + "-" + args[1] + "-" + args[2] + "-"
+ args[3] + "-" + args[4] + "-" + args [5] + "-" + args[6] + "-" + args[7];
```

```
        outputDirectory = outputDirectory + "/";
        boolean success = (new File(outputDirectory)).mkdir();
if (!success)
        {
        System.out.println("Program cannot create a directory to output results.");
        System.out.println("Please make sure there is the program has
        permission to do so.");
        System.exit(0);
}
        // VARIOUS FILE WRITERS TO SAVE THE OUTPUT OF FEEDBACK AND TOP PASSAGES.
        BufferedWriter feedbackWriter = new BufferedWriter(new FileWriter
        (outputDirectory + "feedback-terms"));
        BufferedWriter topPassageWriter = new BufferedWriter(new FileWriter
        (outputDirectory + "top-passageIDs"));

        // WE READ THE INFORMATION ABOUT THE STARTING INDEX FILES FROM A FILE
        BufferedReader expandedIndexFileReader = new BufferedReader (new FileReader
        (expandedTermsIndexFile));
        String line = null;

        LinkedList<Integer> temp = new LinkedList<Integer>();
        while((line = expandedIndexFileReader.readLine()) != null)
        {
                temp.add(Integer.parseInt(line.substring(line.indexOf(':') + 1).trim()));
        }
        expandedTermsIndecies = new int[temp.size()];
        for (int i = 0; i < expandedTermsIndecies.length; i++)
                expandedTermsIndecies[i] = temp.get(i);

        BufferedReader input = new BufferedReader (new FileReader (topicsFile));
```

```
BufferedWriter bwr = new BufferedWriter (new FileWriter (outputDirectory +
"structured-query-of-" + topicsFile));

// POINTS TO THE CORRECT INDEX FOR EACH TOPIC
int expandedTermsIndex = 0;

while ((line = input.readLine ()) != null)
{
if (line.trim ().equals (""))
        {
                continue;
        }
String[]parts = line.split (">");
String num = parts[0].replaceAll ("<", "");
//remove 's and i.e and i.e.
String contents = parts[1].replaceAll ("'s", "").replaceAll ("i\\.e", " ")
.replaceAll ("i\\.e\\.", " ");
bwr.write ("<ID>" + new Integer (num) + "</ID>\n");
bwr.flush ();

        /*
 * using the orignal query to build a linkedlist of snip_seq
 * using the snip_seq,a linkedlist of snoppet which built before, to build
 * a bssset linkedlist
 */
squery squ = new squery (contents, okapi_interface, ink, passageIndex,
passageIDIndex, feedbackWeight, outputNumber, numberOfTopPassages,
topPassageWriter, feedbackWriter, expandedTermsIndecies
[expandedTermsIndex], expandedTermsWeight);
```

```
                System.out.print ("" + new Integer (num) + "::");
                squ.fb_qoutput (new Integer (num).intValue (), outputDirectory + "doc" +
                num, outputDirectory + "txt" + num);

                System.out.println ();
                squ.outputq (new Integer (num).intValue (), bwr, null);
                okapi_interface.comm ("delete all");
                expandedTermsIndex++;
                }
                bwr.close ();
                topPassageWriter.close();
                input.close();
                feedbackWriter.close();
                expandedIndexFileReader.close();
        }
}
```

Ssearch.java in 2005 and 2004

```
import java.io.*;
import java.util.*;
import java.text.*;

public class ssearch {
    public static relex xy;
    public static void main(String[] args) throws Exception {
        xy = new relex();
        xy.x = new relex();
        xy.x.javainit();
        xy.x.comm("ch 2004gendoc");
        Infokit ink = new Infokit();
```

```
        float adjusted = new Float(args[1]).floatValue();
/*
        if (args[0].indexOf("04") >= 0) {
            adjusted = 0.22f;
            //adjusted = 0.17f;
        }
  */
        File testFile = new File(args[0]);
        BufferedReader input = null;

        //try {
            input = new BufferedReader(new FileReader(testFile));
            String line = null;
            BufferedWriter bwr = new BufferedWriter(new
                    FileWriter("result/q" + args[0]));

            while ((line = input.readLine()) != null) {
                if (line.trim().equals("")) {
                    continue;
                }
                String[] parts = line.split(">");
                String num = parts[0].replaceAll("<", "");
                String contents = parts[1].replaceAll("'s", "").replaceAll("i\\.e", " ")
                .replaceAll("i\\.e\\.", " ");
                int qrynumber = new Integer(num).intValue();
                // System.out.println(contents);
                 if(qrynumber==128){
                // squery squ = new squery(contents,xy.x,ink,adjusted,true);
                System.out.println(contents);
                squery squ = new squery(contents, xy.x, ink, adjusted);
```

```
                // squ.qoutput(new Integer(num).intValue(),"newdoc1/doc"+num,null);
                bwr.write("<ID>" + new Integer(num) + "</ID>\n");
                System.out.println("XXX:<ID>" + new Integer(num) + "</ID>");
                bwr.flush();

                squ.fb_qoutput(new Integer(num).intValue(),
                        "result/doc" + num,
                        "result/txt" + num);

                squ.outputq(new Integer(num).intValue(), bwr, null);
                 }
                xy.x.comm("delete all");
            }
            bwr.close();

        //} catch (Exception ec) {}

    }
}
```

A.8 Re-Ranking

Result combination and the Bayesian learning approach are applied on the retrieved results for re-ranking. The input files for all the algorithm proposed in Chapter 4 and 5 has been generated. Taking one of the runs as an example, we present the improved combination script as follows. Note that the baseline-0.5-1.3-2007 file is the original re-

sult retrieved by Okapi, and terms-0.5-1.3-2007 is the keyword file such that it indicates which keyword in the topic is found in which passage.

```
input files:
    baseline-0.5-1.3-2007
    terms-0.5-1.3-2007
output file:
    improved-0.5-1.3-2007

In Eclipse, the configuration files are in the folder of
../result-combine/config

The configuration file for matrix generation is:
    terms-input-file = data/terms-0.5-1.3-2007
    baseline-input-file = data/baseline-0.5-1.3-2007
    output-line = 1000
    output-file-name = data/matrix-0.5-1.3-2007

The configuration file for recursive re-ranking is:
    matrix-input-file = data/matrix-0.5-1.3-2007
    baseline-input-file = data/baseline-0.5-1.3-2007
    terms-fa-weight-name = data/fa-weights-0.5-1.3-2007
    output-line = 1000
    output-file-name = data/fa-new-0.5-1.3-2007

The configuration file for the improved combination method is:
    improved-combine-input-file = data/fa-weights-0.5-1.3-2007,
        data/baseline-0.4-2.0-2007, data/baseline-0.5-1.3-2007,
        data/baseline-1.0-1.0-2007, data/baseline-1.2-0.75-2007
    data/baseline-.0-0.4-2007
```

```
output-line = 1000

output-file-name = data/improved-0.5-1.3-2007
```

A.9 Scripts for the TREC 2007 and 2006 Genomics Track

Taking the wordSentenceParsed index as an example, we present the script for the TREC 2007 Genomics Track as follows, after we have done to preprocess the data sets and build up the index.

```
[qinmin@teln711-027 bin]$ tcsh

[qinmin@teln711-027 bin]$ source env.cshrc

[qinmin@teln711-027 bin]$ javac expandedTermIndexFinder.java

[qinmin@teln711-027 bin]$ java -Djava.library.path=. expandedTermIndexFinder

2007topics.txt 2007gendoc_word tx

[qinmin@teln711-027 bin]$ ./run-okapi-2007.sh 2007gendoc_word tx dn

2007topics.txt 0.25 1000 1000 1 expansion-index-of-2007topics.txt

yorku07gal &

[qinmin@teln711-027 2007gendoc_word-tx-dn-2007topics.txt-0.25-1000-1000-1]$ ll

total 22544

-rw-r--r--    1 qinmin   qinmin    1358201 Sep 19 20:10 combined-doc

-rw-r--r--    1 qinmin   qinmin    1356532 Sep 19 20:10 combined-txt

-rw-r--r--    1 qinmin   qinmin       8934 Sep 19 20:07 feedback-terms

-rw-r--r--    1 qinmin   qinmin    1213820 Sep 19 20:21 formatted-combined-doc

-rw-r--r--    1 qinmin   qinmin    1213325 Sep 19 20:21 formatted-combined-txt

-rw-r--r--    1 qinmin   qinmin    3476712 Sep 19 20:10 ssearch-output

-rw-r--r--    1 qinmin   qinmin       5094 Sep 19 20:21 statistics-of-term-counts

-rw-r--r--    1 qinmin   qinmin       2633 Sep 19 20:10 structured-query-of-2007topics.txt

-rw-r--r--    1 qinmin   qinmin    1779813 Sep 19 20:16 term-count-of-docs
```

```
-rw-r--r--    1 qinmin   qinmin    2359028 Sep 19 20:21 term-count-of-feedbacks
-rw-r--r--    1 qinmin   qinmin    1763485 Sep 19 20:19 term-count-of-txts
-rw-r--r--    1 qinmin   qinmin    1382840 Sep 19 20:10 top-passageIDs
-rw-r--r--    1 qinmin   qinmin    7088825 Sep 19 20:21 top-passages
```

As the similar way, we present the script for the TREC 2006 Genomics Track below.

```
[qinmin@teln711-027 bin]$ tcsh
[qinmin@teln711-027 bin]$ source env.cshrc
[qinmin@teln711-027 bin]$ javac expandedTermIndexFinder.java
[qinmin@teln711-027 bin]$ java -Djava.library.path=. expandedTermIndexFinder
2006topics.txt 2007gendoc_word tx
[qinmin@teln711-027 bin]$ ./run-okapi-2007.sh 2007gendoc_word tx dn 2006topics.txt
0.25 1000 1000 1 expansion-index-of-2006topics.txt yorku07gal &
[qinmin@teln711-027 2007gendoc_word-tx-dn-2006topics.txt-0.25-1000-1000-1]$ ll
total 22544
-rw-r--r--    1 qinmin   qinmin    1358201 Sep 19 20:10 combined-doc
-rw-r--r--    1 qinmin   qinmin    1356532 Sep 19 20:10 combined-txt
-rw-r--r--    1 qinmin   qinmin       8934 Sep 19 20:07 feedback-terms
-rw-r--r--    1 qinmin   qinmin    1213820 Sep 19 20:21 formatted-combined-doc
-rw-r--r--    1 qinmin   qinmin    1213325 Sep 19 20:21 formatted-combined-txt
-rw-r--r--    1 qinmin   qinmin    3476712 Sep 19 20:10 ssearch-output
-rw-r--r--    1 qinmin   qinmin       5094 Sep 19 20:21 statistics-of-term-counts
-rw-r--r--    1 qinmin   qinmin       2633 Sep 19 20:10 structured-query-of-2006topics.txt
-rw-r--r--    1 qinmin   qinmin    1779813 Sep 19 20:16 term-count-of-docs
-rw-r--r--    1 qinmin   qinmin    2359028 Sep 19 20:21 term-count-of-feedbacks
-rw-r--r--    1 qinmin   qinmin    1763485 Sep 19 20:19 term-count-of-txts
-rw-r--r--    1 qinmin   qinmin    1382840 Sep 19 20:10 top-passageIDs
-rw-r--r--    1 qinmin   qinmin    7088825 Sep 19 20:21 top-passages
```

A.10 Scripts for the TREC 2005 and 2004 Genomics Track

The script for the TREC 2005 Genomics Track is:

```
javac -cp .:xmlrpc.jar BioNLP.java

javac -cp .:xmlrpc.jar ssearch.java

javac relex.java

cd /dbms/scratch/gen_demo/parse

java -cp .:Piccolo.jar gen_parser /dbms/scratch/genomic/parse/04trec

cd /dbms/scratch/gen_demo/okapi/bin

./convert_runtime -c ../databases/ 2004gendoc<../../parse/gen_exch.exch

./ix1 -mem 500 -delfinal -doclens 2004gendoc 0 | ./ixf 2004gendoc 0

java -Djava.library.path=. -cp .:xmlrpc.jar ssearch adhoc05 0.11

rm -f result05

cat result/txt*>result05

trec_eval genomics.qrels.txt result05
```

The script for the TREC 2004 Genomics Track is:

```
javac -cp .:xmlrpc.jar BioNLP.java

javac -cp .:xmlrpc.jar ssearch.java

javac relex.java cd /dbms/scratch/gen_demo/parse

java -cp .:Piccolo.jar gen_parser /dbms/scratch/genomic/parse/04trec

cd /dbms/scratch/gen_demo/okapi/bin

./convert_runtime -c ../databases/ 2004gendoc <../../parse/gen_exch.exch

./ix1 -mem 500 -delfinal -doclens 2004gendoc 0 | ./ixf 2004gendoc 0

java -Djava.library.path=. -cp .:xmlrpc.jar ssearch adhoc04 0.22

rm -f result04 cat result/txt*>result04

trec_eval 04.qrels.txt result04
```

B Topics in the Experiments

B.1 Topics of the TREC Genomics Tracks

The topics are presented in the figures as follows.

- The topics from 200 to 235, are the tasks in 2007.
- The topics from 160 to 187, are the tasks in 2006.
- The topics from 100 to 149, are the tasks in 2005.
- The topics from 1 to 50, are the tasks in 2004.

B.2 Topics of the TREC Entity Track

```
<query>
<num>1</num>
<entity_name>Blackberry</entity_name>
<entity_URL>clueweb09-en0004-50-39593</entity_URL>
<target_entity>organization</target_entity>
<narrative>Carriers that Blackberry makes phones for.</narrative>
```

```
</query>

<query>
<num>2</num>
<entity_name>ACM Athena award</entity_name>
<entity_URL>clueweb09-en0004-21-12770</entity_URL>
<target_entity>person</target_entity>
<narrative>Winners of the ACM Athena award.</narrative>
</query>

<query>
<num>3</num>
<entity_name>Claire Cardie</entity_name>
<entity_URL>clueweb09-en0009-89-01791</entity_URL>
<target_entity>person</target_entity>
<narrative>Students of Claire Cardie.</narrative>
</query>

<query>
<num>4</num>
<entity_name>Philadelphia, PA</entity_name>
<entity_URL>clueweb09-en0011-13-07330</entity_URL>
<target_entity>organization</target_entity>
<narrative>Professional sports teams in Philadelphia.</narrative>
</query>

<query>
<num>5</num>
<entity_name>Medimmune, Inc.</entity_name>
<entity_URL>clueweb09-en0008-26-39300</entity_URL>
```

<target_entity>product</target_entity>

<narrative>Products of Medimmune, Inc.</narrative>

</query>

<query>

<num>6</num>

<entity_name>Nobel Prize</entity_name>

<entity_URL>clueweb09-en0002-23-19459</entity_URL>

<target_entity>organization</target_entity>

<narrative>Organizations that award Nobel prizes.</narrative>

</query>

<query>

<num>7</num>

<entity_name>Boeing 747</entity_name>

<entity_URL>clueweb09-en0005-75-02292</entity_URL>

<target_entity>organization</target_entity>

<narrative>Airlines that currently use Boeing 747 planes.</narrative>

</query>

<query>

<num>8</num>

<entity_name>The King's Singers</entity_name>

<entity_URL>clueweb09-en0002-63-29621</entity_URL>

<target_entity>product</target_entity>

<narrative>CDs released by the King's Singers.</narrative>

</query>

<query>

<num>9</num>

<entity_name>The Beaux Arts Trio</entity_name>

<entity_URL>clueweb09-en0005-08-02741</entity_URL>

<target_entity>person</target_entity>

<narrative>Members of The Beaux Arts Trio.</narrative>

</query>

<query>

<num>10</num>

<entity_name>Indiana University</entity_name>

<entity_URL>clueweb09-en0007-37-37513</entity_URL>

<target_entity>organization</target_entity>

<narrative>Campuses of Indiana University.</narrative>

</query>

<query>

<num>11</num>

<entity_name>Home Depot Foundation</entity_name>

<entity_URL>clueweb09-en0009-23-04855</entity_URL>

<target_entity>organization</target_entity>

<narrative>Donors to the Home Depot Foundation.</narrative>

</query>

<query>

<num>12</num>

<entity_name>Air Canada</entity_name>

<entity_URL>clueweb09-en0004-24-03450</entity_URL>

<target_entity>organization</target_entity>

<narrative>Airlines that Air Canada has code share flights with.</narrative>

</query>

<query>
<num>13</num>
<entity_name>American Veterinary Medical Association (AVMA)</entity_name>
<entity_URL>clueweb09-en0004-39-32528</entity_URL>
<target_entity>product</target_entity>
<narrative>Journals published by the AVMA.</narrative>
</query>

<query>
<num>14</num>
<entity_name>Bouchercon 2007</entity_name>
<entity_URL>clueweb09-en0005-48-25203</entity_URL>
<target_entity>person</target_entity>
<narrative>Authors awarded an Anthony Award at Bouchercon in 2007.</narrative>
</query>

<query>
<num>15</num>
<entity_name>SEC conference</entity_name>
<entity_URL>clueweb09-en0010-56-11826</entity_URL>
<target_entity>organization</target_entity>
<narrative>Universities that are members of the SEC conference for football.
</narrative>
</query>

<query>
<num>16</num>
<entity_name>Mancuso Quilt Festivals</entity_name>
<entity_URL>clueweb09-en0011-22-08631</entity_URL>
<target_entity>organization</target_entity>

<narrative>Sponsors of the Mancuso quilt festivals.</narrative>

</query>

<query>

<num>17</num>

<entity_name>The Food Network</entity_name>

<entity_URL>clueweb09-en0006-55-17239</entity_URL>

<target_entity>person</target_entity>

<narrative>Chefs with a show on the Food Network.</narrative>

</query>

<query>

<num>18</num>

<entity_name>Jefferson Airplane</entity_name>

<entity_URL>clueweb09-en0009-25-04698</entity_URL>

<target_entity>person</target_entity>

<narrative>Members of the band Jefferson Airplane.</narrative>

</query>

<query>

<num>19</num>

<entity_name>John L. Hennessy</entity_name>

<entity_URL>clueweb09-en0011-14-04774</entity_URL>

<target_entity>organization</target_entity>

<narrative>Companies that John Hennessy serves on the board of.</narrative>

</query>

<query>

<num>20</num>

<entity_name>Isle of Islay</entity_name>

<entity_URL>clueweb09-en0008-96-25389</entity_URL>

<target_entity>organization</target_entity>

<narrative>Scotch whisky distilleries on the island of Islay.</narrative>

</query>

<query>

<num>21</num>

<entity_name>Bethesda, Maryland</entity_name>

<entity_URL>clueweb09-en0004-43-35557</entity_URL>

<target_entity>location</target_entity>

<narrative>What art galleries are located in Bethesda, Maryland?</narrative>

</query>

<query>

<num>22</num>

<entity_name>Organization of Petroleum Exporting Countries (OPEC)</entity_name>

<entity_URL>clueweb09-en0010-21-28880</entity_URL>

<target_entity>location</target_entity>

<narrative>Find countries that are members of OPEC (the Organization of Petroleum Exporting Countries).</narrative>

</query>

<query>

<num>23</num>

<entity_name>The Kingston Trio</entity_name>

<entity_URL>clueweb09-en0009-81-29533</entity_URL>

<target_entity>organization</target_entity>

<narrative>What recording companies now sell the Kingston Trio's songs? </narrative>

</query>

<query>
<num>24</num>
<entity_name>Jazz at Lincoln Center Orchestra</entity_name>
<entity_URL>clueweb09-en0008-04-03983</entity_URL>
<target_entity>person</target_entity>
<narrative>Find homepages of the members of the Jazz at Lincoln Center Orchestra.
</narrative>
</query>

<query>
<num>25</num>
<entity_name>U.S. Supreme Court</entity_name>
<entity_URL>clueweb09-en0012-87-19363</entity_URL>
<target_entity>organization</target_entity>
<narrative>From what schools did the Supreme Court justices receive their
undergraduate degrees?</narrative>
</query>

<query>
<num>26</num>
<entity_name>Cray XT computer</entity_name>
<entity_URL>clueweb09-en0021-77-27493</entity_URL>
<target_entity>organization</target_entity>
<narrative>Who has installed (taken delivery of) a Cray XT computer?</narrative>
</query>

<query>
<num>27</num>

<entity_name>Department of Mathematics, Montgomery College, Rockville Campus
</entity_name>
<entity_URL>clueweb09-en0024-21-25742</entity_URL>
<target_entity>organization</target_entity>
<narrative>Who are the publishers of the text books used in this department?
</narrative>
</query>

<query>
<num>28</num>
<entity_name>IEEE Engineering in Medicine and Biology Society</entity_name>
<entity_URL>clueweb09-en0007-95-06573</entity_URL>
<target_entity>product</target_entity>
<narrative>Find journals published by the IEEE Engineering in Medicine and
Biology society.</narrative>
</query>

<query>
<num>29</num>
<entity_name>Dow Jones</entity_name>
<entity_URL>clueweb09-en0006-73-08332</entity_URL>
<target_entity>organization</target_entity>
<narrative>Find companies that are included in the Dow Jones industrial average.
</narrative>
</query>

<query>
<num>30</num>
<entity_name>Ocean Spray Cranberries, Inc.</entity_name>
<entity_URL>clueweb09-en0132-45-30062</entity_URL>

<target_entity>location</target_entity>

<narrative>Find U.S. states and Canadian provinces where Ocean Spray growers are located.</narrative>

</query>

<query>

<num>31</num>

<entity_name>American Institute of Architects</entity_name>

<entity_URL>clueweb09-en0004-38-10748</entity_URL>

<target_entity>organization</target_entity>

<narrative>Find chapters of the American Institute of Architects.</narrative>

</query>

<query>

<num>32</num>

<entity_name>National Endowment for the Humanities</entity_name>

<entity_URL>clueweb09-en0010-48-02248</entity_URL>

<target_entity>organization</target_entity>

<narrative>What schools received collaborative research awards from the National Endowment for the Humanities (NEH) in 2008?</narrative>

</query>

<query>

<num>33</num>

<entity_name>DARPA Grand Challenge</entity_name>

<entity_URL>clueweb09-en0021-49-06565</entity_URL>

<target_entity>organization</target_entity>

<narrative>What organizations were able to complete the DARPA Grand Challenge in the 2007 contest?</narrative>

</query>

<query>
<num>34</num>
<entity_name>Access America Travel Insurance</entity_name>
<entity_URL>clueweb09-en0004-42-13808</entity_URL>
<target_entity>organization</target_entity>
<narrative>What airlines have financial default coverage with Access America Travel Insurance and Assistance?</narrative>
</query>

<query>
<num>35</num>
<entity_name>University of Maryland</entity_name>
<entity_URL>clueweb09-en0008-34-21132</entity_URL>
<target_entity>organization</target_entity>
<narrative>Which companies have representation on the board of directors of the University of Maryland?</narrative>
</query>

<query>
<num>36</num>
<entity_name>Ford Motor Company</entity_name>
<entity_URL>clueweb09-en0120-17-13549</entity_URL>
<target_entity>organization</target_entity>
<narrative>What companies build parts used in production of Ford vehicles?</narrative>
</query>

<query>
<num>37</num>
<entity_name>Tavis Smiley Show</entity_name>

<entity_URL>clueweb09-en0009-57-35033</entity_URL>

<target_entity>person</target_entity>

<narrative>Find homepages of people that appeared on the Tavis Smiley show in December 2008.</narrative>

</query>

<query>

<num>38</num>

<entity_name>Richard Petty Motorsports</entity_name>

<entity_URL>clueweb09-en0133-45-25168</entity_URL>

<target_entity>person</target_entity>

<narrative>Who are the drivers and crew chiefs for Richard Petty Motorsports?</narrative>

</query>

<query>

<num>39</num>

<entity_name>Maryland Farm Bureau</entity_name>

<entity_URL>clueweb09-en0008-95-39976</entity_URL>

<target_entity>organization</target_entity>

<narrative>Find companies that offer benefits to members of the Maryland Farm Bureau.

</narrative>

</query>

<query>

<num>40</num>

<entity_name>Costco</entity_name>

<entity_URL>clueweb09-en0006-60-20817</entity_URL>

<target_entity>organization</target_entity>

<narrative>Find homepages of manufacturers of LCD televisions sold by Costco.</narrative>

</query>

<query>

<num>41</num>

<entity_name>New York Times</entity_name>

<entity_URL>clueweb09-en0012-63-17728</entity_URL>

<target_entity>person</target_entity>

<narrative>Find homepages of the regular opinion columnists of the New York Times.

</narrative>

</query>

<query>

<num>42</num>

<entity_name>General Electric</entity_name>

<entity_URL>clueweb09-en0007-76-25787</entity_URL>

<target_entity>organization</target_entity>

<narrative>Find homepages of subsidiaries of General Electric.</narrative>

</query>

<query>

<num>43</num>

<entity_name>Mystery Writers of America</entity_name>

<entity_URL>clueweb09-en0002-73-25630</entity_URL>

<target_entity>person</target_entity>

<narrative>Find homepages of presidents of the Mystery Writers Association (MWA).

</narrative>

</query>

<query>

<num>44</num>

<entity_name>Museum of Art and Design</entity_name>

<entity_URL>clueweb09-en0127-01-43076</entity_URL>

<target_entity>location</target_entity>

<narrative>I want to see the homepages of places visited during the June 2008 MADtrip to Seattle, Washington.</narrative>

</query>

<query>

<num>45</num>

<entity_name>National Zoo</entity_name>

<entity_URL>clueweb09-en0003-82-03073</entity_URL>

<target_entity>organization</target_entity>

<narrative>Who are the corporate partners of the National Zoo?</narrative>

</query>

<query>

<num>46</num>

<entity_name>Fidos for Freedom</entity_name>

<entity_URL>clueweb09-en0006-65-25191</entity_URL>

<target_entity>organization</target_entity>

<narrative>What facilities are visited by the therapy dogs of Fidos for Freedom?

</narrative>

</query>

<query>

<num>47</num>

<entity_name>Albuquerque International Balloon Fiesta</entity_name>

<entity_URL>clueweb09-en0004-73-28331</entity_URL>

<target_entity>organization</target_entity>

<narrative>Who are the balloon manufacturers associated with the Albuquerque International Balloon Fiesta?</narrative>

</query>

<query>
<num>48</num>
<entity_name>AFL-CIO</entity_name>
<entity_URL>clueweb09-en0004-59-11324</entity_URL>
<target_entity>organization</target_entity>
<narrative>Find unions that are members of the AFL-CIO.</narrative>
</query>

<query>
<num>49</num>
<entity_name>Eurail</entity_name>
<entity_URL>clueweb09-en0006-86-35399</entity_URL>
<target_entity>location</target_entity>
<narrative>What countries does Eurail operate in?</narrative>
</query>

<query>
<num>50</num>
<entity_name>Abu Dhabi</entity_name>
<entity_URL>clueweb09-en0004-39-08867 </entity_URL>
<target_entity>organization</target_entity>
<narrative>Find companies owned or controlled by Abu Dhabi.</narrative>
</query>

<query>
<num>51</num>
<entity_name>National Institutes of Health (NIH)</entity_name>
<entity_URL>clueweb09-en0009-58-39495</entity_URL>

<target_entity>organization</target_entity>

<narrative>What organizations comprise the National Institutes of Health (NIH)?

</narrative>

</query>

<query>

<num>52</num>

<entity_name>Smithsonian Institution</entity_name>

<entity_URL>clueweb09-en0011-99-06195</entity_URL>

<target_entity>person</target_entity>

<narrative>Find the members of the Board of Regents of the Smithsonian Institution.

</narrative>

</query>

<query>

<num>53</num>

<entity_name>Foundation Morgan horses</entity_name>

<entity_URL>clueweb09-en0007-32-10466</entity_URL>

<target_entity>organization</target_entity>

<narrative>Who are breeders of Foundation Morgan horses? </narrative>

</query>

<query>

<num>54</num>

<entity_name>Knockout Mouse Project (KOMP)</entity_name>

<entity_URL>clueweb09-en0009-58-40020</entity_URL>

<target_entity>organization</target_entity>

<narrative>What organizations are the participants in the NIH sponsored Knockout

Mouse Project?</narrative>

</query>

<query>

<num>55</num>

<entity_name>Edgars Awards</entity_name>

<entity_URL>clueweb09-en0019-86-23998</entity_URL>

<target_entity>person</target_entity>

<narrative>Who are the 2008 Edgars Awards nominees?</narrative>

</query>

<query>

<num>56</num>

<entity_name>Nature Conservancy</entity_name>

<entity_URL>clueweb09-en0010-64-18232</entity_URL>

<target_entity>organization</target_entity>

<narrative>What companies or organizations have officers that sit on the Nature

Conservancy's board of directors?</narrative>

</query>

<query>

<num>57</num>

<entity_name>Kennedy Center</entity_name>

<entity_URL>clueweb09-en0002-36-00572</entity_URL>

<target_entity>person</target_entity>

<narrative>Find the homepages of the Kennedy Center honorees.</narrative>

</query>

<query>

<num>58</num>

<entity_name>University of Michigan</entity_name>

<entity_URL>clueweb09-en0003-16-28592</entity_URL>

<target_entity>organization</target_entity>

<narrative>What are some of the spin-off companies from the University of Michigan?

</narrative>

</query>

<query>

<num>59</num>

<entity_name>Fulbright Scholars Program</entity_name>

<entity_URL>clueweb09-en0005-77-19875</entity_URL>

<target_entity>person</target_entity>

<narrative>Who are the Fulbright scholars in computer sciences?</narrative>

</query>

<query>

<num>60</num>

<entity_name>Peabody Essex Museum</entity_name>

<entity_URL>clueweb09-en0009-48-18543</entity_URL>

<target_entity>organization</target_entity>

<narrative>I would like to see the homepages of the corporate partners of the Peabody

Essex Museum.</narrative>

</query>

<query>

<num>61</num>

<entity_name>The Association for Symbolic Logic</entity_name>

<entity_URL>clueweb09-en0004-85-27331</entity_URL>

<target_entity>organization</target_entity>

<narrative>Who are institutional members of the Association for Symbolic Logic (ASL)?

</narrative>

</query>

<query>
<num>62</num>
<entity_name>Baltimore</entity_name>
<entity_URL>clueweb09-en0004-40-10287</entity_URL>
<target_entity>organization</target_entity>
<narrative>What cruise lines have cruises originating in Baltimore?</narrative>
</query>

<query>
<num>63</num>
<entity_name>Drew Gilpin Faust</entity_name>
<entity_URL>clueweb09-en0010-43-02957</entity_URL>
<target_entity>organization</target_entity>
<narrative>What institutions Drew Gilpin Faust been affiliated with, for example as a trustee, board member, president, etc?</narrative>
</query>

<query>
<num>64</num>
<entity_name>Pulitzer Prize</entity_name>
<entity_URL>clueweb09-en0011-51-07177</entity_URL>
<target_entity>organization</target_entity>
<narrative>What newspapers won the Pulitzer Prize for journalism in 2007?
</narrative>
</query>

<query>
<num>65</num>
<entity_name>Patricia Cornwell</entity_name>

<entity_URL>clueweb09-en0010-97-31272</entity_URL>

<target_entity>organization</target_entity>

<narrative>What institutions is Patricia Cornwell a member of?</narrative>

</query>

<query>

<num>66</num>

<entity_name>Southwest Airlines</entity_name>

<entity_URL>clueweb09-en0010-03-33074</entity_URL>

<target_entity>location</target_entity>

<narrative>I want to see the homepages of places to see in Cleveland listed by Southwest Airlines.</narrative>

</query>

<query>

<num>67</num>

<entity_name>Vacations By Rail</entity_name>

<entity_URL>clueweb09-en0013-76-14954</entity_URL>

<target_entity>location</target_entity>

<narrative>Find national parks visited on the Vacations by Rail tours.</narrative>

</query>

<query>

<num>68</num>

<entity_name>Association of American Medical Colleges</entity_name>

<entity_URL>clueweb09-en0004-91-09128</entity_URL>

<target_entity>organization</target_entity>

<narrative>Find medical schools that are members of the AAMC.</narrative>

</query>

<query>

<num>69</num>

<entity_name>Newseum</entity_name>

<entity_URL>clueweb09-en0010-82-23701</entity_URL>

<target_entity>organization</target_entity>

<narrative>Who are the founding partners of the Newseum?</narrative>

</query>

<query>

<num>70</num>

<entity_name>Teach For America</entity_name>

<entity_URL>clueweb09-en0011-19-31196</entity_URL>

<target_entity>organization</target_entity>

<narrative>Find the Teach For America National Growth Fund Investors for 2006-2010

</narrative>

</query>

Figure B.1: 2007 Topics (1/2)

200 What serum [PROTEINS] change expression in association with high disease activity in lupus?

201 What [MUTATIONS] in the Raf gene are associated with cancer?

202 What [DRUGS] are associated with lysosomal abnormalities in the nervous system?

203 What [CELL OR TISSUE TYPES] express receptor binding sites for vasoactive intestinal peptide (VIP) on their cell surface?

204 What nervous system [CELL OR TISSUE TYPES] synthesize neurosteroids in the brain?

205 What [SIGNS AND SYMPTOMS] of anxiety disorder are related to coronary artery disease?

206 What [TOXICITIES] are associated with zoledronic acid?

207 What [TOXICITIES] are associated with etidronate?

208 What [BIOLOGICAL SUBSTANCES] have been used to measure toxicity in response to zoledronic acid?

209 What [BIOLOGICAL SUBSTANCES] have been used to measure toxicity in response to etidronate?

210 What [MOLECULAR FUNCTIONS] are attributed to glycan modification?

211 What [ANTIBODIES] have been used to detect protein PSD-95?

212 What [GENES] are involved in insect segmentation?

213 What [GENES] are involved in Drosophila neuroblast development?

214 What [GENES] are involved axon guidance in C.elegans?

215 What [PROTEINS] are involved in actin polymerization in smooth muscle?

216 What [GENES] regulate puberty in humans?

Figure B.2: 2007 Topics (2/2)

217 What [PROTEINS] in rats perform functions different from those of their human homologs?

218 What [GENES] are implicated in regulating alcohol preference?

219 In what [DISEASES] of brain development do centrosomal genes play a role?

220 What [PROTEINS] are involved in the activation or recognition mechanism for PmrD?

221 Which [SIGNALING PATHWAYS] are mediated by CD44?

222 What [MOLECULAR FUNCTIONS] is LITAF involved in?

223 Which anaerobic bacterial [STRAINS] are resistant to Vancomycin?

224 What [GENES] are involved in the melanogenesis of human lung cancers?

225 What [BIOLOGICAL SUBSTANCES] induce clpQ expression?

226 What [PROTEINS] make up the murine signal recognition particle?

227 What [GENES] are induced by LPS in diabetic mice?

228 What [GENES] when altered in the host genome improve solubility of heterologously expressed proteins?

229 What [SIGNS OR SYMPTOMS] are caused by human parvovirus infection?

230 What [PATHWAYS] are involved in Ewing's sarcoma?

231 What [TUMOR TYPES] are found in zebrafish?

232 What [DRUGS] inhibit HIV type 1 infection?

233 What viral [GENES] affect membrane fusion during HIV infection?

234 What [GENES] make up the NFkappaB signaling pathway?

235 Which [GENES] involved in NFkappaB signaling regulate iNOS?

Figure B.3: 2006 Topics (1/2)

<160>What is the role of PrnP in mad cow disease

<161>What is the role of IDE in Alzheimer's disease

<162>What is the role of MMS2 in cancer

<163>What is the role of APC (adenomatous polyposis coli) in colon cancer

<164>What is the role of Nurr-77 in Parkinson's disease

<165>How do Cathepsin D (CTSD) and apolipoprotein E (ApoE) interactions contribute to Alzheimer's disease

<166>What is the role of Transforming growth factor-beta1 (TGF-beta1) in cerebral amyloid angiopathy (CAA)

<167>How does nucleoside diphosphate kinase (NM23) contribute to tumor progression

<168>How does BARD1 regulate BRCA1 activity

<169>How does APC (adenomatous polyposis coli) protein affect actin assembly

<170>How does COP2 contribute to CFTR export from the endoplasmic reticulum

<171>How does Nurr-77 delete T cells before they migrate to the spleen or lymph nodes and how does this impact autoimmunity

<172>How does p53 affect apoptosis

<173>How do alpha7 nicotinic receptor subunits affect ethanol metabolism

<174>How does BRCA1 ubiquitinating activity contribute to cancer

<175>How does L2 interact with L1 to form HPV11 viral capsids

<176>How does Sec61-mediated CFTR degradation contribute to cystic fibrosis

<177>How do Bop-Pes interactions affect cell growth

Figure B.4: 2006 Topics (2/2)

<178>How do interactions between insulin-like GFs and the insulin receptor affect skin biology

<179>How do interactions between HNF4 and COUP-TF1 suppress liver function

<180>How do Ret-GDNF interactions affect liver development

<181>How do mutations in the Huntingtin gene affect Huntington's disease

<182>How do mutations in Sonic Hedgehog genes affect developmental disorders

<183>How do mutations in the NM23 gene affect tracheal development

<184>How do mutations in the Pes gene affect cell growth

<185>How do mutations in the hypocretin receptor 2 gene affect narcolepsy

<186>How do mutations in the Presenilin-1 gene affect Alzheimer's disease

<187>How do mutations in familial hemiplegic migraine type 1 (FHM1) gene affect calcium ion influx in hippocampal neurons

Figure B.5: 2005 Topics (1/3)

<100>procedure or methods for how to open up a cell through a process called electroporation.

<101>procedure or methods for exact reactions that take place when you do glutathione S-transferase (GST) cleavage during affinity chromatography.

<102>procedure or methods for different quantities of different components to use when pouring a gel to make it more or less porous.

<103>procedure or methods for green fluorescent protein (GFP) tagged proteins to do experiments with tagged proteins.

<104>procedure or methods for how to do a microsomal budding assay, i.e., budding of vesicles from microsomes in vitro.

<105>procedure or methods for purification of rat IgM.

<106>procedure or methods for chromatin IP (Immuno Precipitations) to isolate proteins that are bound to DNA in order to precipitate the proteins out of the DNA.

<107>procedure or methods for normalization procedures that are used for microarray data.

<108>procedure or methods for identifying in vivo protein-protein interactions in time and space in the living cell.

<109>procedure or methods for fluorogenic 5-nuclease assay.

<110>role of the gene Interferon-beta in the disease Multiple Sclerosis.

<111>role of the gene PRNP in the disease Mad Cow Disease.

<112>role of the gene IDE gene in the disease Alzheimer's Disease.

<113>role of the gene MMS2 in the disease Cancer.

<114>role of the gene APC (adenomatous polyposis coli) in the disease Colon Cancer.

Figure B.6: 2005 Topics (2/3)

<115>role of the gene Nurr-77 in the disease Parkinson's Disease.

<116>role of the gene Insulin receptor gene in the disease Cancer.

<117>role of the gene Apolipoprotein E (ApoE) in the disease Alzheimer's Disease.

<118>role of the gene Transforming growth factor-beta1 (TGF-beta1) in the disease Cerebral Amyloid Angiopathy (CAA).

<119>role of the gene GSTM1 in the disease Breast Cancer.

<120>role of the gene nucleoside diphosphate kinase (NM23) in the process of tumor progression.

<121>role of the gene BARD1 in the process of BRCA1 regulation.

<122>role of the gene APC (adenomatous polyposis coli) in the process of actin assembly.

<123>role of the gene COP2 in the process of transport of CFTR out of the endoplasmic reticulum.

<124>role of the gene casein kinase II in the process of ribosome assembly.

<125>role of the gene Nurr-77 in the process of preventing auto-immunity by deleting reactive T-cells before they migrate to the spleen or the lymph nodes.

<126>role of the gene P53 in the process of apoptosis.

<127>role of the gene alpha7 nicotinic receptor subunit gene in the process of ethanol metabolism.

<128>role of the gene gamma-aminobutyric acid receptors (GABABRs) in the process of inhibitory synaptic transmission.

<129>role of the gene Interferon-beta in the process of viral entry into host cell.

<130>genes BRCA1 regulation of ubiquitin in cancer.

<131>genes L1 and L2 in the HPV11 virus in the role of L2 in the viral capsid.

Figure B.7: 2005 Topics (3/3)

<132>genes APC (adenomatous polyposis coli) and wnt in colon cancer.

<133>genes phospholipase A2 (PLA2) and SAR1 in Endoplasmic reticulum transport (i.e. vesicle budding from the ER).

<134>genes CFTR and Sec61 in degradation of CFTR which leads to cystic fibrosis.

<135>genes Bop and Pes in cell growth.

<136>genes alpha7 nicotinic receptor gene and ApoE gene in the neurotoxic effects of ethanol.

<137>genes Insulin-like GF and insulin receptor gene in the function in skin.

<138>genes HNF4 and COUP-TF I in the suppression in the function of the liver.

<139>genes Ret and GDNF in kidney development.

<140>BRCA1 185delAG mutation and its/their role in ovarian cancer.

<141>Huntingtin mutations and its/their role in Huntington's Disease.

<142>Sonic hedgehog mutations and its/their role in developmental disorders.

<143>Mutations of NM23 and its/their impact on tracheal development.

<144>Mutations in metazoan Pes and its/their effect on cell growth.

<145>Mutations of hypocretin receptor 2 and its/their role in narcolepsy.

<146>Mutations of presenilin-1 gene and its/their biological impact in Alzheimer's disease.

<147>Mutations of alpha7 nAChR gene and its/their biological impact in alcoholism.

<148>Mutation of familial hemiplegic migraine type 1 (FHM1) and its/their neuronal Ca2+ influx in hippocampal neurons.

<149>Mutations of the alpha 4-GABAA receptor and its/their impact on behavior.

Figure B.8: 2004 Topics (1/4)

<01>Ferroportin-1 in humans, Find articles about Ferroportin-1, an iron transporter, in humans.

<02>Generating transgenic mice, Find protocols for generating transgenic mice.

<03>Time course for gene expression in mouse kidney, What is the time course of gene expression in the murine developing kidney?

<04>Gene expression profiles for kidney in mice, What mouse genes are specific to the kidney?

<05>Protocols for isolating cell nuclei, Articles are relevant if they describe methods for subcellular fractionation of nuclei.

<06>FancD2, Find articles about function of FancD2.

<07>DNA repair and oxidative stress, Find correlation between DNA repair pathways and oxidative stress.

<08>Correlation between DNA repair pathways and skin cancer, Genes and proteins (pathways) common to DNA repair, oxidative diseases, skin-carcinogenesis, and skin-carcinogenesis, .

<09>mutY,Find articles about the function of mutY in humans.

<10>NEIL1, Find articles about the role of NEIL1 in repair of DNA.

<11>Carcinogenesis and hairless mice, Find articles regarding carcinogenesis induced in hairless mice.

<12>Genes regulated by Smad4, Find articles describing genes that are regulated by the signal transducing molecule Smad4.

<13>Role of TGFB in angiogenesis in skin,Documents regarding the role of TGFB in angiogenesis in skin with respect to homeostasis and development.

<14>Expression or Regulation of TGFB in HNSCC cancers, Documents regarding TGFB expression or regulation in HNSCC cancers.

Figure B.9: 2004 Topics (2/4)

<15>ATPase and apoptosis, Find information on role of ATPases in apoptosis

<16>AAA proteins, How do AAA proteins mediate interaction with lipids or DNA and what is their functional impact?

<17>DO1 antibody, Determine binding affinity of anti-p53 monoclonal antibody DO1.

<18>Gis4, Properties of Gis4 with respect to cell cycle and/or metabolism.

<19>Comparison of Promoters of GAL1 and SUC1, What similarities and differences exist between the upstream promoter regions of GAL1 and SUC1? Are there co-repressors or co-activators? If so, are they regulated by SNF 1?

<20>Substrate modification by ubiquitin, Which biological processes are regulated by having constituent proteins modified by covalent attachment to ubiquitin or ubiquitin-like proteins?

<21>Role of p63 and p73 in relation to DNA damage, Do p63 and p73 cause cell cycle arrest or apoptosis related to DNA damage?

<22>Relative response of p53 family members to agents causing single-stranded versus double-stranded DNA breaks, Does p53 respond differently to different DNA-damaging agents? Do they respond differently to single-strand versus double-strand breaks?

<23>Saccharomyces cerevisiae proteins involved in ubiquitin system, Which Saccharomyces cerevisiae proteins are involved in the ubiquitin proteolytic pathway?

<24>Mouse peptidoglycan recognition proteins (PGRP), Find all reports describing mouse peptidoglycan recognition proteins (PGRP).

<25>Cause of scleroderma, Identify studies that include genome-wide scans and microarray analysis in the investigation of scleroderma.

<26>Function of BUB2/BFA1 in the process of cytokinesis, Retrieval of information regarding the role of BUB2 and BFA1 in cytokinesis in yeast.

Figure B.10: 2004 Topics (3/4)

<27>Role of autophagy in apoptosis, Experiments establishing positive or negative interconnection between autophagy and apoptosis.

<28>Proteases that function in both apoptosis and autophagy cell death, Studies that investigate similarities in morphological changes among apoptosis and autophagy processes.

<29>Phenotypes of gyrA mutations, Documents containing the sequences and phenotypes of E. coli gyrA mutations.

<30>Regulatory targets of the Nkx gene family members, Documents identifying genes regulated by Nkx gene family members.

<31>TOR signaling in neurofibromatosis, Reports that provide possible links between neurofibromatosis and TOR signaling.

<32>Xenograft animal models of tumorogenesis, Find reports that describe xenograft models of human cancers.

<33>Mice, mutant strains, and Histoplasmosis, Identify research on mutant mouse strains and factors which increase susceptibility to infection by Histoplasma capsulatum.

<34>Gene products of Cryptococcus important to fungal survival, Articles reporting experiments allowing annotation of gene products of Cryptococcus.

<35>WD40 repeat-containing proteins, What is the function of proteins containing WD40 repeats?

<36>RAB3A, Background information on RAB3A.

<37>PAM, What research is being done on peptide amidating enzyme, PAM?

<38>Risk factors for stroke, Information concerning genetic loci that are associated with increased risk of stroke, such as apolipoprotein E4 or factor V mutations.

Figure B.11: 2004 Topics (4/4)

<39>Hypertension, Identify genes as potential genetic risk factors candidates for causing hypertension.

<40>Antigens expressed by lung epithelial cells, To identify the antigens expressed by lung epithelial cells and the antibodies available.

<41>Mutations in the Cystic Fibrosis conductance regulator gene, What phenotypes have been described resulting from mutations in the Cystic Fibrosis conductance regulator gene?

<42>Genes altered by chromosome translocations, What genes show altered behavior due to chromosomal rearrangements?

<43>Sleeping Beauty, Studies of Sleeping Beauty transposons.

<44>Proteins involved in the nerve growth factor pathway, Create a list of all the nerve growth factor pathway proteins.

<45>Mental Health Wellness-1, What genetic loci, such as Mental Health Wellness 1 (MWH1) are implicated in mental health?

<46>RSK2, What human biological processes is RSK2 known to be involved in?

<47>Human gene BCL-2 antagonists and inhibitors, Research the human gene BCL-2 to determine if there are antagonists and inhibitors inside of a cell.

<48>Human homologues of C. elegans UNC genes, What is the focus of studies involving the members of the human UNC gene family?

<49>Glyphosate tolerance gene sequence, Find reports and glyphosate tolerance gene sequences in the literature.

<50>Low temperature protein expression in E. coli, Find research on improving protein expressions at low temperature in Escherichia coli bacteria.

C Sample Raw Data

C.1 Sample HTML Raw Data of the TREC 2007 and 2006 Data Set

```
<html>
<body>
 <H2> A Strategy for the Rapid Identification of
Phosphorylation Sites in the Phosphoproteome
 <A NAME="RFN4"></A><SUP><A HREF="#FN4">*</A></SUP> </H2> <STRONG>
</NOBR> <NOBR>Justin A. MacDonald<SUP><IMG SRC="/math/Dagger.gif"
ALT="{ddagger}" BORDER="0"></SUP><A NAME="RFN5"></A><SUP>,<A
HREF="#FN5"><IMG SRC="/math/link//sect.gif" ALT="&#167;"
BORDER="0"></A></SUP></NOBR>, <NOBR>Aaron J. Mackey<A
NAME="RFN5"></A><SUP><A HREF="#FN5"><IMG SRC="/math/link//sect.gif"
ALT="&#167;" BORDER="0"></A></SUP><SUP>,&#182;</SUP></NOBR>,
<NOBR>William R. Pearson<SUP>||</SUP></NOBR> and <NOBR>Timothy A. J.
Haystead<SUP><IMG SRC="/math/Dagger.gif" ALT="{ddagger}"
BORDER="0"></SUP><SUP>,<A HREF="#COR1">**</A></SUP></NOBR>
</STRONG><P> <FONT SIZE=-1> <SUP>&#182;</SUP> Department of
```

Microbiology, University of Virginia, Charlottesville, Virginia 22908
 ^{||} Department of Biochemistry and Molecular Genetics, University of Virginia, Charlottesville, Virginia 22908
 ^{<IMG SRC="/math/Dagger.gif" ALT="{ddagger}" BORDER="0">} Department of Pharmacology and Cancer Biology, Duke University, Durham, North Carolina 27710 </FONT><P> <P> <A NAME="ABS"><!-- null --></A>
<BR CLEAR=right><TABLE WIDTH=100
% BGCOLOR=e1e1e1 CELLPADDING=0 CELLSPACING=0>
<TR><TD ALIGN=left VALIGN=middle WIDTH=5
% BGCOLOR=ffffff><IMG WIDTH=10 HEIGHT=21 HSPACE=5 ALT=" "
SRC="/icons/toc/rarrow.gif"></TD>
<TH ALIGN=left VALIGN=middle WIDTH=95%><FONT SIZE=+2>
ABSTRACT </FONT></TH></TR></TABLE>
<TABLE ALIGN=right CELLPADDING=5
BORDER><TR><TH ALIGN=left><FONT SIZE=-1> <A HREF="#top"><IMG
WIDTH=11 HEIGHT=9 BORDER=0 HSPACE=5 ALT=" "
SRC="/icons/toc/uarrow.gif">TOP
</A> <IMG WIDTH=11 HEIGHT=9
BORDER=0 HSPACE=5 ALT=" " SRC="/icons/toc/dot.gif"><FONT
COLOR=464c53>ABSTRACT</FONT>
 <A HREF="#SEC1"><IMG WIDTH=11
HEIGHT=9 BORDER=0 HSPACE=5 ALT=" "
SRC="/icons/toc/darrow.gif">EXPERIMENTAL PROCEDURES
</A> <A
HREF="#SEC2"><IMG WIDTH=11 HEIGHT=9 BORDER=0 HSPACE=5 ALT=" "
SRC="/icons/toc/darrow.gif">RESULTS
</A> <A HREF="#SEC3"><IMG

WIDTH=11 HEIGHT=9 BORDER=0 HSPACE=5 ALT=" "
SRC="/icons/toc/darrow.gif">DISCUSSION
</A> <A HREF="#BIBL"><IMG
WIDTH=11 HEIGHT=9 BORDER=0 HSPACE=5 ALT=" "
SRC="/icons/toc/darrow.gif">REFERENCES
</A>
</FONT></TH></TR></TABLE>
 Edman phosphate
(³²P) release sequencing provides a high sensitivity<SUP>
</SUP>means of identifying phosphorylation sites in proteins
thatcomplements mass spectrometry techniques. We have
developeda bioinformatic assessment tool, the cleavage
of radiolabeledprotein (CRP) program, which enables
experimental identificationof phosphorylation sites via
³²P labeling and Edman degradationof cleaved
proteins obtained at femtomole levels. By observingthe
Edman cycle(s) in which radioactivity is found, candidate<SUP>
</SUP>phosphorylation sites are identified by determining which
residuesoccur at the observed number of cycles
downstream from a peptidecleavage site. In cases where
more than one residue could beresponsible for the
observed radioactivity, additional experimentswith
cleavage reagents having alternative specificities may<SUP>
</SUP>resolve the ambiguity. Given a protein sequence and a
cleavagesite, CRP performs these experiments <I>in
silico</I>, identifying<SUP>
.....

```
<HR NOSHADE SIZE=1 WIDTH="50%" ALIGN="CENTER">
Proteomic technologies have transformed the manner in which<SUP>
</SUP>proteins and their contributions ..... access via
<SUP>32</SUP>P labeling<SUP> </SUP>and Edman sequencing to
concentrations of phosphorylation sites<SUP> </SUP>that are below
the femtomole level.<SUP> </SUP><P>
 <A NAME="SEC1"><!-- null --></A>
<BR CLEAR=right><TABLE WIDTH=100% BGCOLOR=e1e1e1 CELLPADDING=0 CELLSPACING=0>
 <TR><TD ALIGN=left VALIGN=middle WIDTH=5% BGCOLOR=ffffff>
 <IMG WIDTH=10 HEIGHT=21 HSPACE=5 ALT=" "
SRC="/icons/toc/rarrow.gif"></TD>
 <TH ALIGN=left VALIGN=middle WIDTH=95%><FONT SIZE=+2>  
EXPERIMENTAL PROCEDURES </FONT></TH></TR></TABLE> <TABLE ALIGN=right
CELLPADDING=5 BORDER><TR><TH ALIGN=left><FONT SIZE=-1> <A
HREF="#top"><IMG WIDTH=11 HEIGHT=9 BORDER=0 HSPACE=5 ALT=" "
SRC="/icons/toc/uarrow.gif">TOP<BR></A>
......
</FONT></TH></TR></TABLE>  <BR>
<STRONG><EM>Materials&#151;</EM></STRONG><BR> Mouse submaxillary
gland endoproteinase Arg-C and <I>Staphylococcus ..... panels
represent the worst and best combination of cleavage sites.<P>
</TD></TR></TABLE> </TD></TR></TABLE></CENTER> <BR> Additional
cleavage ..... obtained from an endoproteinase Glu-C digest of
immunoprecipitated Hsp27 are presented.<P> </TD></TR></TABLE>
```

```
</TD></TR></TABLE></CENTER> <BR> </BODY> </HTML>
```

C.2 Sample HTML Raw Data of the TREC 2005 and 2004 Data Set

```
<html>
<body>
SIZE=1><H2><FONT COLOR=00337F>Perspectives and
Editorials</FONT></H2></TD></TR></TABLE>
<H2><FONT FACE="arial,verdana,helvetica"><EM>
Editorial Commentary
</EM></FONT></H2>
<EM>
</NOBR><NOBR>Ben Avi Weissman</NOBR>
</EM><BR> <I>Department of Pharmacology<SUP> </SUP><BR> Israel
Institute for Biological Research<SUP> </SUP><BR> PO Box 19<SUP>
</SUP><BR> Ness Ziona 74100<SUP> </SUP><BR>
Israel</I></FONT></TD></TR></TABLE>
<P><FONT SIZE=+1 FACE="arial,verdana,helvetica"><A NAME=""><!-- null
--></A>
Culty M, Luo L, Yao Z, Chen H, Papadopoulos V, Zirkin B. Cholesterol
transport, peripheral benzodiazepine receptor, and steroidogenesis
in aging Leydig cells. <I>J Androl</I>.
2002 ;23:439&#150;447.<!-- HIGHWIRE ID="23:3:326:1" --><A
HREF="/cgi/ijlink?linkType=ABST&journalCode=jandrol&resid=23/3/439"
```

```
><nobr>[Abstract/<font COLOR="CC0000">Free</font> Full Text]
</nobr></A><!-- /HIGHWIRE --></FONT><P>
<HR SIZE=2><BR>Using the specific ligand ..... <\BR>
</BODY>
</HTML>
```

D Evaluation Scripts

Once all the relevant passage are retrieved, we will start evaluating their relevance by comparing them to a set of relevant passages determined by a panel of expert judges, which is introduced in Chapter 6 as gold standard.

To evaluate the relevance of retrieved passages, TREC also provides a simple Python program, which compares the gold standard file and the retrieved passage file, and then prints to the standard output under the evaluation measures. All the programs are available in the CD. The environment for the programs is to install the latest release of Python (currently 2.5.1). The script for the evaluation is presented as:

```
#!/bin/sh
#first argument is the path + name of combined output to be evaluated
#second argument is the output path
#third argument is the output file name
ls -l $2
./Python-2.5.1/python trecgen200X_score.py 200X.goldstd.tsv.txt
$1 > $3 mv $3 $2/$3
```

Taking the 2007 experimental result as an example, we call python from their respective folders as follows:

```
$ LOCATION_OF/Python-2.5.1/python LOCATION_OF/trecgen2007_score.py
LOCATION_OF/gold-stand-2007 LOCATION_OF/output-file-2007 >
LOCATION_OF/evaluation-result-of-output-file-2007
```

The gold standard file clearly changes year to year, and so does the evaluating program. The details of changes and how to run the programs are also published on the TREC web site and in my CD which is enclosed with this thesis.

E Research Publications

E.1 Refereed Journal Papers

1. Hu, Qinmin, Nie, Z., Huang, X. and Cercone, N., Integrating Proximity Information into Topic Models for Entity Search (23 pages), submitted to journal of ACM Transactions on Knowledge Discovery from Data (TKDD), 2012.

2. Hu, Qinmin and Huang, X., Enhancing Genomics Information Retrieval Through Dimensional Analysis (14 pages), accepted by journal of Bioinformatics and Computational Biology, 2012.

3. Hu, Qinmin, Huang, X. and Hu, X., Modeling and Mining Term Association for Improving Biomedical Information Retrieval Performance (35 pages), in journal of BMC Bioinformatics, 2012, 13(Suppl 9): S2 (11 June 2012). ISSN: 1471-2105.

4. Hu, Qinmin, Huang, X. and Miao J., A Robust Approach to Optimizing Multi-Source Information for Enhancing Genomics Retrieval Performance (18 pages), in journal of BMC Bioinformatics, 2011, 12(Suppl 5):S6 (27 July 2011). ISSN:

1471-2105.

5. Hu, Qinmin and Huang, X., Passage Extraction and Result Combination for Genomics Information Retrieval (23 pages), in journal of Intelligent Information Systems (JIIS), Springer-Verlag Publisher. ISSN (Printed): 0925-9902 and ISSN (Online): 1573-7675. Vol.34, No.3, 2010. pp.249-274.

E.2 Refereed Book Chapter

Andreopoulos, B., Huang, X., An, A., Labudde, D. and Hu, Qinmin, Promoting Diversity in Top Hits for Biomedical Passage Retrieval, in Z. Ras and A. Dardzinska (eds.): Advances in Data Management, Springer. 2009. pp 371-393. ISSN (Printed): 1860-949X and ISSN (Online): 1860-9503.

E.3 Refereed Conference Papers

1. Hu, Qinmin and Huang, X., When Crowdsourcing Meets Information Retrieval, submitted to the 36th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 2013.

2. Hu, Qinmin and Huang, X., Positional Topic Modelling for Entity Retrieval, submitted to the 36th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 2013.

3. Hu, Qinmin, Xu, Z. and Huang, X., York University at TREC 2012: CrowdSourcing Track, in proceedings of the 21st Text Retrieval Conference, NIST Special Publication, 2012.

4. Hu, Qinmin, Huang, X. and Hu, X., A Term Association Approach for Genomics Information Retrieval, in proceedings of the 2011 IEEE International Conference on Bioinformatics & Biomedicine, pages 532-537, 2011.

5. Hu, Qinmin, Huang, X. and Miao, J., Exploring a Multi-Source Fusion Approach for Genomics Information Retrieval, in proceedings of the 2010 IEEE International Conference on Bioinformatics & Biomedicine, pages 669-672, 2010.

6. Hu, Qinmin and Huang, X., Genomics Information Retrieval Using a Bayesian Model for Learning and Re-ranking, in proceedings of the 1st ACM International Conference on Bioinformatics and Computational Biology, pages 426-429, 2010.

7. Huang, X., An, A. and Hu, Qinmin, Medical Search and Classification Tools for Recommendation, in proceedings of the 33rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, page 707, 2010.

8. Hu, Qinmin, Ye, Z. and Huang, X., Enhancing Content-Based Image Retrieval Using Machine Learning Techniques, in proceedings of the 2010 International Conference on Active Media Technology, pages 383-394, 2010.

9. Miao, J., Huang, X. and Hu, Qinmin, York University at TRECVID 2010, in proceedings of the 19th Text Retrieval Conference, NIST Special Publication, 2010.

10. Hu, Qinmin and Huang, X., A Time Series Based Method for Analyzing and Predicting Personalized Medical Data, in proceedings of the 2010 International Conference on Brain Informatics, pages 288-298, 2010.

11. Huang, X. and Hu, Qinmin, A Bayesian Learning Approach to Promoting Diversity in Ranking for Biomedical Information Retrieval, in proceedings of the 32nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pages 307-314, 2009 (15.8% acceptance rate: 78 regular papers accepted out of 494 submissions).

12. Yin, X., Huang, X., Hu, Qinmin and Li, Z., Boosting Biomedical Information Retrieval Performance through Citation Graph: An Empirical Study, in proceedings of the 13th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD'09), pages 949-956, 2009.

13. Ye, Z., Huang, X., Hu, Qinmin and Lin, H., An Integrated Approach for Medical Image Retrieval through Combining Textual and Visual Features, in proceedings of Multilingual Information Access Evaluation II. Multimedia Experiments - 10th Workshop of the Cross-Language Evaluation Forum, CLEF 2009, pages 195-202, 2009.

14. Huang, X., An, A., Hu, Qinmin and Tu, K., Medical Text Analytics Tools for Search and Classification, in proceedings of the 2009 Annual International Conference on Information Technology and Communications in Health (ITCH'09), pages 19-24, 2009.

15. Hu, Qinmin and Huang, X., A Reranking Model for Genomics Aspect Search, in proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pages 783-784, 2008.

16. Hu, Qinmin and Huang, X., A Dynamic Window Based Passage Extraction Algorithm for Genomic Information Retrieval, in proceedings of the 17th International Symposium on Methodologies for Intelligent Systems (ISMIS'08), pages 434-444, 2008.