USING LEARNING TO RANK APPROACH TO PROMOTING DIVERSITY FOR BIOMEDICAL INFORMATION RETRIEVAL WITH WIKIPEDIA

JIAJIN WU

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

MASTER OF ARTS

GRADUATE PROGRAM IN INFORMATION SYSTEMS & TECHNOLOGY
YORK UNIVERSITY
TORONTO, ONTARIO
APRIL 2014

Abstract

In most of the traditional information retrieval (IR) models, the independent relevance assumption is taken, which assumes the relevance of a document is independent of other documents. However, the pitfall of this is the high redundancy and low diversity of retrieval result. This has been seen in many scenarios, especially in biomedical IR, where the information need of one query may refer to different aspects. Promoting diversity in IR takes the relationship between documents into account. Unlike previous studies, we tackle this problem in the learning to rank perspective. The main challenges are how to find salient features for biomedical data and how to integrate dynamic features into the ranking model. To address these challenges, Wikipedia is used to detect topics of documents for generating diversity biased features. A combined model is proposed and studied to learn a diversified ranking result. Experiment results show the proposed method outperforms baseline models.

Acknowledgements

First and foremost, thank God who created me and redeemed me for His economy.

I would give my special thanks to my supervisor Professor Jimmy Huang. He is always available and thoughtful and very concerned with students. He is willing to start a conversation either on research or life with us at any time. He encouraged, directed, and helped me throughout the research and writing of this thesis. I also would like to thank Professor Zijiang Yang, who taught me the course of research method and is my supervisory committee member. Her attitude at work impressed and inspired me. I am also very thankful to my oral examination committee members, without whose suggestions, this thesis would not be like this.

In addition, thanks to my friends at York University, especially Jeff Ye, Jessie Zhao and others in my laboratory and class for their friendship. I would also like to express my thanks and love to my parents for their love and support all the time. At last but not least, thanks to all the brothers and sisters in the church.

Table of Contents

List of Tables

List of Figures

1 Introduction

1.1 Motivation

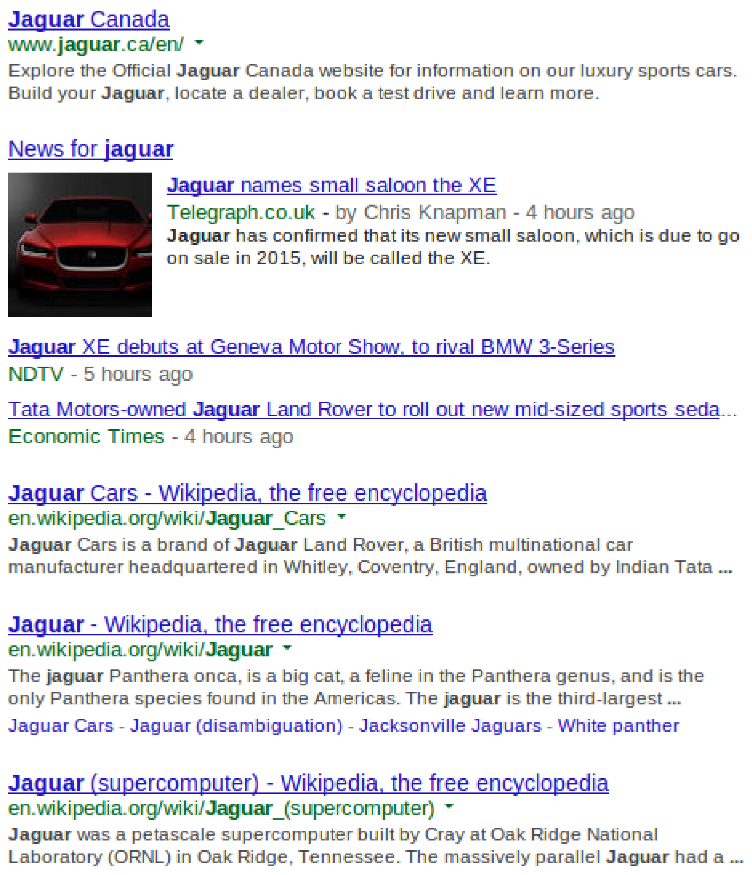
As more and more data and information are made available digitally and on the Internet, the technique of information retrieval (IR) has been developed substantially. It helps indexing and retrieving data for obtaining knowledge and has already played a crucial role in both daily routine and academia in many fields. One of the fundamental concepts of IR is relevance, which refers to how much the document meets the information need of the query. With this basis, the main goal of IR is to determining the relevance between documents and query and presenting the documents in the descending order of relevance.

However, with the extensive usage of IR systems, there is a growing demand of increasing results novelty from the user end and disambiguation of query from the IR system end, so the traditional IR is facing some challenges. In the past decade, to promoting diversity in ranking has emerged as a very hot topic in IR to meet this need. The restriction of traditional IR is that each document is treated independently. One manifestation of this drawback in the preliminary IR systems is that multiple similar documents will be returned on the top of the ranking list. The goal of this research is to explore how IR can move beyond the assumption that the relevance of a document is independent of other documents.

Figure 1.1 shows a typical example of an ambiguous query “Jaguar”. One would wish for diverse results for this query since it is not clear if the user is interested in the animal, the car or another meaning of this query. In this example, the uncertainty comes from the ambiguity of the entity the query refers to. In another example, “swine flu”, the uncertainty comes from the user. Since the doctors and patients who search this query may concern with different aspects of this topic (eg. vaccine or case of swine flu for doctors, and symptoms for patients).

The application of diversity IR has shown beneficial in the scenario of biomedical IR, where biologists tend to query a certain type of entities covering different aspects, such as genes, proteins, diseases, and mutations [31]. The biomedical IR has been studied in TREC[1] for several years. And in 2006 and 2007 Genomics tracks, a new task was proposed focusing on passage retrieval for question answering using full-text documents from the biomedical literature. Systems were required

[1] Text REtrieval Conference is an on-going series of workshops focusing on a list of different IR research areas, or tracks since 1992.



Jaguar Canada
www.jaguar.ca/en/
Explore the Official **Jaguar** Canada website for information on our luxury sports cars. Build your **Jaguar**, locate a dealer, book a test drive and learn more.

News for **jaguar**

Jaguar names small saloon the XE
Telegraph.co.uk - by Chris Knapman - 4 hours ago
Jaguar has confirmed that its new small saloon, which is due to go on sale in 2015, will be called the XE.

Jaguar XE debuts at Geneva Motor Show, to rival BMW 3-Series
NDTV - 5 hours ago

Tata Motors-owned **Jaguar** Land Rover to roll out new mid-sized sports seda...
Economic Times - 4 hours ago

Jaguar Cars - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/**Jaguar**_Cars
Jaguar Cars is a brand of **Jaguar** Land Rover, a British multinational car manufacturer headquartered in Whitley, Coventry, England, owned by Indian Tata ...

Jaguar - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/**Jaguar**
The **jaguar** Panthera onca, is a big cat, a feline in the Panthera genus, and is the only Panthera species found in the Americas. The **jaguar** is the third-largest ...
Jaguar Cars - Jaguar (disambiguation) - Jacksonville Jaguars - White panther

Jaguar (supercomputer) - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/**Jaguar**_(supercomputer)
Jaguar was a petascale supercomputer built by Cray at Oak Ridge National Laboratory (ORNL) in Oak Ridge, Tennessee. The massively parallel **Jaguar** had a ...

Figure 1.1: An explanatory of diverse results given an ambiguous query

to return passages that contain answers to the questions. The task is essentially to return different aspects with best coverage of the query to answer the question. Thus how to promote the diversity of the retrieval result is crucial to this problem.

The objective of this thesis is to propose solutions that make use of machine learning techniques in IR, namely learning-to-rank, to promote diversity for biomedical IR with the help of Wikipedia. Learning-to-rank is a new technique in IR that has been developed in the past decade, which adopts machine learning techniques to advance traditional IR. It is feature based and has the natural advantage of learning an optimized ranking formula from different heterogeneous or homogeneous knowledge (eg. features), whereas it is impossible to integrate different elements into one single ranking model in the traditional way (eg. to integrate probabilistic elements into the formula of language model is impracticable, and vice versa).

1.2 Main Contributions

In this study, a novel IR approach is proposed to address the challenges of biomedical IR such as high redundancy and low diversity in the retrieval ranking lists. Traditional IR algorithms have the relevance independence assumption stating that the relevance of a document is independent with other retrieved documents. However, this restriction leads to the high redundancy because similar documents would be ranked alike. An ideal ranking list should have the top ranked documents as relevant as possible and meanwhile the documents should cover as many different aspects of the query as possible.

Wikipedia is a free online encyclopedia, written collaboratively by a large amount of participants. By March 2014, there are a total number of 32 million pages among which 4 million in English having been created. Each of a Wikipedia page could be considered as an entity. Each page contains multiple links, including wikilink, interwiki link and external web link connections which lead readers to other Wikipedia pages, other Wikipedia projects, and external websites, respectively. The rich knowledge resource together with the semantics meaning implicit in the linkage structure enable Wikipedia the capability of being used as external knowledge for analyzing the content of a given document. In this research, a Wikipedia mining tool, Wikipedia Miner [47], is exploited for detecting the topics of documents.

Learning-to-rank is a type of method that is based on features. In the previous studies of its application to biomedical IR, only traditional IR features are used. These include but are not limited to: the scores of conventional IR models (e.g. BM25, Language Model) and linkage information (e.g. PageRank, HITS). However, none of these are domain-related nor diversity-favored. This limits the performance of the ranking model. In this research, a family of diverse features will be integrated into the ranking model.

On top of feature study of learning-to-rank, there are many paradigms of how to learn an optimized ranking model from these features. For example, three major categories of learning methods namely Pointwise, Pairwise and Listwise exist in the literature. In this research, the method of [46] is selected, which learns a linear model using coordinate ascent. It is a Listwise method and learns model by directly optimizing IR metrics.

Overall, the key unique contributions in this research include (1) using Wikipedia to determine the top k aspects/topics of the retrieved documents with respect to the query, (2) defining diversity features based on topics coverage of individual document and all other documents that are ranked higher in the ranking result, (3) based on two models, one of which emphasizes the overall relevance precision, and the other focuses on promoting the overall topics coverage, proposing a learning framework that integrates the two models to provide diverse results.

1.3 Thesis Organization

This thesis is organized in the following way. Chapter 2 goes through background information and related work in the area of biomedical IR, including conventional IR models, TREC Genomics Tracks, diversity IR and learning to rank. Chapter 3 presents preprocess procedures in addressing the challenges of biomedical IR and describes the method explored in this research step by step. Chapter 4 lists the experimental setting for this research. Chapter 5 provides the experimental results and discussion on how the proposed method work. Finally the conclusion and future work are given in Chapter 6.

2 Background and Related Work

Modern information retrieval (IR) basically involves with indexing and retrieval, and since indexing technique is beyond the study of this research, only retrieval from inverted index is discussed in this thesis. IR has historically focused on document retrieval, but the field has expanded in recent years with the growth of new information needs. The growing amount of scientific discovery in genomics and related biomedical disciplines has led to a corresponding growth in the amount of online data and information. A growing challenge for biomedical researchers is how to access and manage this ever-increasing quantity of information. This situation presents opportunities and challenges for the IR field to propose strategies for retrieving information in biomedical domain.

2.1 Information Retrieval

The modern IR dates back to as early as 1930's, when Goldberg et al. submitted patents of a document search engine using photoelectric cells and pattern recognition to search the metadata on microfilmed documents [18]. The term of IR was coined in 1950s [49]. The key concept of relevance in IR was defined as a measure of the probability that the document will satisfy the information need of a given request [45]. Ever since, most IR systems strive to assign a score to each document for the measurement of relevance and rank documents according to this score.

2.1.1 Vector Space Model

The first most frequently used IR model is vector space model (VSM). Like most of the IR models, VSM uses bag-of-words notation. It utilizes a vocabulary and each term of which could be a word or a phrase. Based on this, each document is represented as a vector of terms [61]. It is not hard to imagine that the space dimension of this vector space is very high. Since each document contains only a limited set of terms, most of the vectors would be very sparse. In VSM, the query is treated as a short document and is represented in a similar way.

In order to assign a relevance score to each document, VSM measures similarity between document vector and query vector. For this end, many similarity measurements have been taken into consideration, such as cosine of the angle formed by the two vectors and inner-product of the two vectors. The most frequently used one is the cosine of the angle because it has the nice property of being 1 for identical vectors and 0 for orthogonal vectors.

2.1.2 Probabilistic Model

Another family of the most famous IR models is the probabilistic model, which is based on the Probabilistic Relevance Framework (PRF). In this framework, the relevance is taken as the degree of a document's meeting the information need that is judged by user [58]. The assumptions of relevance required for this framework are:

- Relevance is assumed to be a property of the document based on given information need only, assessable without reference to other documents; and
- The relevance property is assumed to be binary.

Given that the IR system does not know the relevance property of each document, it is assumed that the information known to the system will suggest the best probabilistic or statistical evidence as to the relevance of the document satisfying the underlying need. In PRF, all the statistical information will be encapsulated in the probability of relevance. And then following the statement of Probability Ranking Principle (PRP), retrieved documents will be ordered descendingly by probability of relevance [39].

The Binary Independence Model (BIM) is one of the derivants of PRF. It assumes that documents are binary vectors, that is, only presence or absence of terms in documents are measured. Moreover, it assumes that terms are independently distributed in the document, and no association between terms is modeled. This allows the representation to be treated as an instance of VSM.

The most successful algorithm in the family of PRF is BM25 [59]. BM25 is an extension of BIM, but instead of taking the independent assumption of within document terms, it introduces a hidden attribute called "eliteness" and assumes the independent relationship between relevance and eliteness. And the frequency of a term (term frequency, a.k.a., TF) is assumed to depend on eliteness. Eliteness is "aboutness" of a document for a term. Those documents that are talking about the concept represented by the term are described as "elite" for the term. 2 poisson models (with different means) are used to model within-document term frequency for elite documents and non-elite documents.

Although BM25 has proven to be effective and are defaulted baselines methods in many applications, recent research shows that there is still room for improving it. For example, BM25 does not take structure information into account. However, it is noticeable that in many types of documents, there are different fields, for example, title, abstract, introduction and method in scientific papers. It is common that text appearing in different fields contribute differently in predicting relevance. BM25F [76] is a variant of BM25 that addresses this issue. In BM25F, parameters of BM25 vary in different streams of text.

Another restriction of BM25 is that only textual features are used in the formula. One stream of methods incorporate BM25 with linkage features for web searching [14; 15].

In BM25, there is a verbosity hypothesis stating that the document length is not correlated to relevance. While, many studies suggest that the document length could have impact on relevance [63; 83]. In [83], density analysis is explored to measure the document length distribution and a length-based BM25 weight model is proposed.

BM25 also assumes the independence of terms in documents. However, studies [16; 19; 64] show that query terms co-occurrences, or proximities, have an impact on relevance. In [28], window-based N-gram counting and survival analysis methods are used to measure proximity. The proximity are then incorporated into BM25 to boost retrieval performance. In [81], a shape function is used to characterize the impact of an occurrence of a query term and any other term in the document. A pseudo term (cross term) is defined out of two query terms, when they are close to each other and their shape functions intersects. Several kernel functions are used as impact shape functions to measure the impact of query terms. A cross term retrieval model is proposed to integrate cross terms and query terms into BM25 and improvement is seen.

2.1.3 Language Model

Language modeling approach was first introduced into IR by Ponte and Croft [54]. The term "language model" refers to a probabilistic model of text, that is, it defines a probability distribution over sequences of words. The method is often referred to as "query likelihood" scoring method. The underlying idea of this method is to first estimate a language model for each document, and then, according to the probability of query generated by each document (i.e., likelihood), the documents are ordered descendingly. A basic language modeling approach formulation is shown in Equation 2.1:

$$p(D|Q) = \frac{p(Q|D)p(D)}{p(Q)} \propto p(Q|D)p(D) \tag{2.1}$$

where Q is a query, D a document, and $p(Q|D)$ the probability that a user who likes document D would pose query Q. Let θ_D be a "language model" estimated based on document D, and $p(Q|D)$ could be interpreted as $p(Q|\theta_D)$. In the query likelihood method, θ_D was defined as a multiple Bernoulli model. The pitfall of this assumption is that the TF is ignored, only presence/absence of the term in the document is accounted. A variant of this is to assume multinomial distribution of terms in document.

One important issue of the language modeling approach is the estimation of θ_D. Imagine an unseen term in D appears in Q, which is quite common due to data sparseness, using maximum likelihood (ML) estimation to calculate probability of document D generating query Q will cause zero probability problem, i.e., the $p(Q|\theta_D)$ will be zero. It is important to solve this by smoothing ML estimate of probabilities. Different smoothing strategies lead to different smoothing methods [11; 37; 43; 77].

$p(D)$ is another factor that will usher into variants of language modeling approaches. It can be used to incorporate additional retrieval criteria, such as page quality in web search, to favor documents with certain features.

2.2 Biomedical Information Retrieval

2.2.1 TREC Genomics Track

TREC (Text REtrieval Conference)[2] is an annual activity of the IR community aiming to evaluate systems and users. It is sponsored by the National Institute for Standards and Technology. IR has historically focused on document retrieval. However, some special interests have expanded in recent years with the growth of new information needs (e.g., question-answering, cross-lingual), data types (e.g., video) and platforms (e.g., the Web). The role of TREC is to make research groups work on a common source of data and a common set of queries or tasks.

TREC activity is organized into tracks of common interest, such as question-answering, multi-lingual IR, Web searching, and interactive retrieval. TREC generally works on an annual basis, with data distributed in the spring, experiments run in the summer, and the results presented at the annual conference which usually takes place in November.

The goal of the TREC Genomics Track[3] is to create test collections for evaluation of IR and related tasks in the genomics domain. The Genomics Track differs from other TREC tracks in that it is focused on retrieval in a specific domain as

[2] http://trec.nist.gov/

[3] http://ir.ohsu.edu/genomics/

opposed to general retrieval tasks, such as web searching or question answering. There are many reasons why a focus on this domain is important. New advances in biotechnologies have changed the face of biomedical research, particularly high-throughput techniques such as gene microarrays. They not only generates massive amounts of data but also have led to an explosion of new scientific knowledge. As a result, this domain is ripe for improved information access and management. The scientific literature plays a key role in the growth of biomedical research data and knowledge. Experiments identify new genes, diseases, and other biomedical processes that require further investigation. Furthermore, the literature itself becomes a source of experiments as researchers turn to it to search for knowledge that drives new hypotheses and research. Thus, there are considerable challenges not only for better IR systems, but also for improvements in related techniques, such as information extraction and text mining.

The Genomic Track started from 2003 and ended at 2007. The ad-hoc task for 2003 focuses on extracting the documents which describe the function of genes. For the second year, the task focuses on extracting the documents according to the queries which simulate the real need from biologists. In the third year, the task puts more energy on how to categorize the queries and provide some different processing for different query categories. In 2005, 32 groups from all over the world submitted 59 runs to the ad-hoc retrieval task.

2.2.2 Genomics Passage Retrieval

In the 2006 Genomics Track, the Genomics passage retrieval was proposed and it is further investigated in the 2007 Genomics Track. As in the previous tracks, there were a large number of participating groups in these two years' tracks.

In 2006 Genomic Track, the University of Illinois at Urbana-Champaign applied language modeling techniques to the passage retrieval [35]. They used a regularized estimation method to improve the pseudo relevance feedback mechanism in the retrieval model in the KL-divergence retrieval framework. They also used a Hidden Markov Model based passage extraction method to determine the length and boundaries of query-dependent relevant passages.

University of Wisconsin focused on query generation and reranking query results to encourage relevance and diversity [25]. They implemented a query generation method using an in-domain syntactic parser to automatically identify noun phrases in the topic descriptions. Given that it is common to have many entity phrases that refer to the same concept, especially in the biomedical setting, they used online resources to expand the queries with synonyms. They tested two different ways for reranking. One is a clustering-based approach, that they re-ranked the passages by picking out one document from each cluster to promote ranking diversity. And the other one is a graph-theoretic algorithm (GRASSHOPPER). GRASSHOPPER is based on absorbing Markov chain random walks. Specifically, a random walk is defined on a graph over the passages. Passages which have been ranked so far become absorbing states. These absorbing states "drag down" the importance of similar unranked states, thus encouraging diversity.

Purdue University extracted acronyms, aliases, and synonyms from external biomedical resources, and weighted and combined them to expand original queries [42]. They used a hierarchical Dirichlet smoothing method for utilizing passage, document, and collection language models in passage retrieval. A post-processing step was performed to combine the scores from passage retrieval and document retrieval. A query term matching-based method was presented to further improve the search performance. An external database constructed from MEDLINE abstracts was used to assign MeSH (Medical Subject Heading)[4] terms to passages for estimating topical aspects. However, their methods achieved worse aspect-level scores than baseline method.

Later in the TREC 2007 Genomics track, 27 groups participated and 66 runs in total were submitted. Most of the teams tried to obtain the aspect level performance

[4] http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=mesh

through their passage level results, instead of working on the aspect level retrieval directly [17; 30; 82].

University of Neuchatel used two different ways to define passage [12]. One way was using HTML tags such as H1, H2, P, BR, TABLE, and TD as passage delimiters. The other way was to define the passage on sentence level. As for the retrieval model, they used single IR models, such as BM25, language model and Divergence from Randomness as well as combination of them. Furthermore, they applied WordNet thesaurus expansions and orthographic variants resulting from that to their system.

University of Illinois at Chicago considered that a query constitutes of two parts, target and qualification [69]. A target refers to any instance of a certain entity type and the qualification refers to the condition that the target has to meet to be accepted as an answer to the query. The relevance of a document to a query is measured by to what degree the document contains a target and satisfies the qualification. Based on this, they further classified the entity into two types. The difference of the two types is that for type I, resources (such as UMLS) can be found, from which some candidate targets could be retrieved, whereas for type II no such resources are available. For each type, they used different strategies for retrieval. They developed a conceptual retrieval model and incorporated five types of domain knowledge in the domain of genomics to that model.

Recently there are some work showing that Wikipedia can be used as an external knowledge resource to facilitate biomedical IR [73; 74]. In these studies, Wikipedia is used as an encyclopedia to help to detect the topics of documents. The novelty of detected topics are measured by binary novelty measurement and survival models for re-ranking to promote diversity of whole ranking list.

2.3 Diversity in IR

One of the most well-known algorithms used for result set diversification in IR is Maximal Marginal Relevance (MMR) [9]. MRR measures query relevance and information novelty independently and combines them linearly as the metric labeled as "marginal relevance". Then, marginal relevance is to be maximized to reduce redundancy while maintaining query relevance in re-ranking retrieved documents and in selecting appropriate passages for text summarization. Experimental results showed that MMR ranking works well in query-relevant multi-document summarization, especially for longer documents which typically contain more inherent passages redundancy across document sections such as abstract, introduction, conclusion, and results, etc.

Similar work in the language modeling framework was studied in [79]. Two ways of measuring the novelty of a document were presented. One is based on the KL-divergence measure, and the other based on a simple mixture model. Then novelty and relevance are combined in a cost function which measures the cost of (1) user seeing a relevant, but redundant document, and (2) user seeing a non-relevant document. The method was shown slightly outperforming a well-tuned relevance ranking baseline.

Zhang et al. proposed Affinity Ranking (AR) to re-rank search results by optimizing two metrics: (1) diversity – which indicates the variance of topics in a group of documents; (2) information richness – which measures the coverage of a single document to its topic [80]. Both of the two metrics are calculated from a directed link graph named Affinity Graph (AG). AG models the structure of a group of documents based on the asymmetric content similarities between each pair of documents. AR score of each document is obtained as a combination of the information richness and diversity penalty scores. AR scores are then used to re-rank the top search results. Experimental results showed improvement of AR algorithm in both diversity and information richness in the top 10 searched results without loss in precision and recall.

Radlinski et al. used "abandonment" to measure user satisfaction, which refers to the event that a user does not click the document [56]. Abandonment indicates that users are presented with search results of no potential interest. Two algorithms were proposed to directly optimize the abandonment rate based on different assumptions. One assumes user interests and documents do not change over time. Thus a greedy strategy is used to iteratively select documents for each rank, and after each document is selected this decision is never revisited. The other assumes that user interests and documents change over time. The Ranked Bandits Algorithm (RBA) is used under this assumption. RBA leverages standard theoretical results for multi-armed bandits (MAB) [4], which is modeled on casino slot machines. The goal of MAB is to gain the maximal total reward by selecting the optimal sequence of slot machines to play.

Markowitz et al. introduced the Modern Portfolio Theory (MPT) [44] to IR for document ranking [68]. IR ranking problem was compared to investment problem in financial market. According to the MPT, they claimed the principle of PRP that ranking documents in order of decreasing probability of relevance is not the optimal option. The reasons are: (1) during retrieval, the relevance of documents are unknown and cannot be estimated with absolute certainty from IR models, (2) the relevance estimates of individual documents are also correlated, either positively or negatively [26]. As a result, the authors proposed to select a top-n ranked list (portfolio) of documents as a whole. Experimental results showed that their approach can adopt to different risk preferences of evaluation metrics, and as a result, significant performance gain was achieved.

Agrawal et al. stated the problem of result diversification and proposed a set function $P(S|q)$ [1]. They supposed that users only consider the top k returned results of a search engine. And their objective is to maximize the probability that the average user finds at least one useful result within the top k results. They pointed out that the objective is NP-hard [23] to optimize, but the set function admits a simple greedy strategy that will solve the problem quite well. Variances of classical IR performance metrics: Normalized Discounted Cumulative Gain (NDCG), Mean Reciprocal Rank (MRR) and Mean Average Precision (MAP), were proposed as intent aware measures: NDCG-IA, MRR-IA and MAP-IA. They were used to take diversification into account. Experimental results compared with commercial search engines in terms of intent aware measures showed the proposed algorithm outperforms the baselines. However, no comparison with traditional performance metrics were provided.

Santos et al. introduced a probabilistic framework xQuAD (eXplicit Query Aspect Diversification) for search result diversification, which explicitly models an ambiguous query as a set of sub-queries [62]. Given an ambiguous query q and an initial ranking R produced for this query, a new ranking S is built by iteratively selecting the τ highest scored documents from R. The scores of documents are given according to a probability mixture model, which is composed of two probabilities modeling relevance and diversity respectively. Experimental results showed xQuAD is effective at diversifying Web search results.

2.4 Learning to Rank

Ranking is the central problem for many tasks in IR related fields, including document retrieval, entity search, question answering, meta-search, personalized search, online advertisement, collaborative filtering, document summarization, and machine translation. The main goal of ranking for IR is to find the criterion for ranking. The traditional criterion is the relevance of retrieved documents with respect to query. The relevance could be affected by many elements, such as TF of query terms appearing in individual document, inverse document frequency (IDF) of query terms appearing in whole documents set, the probability that the document's language model would generate the terms of the query, the authority of the web page containing the document and other web pages linkage information. Traditionally, as discussed in Section 2.1, a parameterized ranking function would be used to determine the relevance. For example, BM25 of probabilistic model, language models, PageRank and HITS are all of this paradigm.

The limitation of traditional IR models is that it is not straightforward to integrate multiple heterogeneous elements into single formula. And the predefined parameters will not work for all situations. Therefore, intensive parameter tunning is usually required. In the past decade, the learning to rank technique has emerged in the field of IR. It adopts machine learning techniques for performing ranking task. Figure 2.1 gives an illustration of how learning to rank works.

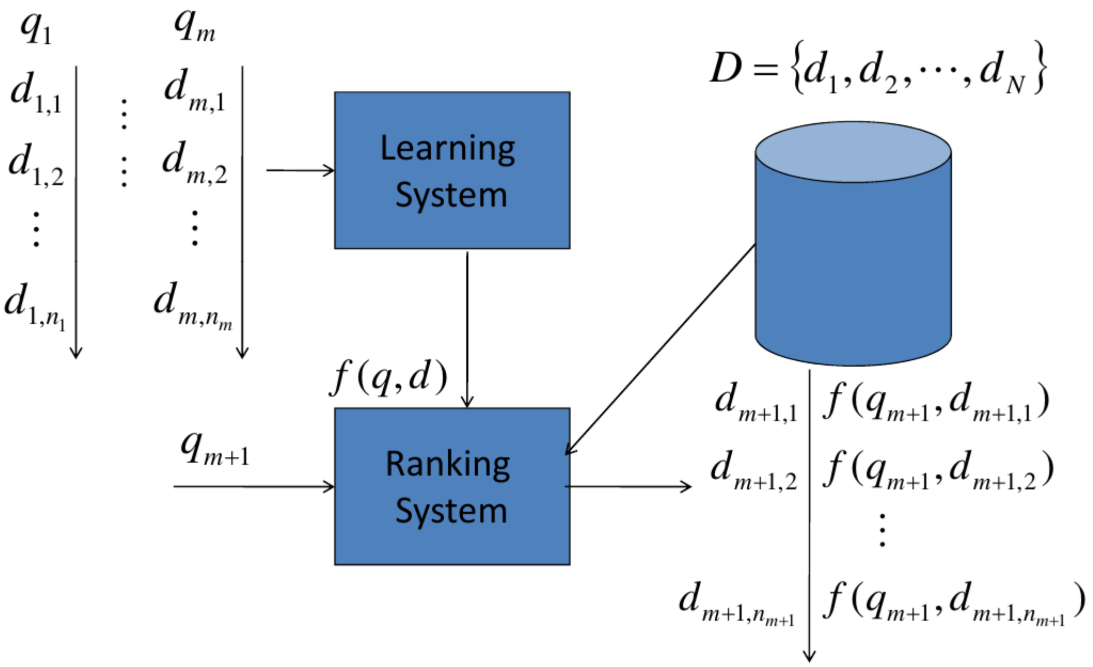

q_1
$d_{1,1}$
$d_{1,2}$
$\vdots$
d_{1,n_1}
$\vdots$
q_m
$d_{m,1}$
$d_{m,2}$
$\vdots$
d_{m,n_m}
$D = \{d_1, d_2, \cdots, d_N\}$
Learning System
$f(q,d)$
q_{m+1}
Ranking System
$d_{m+1,1}$ $f(q_{m+1}, d_{m+1,1})$
$d_{m+1,2}$ $f(q_{m+1}, d_{m+1,2})$
$\vdots$
$d_{m+1,n_{m+1}}$ $f(q_{m+1}, d_{m+1,n_{m+1}})$

Figure 2.1: Learning to Rank Framework

where D is the document set, q_i the ith query from the query set $Q = \{q_1, q_2, ..., q_m\}$, $D_i = \{d_{i,1}, d_{i,2}, ..., d_{i,n_i}\}$ the set of documents associated with query q_i, and $f(q, d)$ a ranking function that can assign a score to a given document pair q and d.

Here are the major characteristics of learning to rank method:

(1) Feature-based: using features defined on the query and the documents as input;

(2) Local ranking model: a local ranking model $f(q, d)$ is utilized;

(3) Supervised learning: the ranking model is usually learned by supervised learning (the machine learning task of inferring a function from labeled training data).

Learning to rank has been intensively studied recently and most of the methods in the literature fall into the following three categories: pointwise approach, pairwise approach and listwise approach. The pointwise and pairwise approaches transform the ranking problem into existing machine learning problems: classification, regression and ordinal classification. The listwise approaches takes ranking lists of objects as instances in learning. Compared with the other two types of methods, the listwise approaches are the real sense of the learning to rank. There are two sorts of listwise approaches: (1) learning by directly optimizing IR performance metrics or their variances, (2) learning by minimizing listwise loss functions (for example, cross entropy [8], or likelihood [70]).

2.4.1 Pointwise

The pointwise approach takes each single document as the input instance, and transforms the ranking problem into classification, regression and ordinal classification problems. Ranking is more about predicting relative order than accurate relevance degree, however, since the group structure of ranking is ignored in the learning process, the relative order between documents will not be naturally reflected in the ranking results. Furthermore, the two intrinsic properties of the IR evaluation measures for ranking (i.e., query-level and position-based) cannot be well considered by the pointwise approach.

Suppose that the learned ranking model $f(x)$ outputs real numbers which will be used to rank documents (sort documents according to the scores given by the model). The loss function used in learning process is pointwise as it is defined on a single object (feature vector).

Nallapati et al. investigated two representative classification models, Maximum Entropy (ME) [27] and Support Vector Machines (SVM) [65; 66], to learn the ranking model [50]. SVM has proven to be one of the best classifiers in many classification tasks performance even when the number of training samples is small.

This is because SVM does not need to use all information of training set, but only the margin on the constraint set of the training data. Also SVM is associated with a nice generalization theory based on the VC dimension [67], and therefore, is theoretically guaranteed to have good performance even if the number of training samples is small. Previous experiments on ad-hoc retrieval indicated that the ME-based algorithm is significantly worse than the baseline language models, but the SVM-based algorithm is comparable with and sometimes slightly better than the language models. Based on this, the author argued that SVM is still preferred because of its ability to learn arbitrary features automatically, to make fewer assumptions, and to be more expressive.

Li et al. proposed the McRank algorithm which uses multi-class classification to solve the ranking problem [40]. The authors were motivated by the fact that the errors in ranking based on discounted cumulative gain (DCG) is bounded by the errors in multi-class classification. The loss function they employed to train the ranking model is the upper bound of the classification error and different upper bounds yield different loss functions; for example, the exponential loss, the hinge loss, and the logistic loss. They studied how to convert classification results to ranking scores. The output class is converted to a probability using a logistic function, which indicates the probability of a document belong to a specific category.

The Gradient Boosting Tree algorithm is used to train the class probabilities.

Crammer et al. used a famous algorithm on ordinal regression, PRanking, to assign a grade to a given object [13]. The goal of PRanking is to find a direction defined by a parameter vector w. After projecting the documents onto the direction, it will be easy to distinguish the documents into different ordered categories by using thresholds. The grades can be used for ranking, and thus their method can also be viewed as a method for ranking. Given training data, Pranking iteratively learns a number of parallel Perceptron models, and each model separates two neighboring grades.

2.4.2 Pairwise

The pairwise approach takes document preference pair as the input instance and transforms ranking problem into pairwise classification and pairwise regression. Although it takes document preference into account, in which sense it is more preferred than pointwise, the ranking structure is still ignored.

Herbrich et al. proposed Ranking SVM which is one of the most well-known learning to rank methods [29]. The basic idea is to treat the ranking problem as pairwise classification and employ SVM technique to perform the learning task. The input instances are created by making document preference pairs according to the difference between documents grades.

Freund et al. proposed RankBoost method based on the Boosting technique [20]. It adopts AdaBoost [21] to perform the classification task over document pairs. The difference between AdaBoost and RankBoost lies in that the former defines distribution on document whereas the latter on document pairs.

Burges et al. proposed RankNet algorithm likewise [7]. RankNet employs Neural Network as ranking model and uses Cross Entropy as loss function. The optimal Neural Network model is then trained by using Gradient Descent algorithm.

2.4.3 Listwise

The listwise approach takes ranking list as input instance in both learning and predicting, such that, the ranking group structure is maintained and the IR performance metrics could be more easily incorporated into the listwise loss functions. Listwise approaches can be further divided into two categories. For the first type, the listwise loss function is defined to measure the difference between the documents permutation given by hypothesis of ranking model and ground truth permutation. For the second type, the loss function is defined based on approximation or bound of IR performance metrics.

Cao et al. pointed out the importance of employing the listwise approach to ranking and proposed ListNet algorithm [8]. ListNet treats lists of documents as input “instances”. They proposed to use the Luce-Plackett model to calculate the permutation probability or top k probability of list of objects. Extended from RankNet, ListNet employs a Neural Network as model, and employs KL divergence as loss function. The permutation probability or top k probability of a list of documents is calculated by the Luce-Plackett model. KL divergence is used to measure the difference between the learned ranking list and the ground truth ranking list using their permutation probability distributions or top k probability distributions. Gradient Descent is used as optimization algorithm.

Yue et al. proposed SVM^{map} which uses structured SVM to globally optimize a hinge-loss relaxation of the IR performance metric MAP [75]. Their algorithm is computationally efficient in finding a globally optimal solution. This idea of were extended to optimize other IR evaluation measures NDCG and MRR [10; 52]. The difference lies in the feature mapping and strategy for searching the optimized model.

3 Learning to Rank for Biomedical Information Retrieval

3.1 Preprocess

Conducting conventional IR experiments requires several key components including: preprocessing, indexing, retrieval and performance measurement. Different from conventional IR, learning to rank is feature-based method and is consist of training and testing processes given a feature-based dataset. However, in this research, we will start with creating the dataset for training and testing for learning to rank model from the conventional IR, so the first three components in conventional IR are still needed for constructing the dataset.

3.1.1 Text Processing

The data used in this research is from the TREC Genomics Track 2006 and 2007 datasets. The raw data comes from full-text HTML biomedical journal papers. The task is to retrieve passages (from part of paragraph) from the data to answer the structured questions from real biologists.

HTML Parsing

The first step for processing the data is to partition the raw HTML papers into paragraphs according to the HTML $< p >$ or $< /p >$ tags. Each paragraph will be identified with their document ID, offset and length. The next step is to convert the HTML to human friendly readable plain text. This is done by removing all the HTML tags.

Stop Words Removal

Stop words in the field of IR refer to those words that contribute little or no to the relevance of documents and can be filtered out. There is no one definite list of stop words which is used by all IR systems. For some of the IR systems, these are the most common, short function words, such as the, is, at, which and a. In this research, a stop word list provided by an open source IR system is adopted[5].

Stemming

[5] http://ir.dcs.gla.ac.uk/resources/linguistic_utils/stop_words

Stemming is to recognize variants of the same word and convert all of them to the stem. It helps reducing the number of indexed terms. Porter stemmer [55] is widely used in IR community and is adopted in this research as well.

TREC Formating

The IR system usually accepts certain data format for indexing. Figure 3.1 shows the TREC format used in this research. The processed plain text is converted into this format and is used for indexing.

Figure 3.1: TREC format

```
< DOC >
< DOCNO > document_number < /DOCNO >
< TEXT >
Index this document text.
< /TEXT >
< /DOC >
```

Where the $< DOC >$ and $< /DOC >$ identify the boundary of the indexed unit, which is a passage in this research. $< DOCNO >$ field is the identification of the index unit which consists of document ID, passage offset and passage length. $< TEXT >$ field is the indexed content which is the processed clean stemmed text.

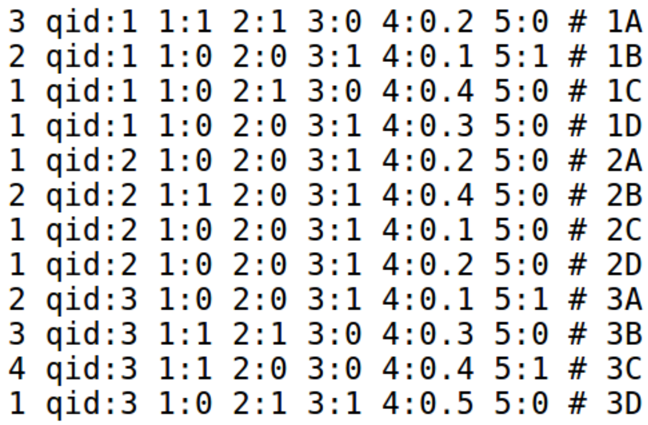
Indexing and Retrieval

In this research, an in-house IR platform is used for indexing and retrieval.

Various ranking models are implemented in the platform and typical IR models were used for retrieval in this research. Passages returned by multiple IR models are selected for training data and represented as feature vector, each of which are scores assigned by different IR models.

3.1.2 Training and Testing

Learning to rank technique is comprised of training and testing processes, as a supervised machine learning task. The data used in learning to rank is similar to, but different from, the data in conventional supervised learning tasks such as classification and regression. The training data contains queries and documents. Each query is associated with a number of documents and they form a group. The groups are independent and identically distributed (i.i.d.) data, while the instances within a group are not i.i.d.. The relevance of the documents with respect to the query is also given. The relevance information can be given in several ways. Here, we take the most widely used approach, and we assume that the relevance of a document with respect to a query is represented by a label. The labels are at several grades (levels). The higher grade a document has, the more relevant the document is. Figure 3.2 shows an example of the training dataset of learning to rank. Where it is composed of 3 queries, and each of them has 4 associated documents, and the



```
3 qid:1 1:1 2:1 3:0 4:0.2 5:0 # 1A
2 qid:1 1:0 2:0 3:1 4:0.1 5:1 # 1B
1 qid:1 1:0 2:1 3:0 4:0.4 5:0 # 1C
1 qid:1 1:0 2:0 3:1 4:0.3 5:0 # 1D
1 qid:2 1:0 2:0 3:1 4:0.2 5:0 # 2A
2 qid:2 1:1 2:0 3:1 4:0.4 5:0 # 2B
1 qid:2 1:0 2:0 3:1 4:0.1 5:0 # 2C
1 qid:2 1:0 2:0 3:1 4:0.2 5:0 # 2D
2 qid:3 1:0 2:0 3:1 4:0.1 5:1 # 3A
3 qid:3 1:1 2:1 3:0 4:0.3 5:0 # 3B
4 qid:3 1:1 2:0 3:0 4:0.4 5:1 # 3C
1 qid:3 1:0 2:1 3:1 4:0.5 5:0 # 3D
```

Figure 3.2: Training Data Sample

labels range from 1 to 4 representing different levels of relevance.

A (local) ranking model is a function of query and document, or equivalently, a function of feature vector derived from query and document. And this is usually gained by training from the dataset using supervised learning method as has been done in this thesis.

In the testing process, new queries and associated set of documents are created. Feature vectors with the same composition as the training data will be generated and scores to the documents will be assigned using the trained ranking model.

3.1.3 Training Data Creation

As a supervised learning task, how to create high quality training data is crucial important to learning to rank. Ideally, the training data should consist of the perfect ranking lists of documents for each query. Currently, there are two common ways to create training data. The first one is human labeling, which is widely used in the IR community. First, a set of queries is randomly selected from the query log of a search system. Suppose that there are multiple search systems. Then the queries are submitted to the search systems, and all the top ranked documents are collected. As a result, each query is associated with documents from multiple search systems (it is called the pooling strategy). Human judges are then asked to make relevance judgments on all the query document pairs. Relevance judgments are usually conducted at five levels, for example, perfect, excellent, good, fair, and bad. Human judges make relevance judgments from the viewpoint of average users.

The other way of generating training data is derivation from click through data. Click-through data at a web search engine records clicks on documents by users after they submit queries. Click-through data represents implicit feedbacks on relevance from users and thus is useful for relevance judgments. One method is to use the differences between numbers of clicks on documents to derive preferences (relative relevance) on document pairs [38].

In this research, we use the dataset given by TREC Genomics Track and the relevance judgment published by NIST based on the pooling strategy. The problem of directly applying the data for learning to rank is that only a limited number of passages' judgments are available. While the data is given as raw, and different strategies of scoping the passages (a span of document) out of documents lead to different spans of passages from the "official passages". So there would be a lot mismatch. We develop a algorithm to generate the labels that will be used for learning to rank using the TREC data.

3.1.4 Feature Construction

The ranking model of learning to rank is feature based, for example the ranking model $f(q,d)$ is in fact defined as $f(x)$ where x is a feature vector based on q and d. This enables the ranking model good generalization ability. Specifically, only a small number of queries and their associated documents are needed for the model training, but any other queries and their associated documents could be applicable to predicting. As in other machine learning tasks, the performance of learning highly depends on the effectiveness of the features used. How to define useful features thus is very important.

In traditional IR, unsupervised ranking models (eg. BM25 and PageRank) are widely used. The definitions of BM25 and PageRank are given as follows.

BM25 is a probabilistic model representing the relevance of document d to query q [59]. It looks at the matching degree between the query terms and document terms and utilizes the numbers of occurrence of query terms in the document to represent relevance. Specifically, BM25 of query q and document d is calculated as:

$$BM25(q,d) = \sum_{\omega \in q \cap d} idf(\omega) \frac{(k+1)tf(\omega)}{tf(\omega) + k((1-b) + b\frac{dl}{avgdl})} \tag{3.1}$$

where ω denotes a word in d and q, $tf(\omega)$ the frequency of ω in d, $idf(\omega)$ the inverse document frequency of ω, dl the length of d, $avgdl$ the average document length, and k and b are parameters.

PageRank represents the importance of web page [53]. It views the web as a directed graph in which pages are vertices and hyperlinks are directed edges. It defines a Markov process on the web graph, and views the stationary distribution (PageRank) of the Markov process as scores of page importance. PageRank of web page d is defined as $P(d)$ in equation 3.2:

$$P(d) = \alpha \sum_{d_i \in M(d)} \frac{P(d_i)}{L(d_i)} + (1-\alpha)\frac{1}{N} \tag{3.2}$$

where $P(d)$ is the probability of visiting page d, $P(d_i)$ the probability of visiting page d_i , $M(d)$ the set of pages linked to d, $L(d_i)$ the number of outlinks from d_i , N the total number of nodes on the graph, and α a weight.

In web search, all of the ranking models could be viewed as features. One of the early practice in web search was to define the ranking model as a linear combination of features. However, when more and more features have been developed, it is no longer straightforward to manually combine many features in one single model because the parameters tuning would become tedious and time-consuming. Thus a more general and principled learning approach is needed for constructing ranking model.

3.2 Aspect Detection

In this section, three methods of aspect detection will be introduced, and based on the application of this research, the selection of the method will be discussed.

3.2.1 Topic Model

Topic models are based upon the idea that documents are mixtures of topics, where a topic is a probability distribution over words [6; 33]. It provides a simple way to analyze large volumes of unlabeled text. A "topic" consists of a cluster of words that occur most frequently in the cluster of documents. Using contextual clues, topic models can connect words with similar meanings and distinguish between uses of words with multiple meanings.

Latent Dirichlet Allocation (LDA) is an example of a topic model and was first presented as a graphical model for topic discovery [6]. LDA is a generative model that allows sets of observations to be explained by unobserved groups that explain why some parts of the data are similar. In LDA, each document may be viewed as a mixture of various topics. This is similar to probabilistic latent semantic analysis, except that in LDA the topic distribution is assumed to have a Dirichlet prior. In practice, this results in more reasonable mixtures of topics in a document. Figure 3.3 shows a probabilistic graphical representation of LDA model.

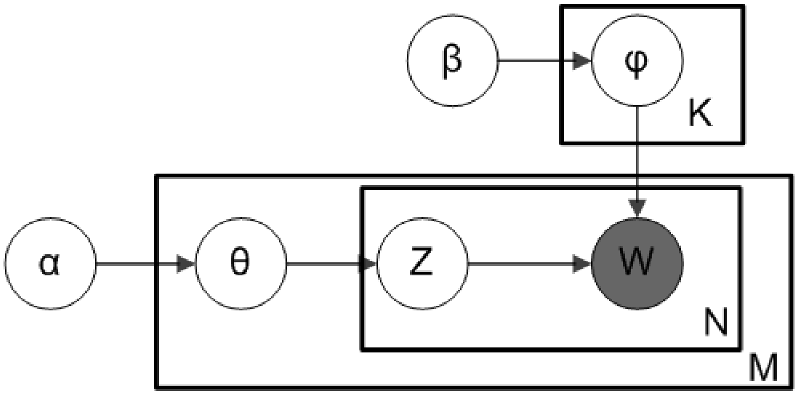


Figure 3.3: A Probabilistic Graphical Representation of LDA Model

The outcome of LDA model conducted on a set of documents is topic-words distribution for input documents. The gensim[6] package could be used for conducting topic modeling.

3.2.2 Clustering

The problem of clustering has been studied widely in the database and statistics literature in the context of a wide variety of data mining tasks [36]. The clustering problem is defined to be that of finding groups of similar objects in the data. The similarity between the objects is measured with the use of a similarity function.

Traditional methods for clustering have generally focused on the case of quantitative data, in which the attributes of the data are numeric [51]. The problem has also been studied for the case of categorical data [24], in which the attributes may take on nominal values.

A text document can be represented in the form of binary data, when we use the presence or absence of a word in the document in order to create a binary vector. In such a case, it is possible to directly use a variety of categorical data clustering algorithms [24] on the binary representation. A more enhanced representation would include refined weighting methods based on the frequencies of the individual words in the document as well as frequencies of words in the entire collection (e.g.,

[6] http://radimrehurek.com/gensim/

TF-IDF weighting [60]).

Based on the natures of generated clusters and techniques and theories behind them, clustering algorithms could be categorized into the following types: Distance and Similarity Measures, Hierarchical, Squared Error-Based, Estimation via Mixture Densities and Graph Theory-Based etc. For more details, please refer to [71].

K-means is the most important flat clustering algorithm. Its objective is to minimize the average squared Euclidean distance of documents from their cluster centers where a cluster center is defined as the mean or centroid $\vec{\mu}$ of the documents in a cluster ω :

$$\vec{\mu}(\omega) = \frac{1}{|\omega|} \sum_{\vec{x} \in \omega} \vec{x} \tag{3.3}$$

The definition assumes that documents are represented as length-normalized vectors in a real-valued space. The ideal cluster in K-means is a sphere with the centroid as its center of gravity. Ideally, the clusters should not overlap. Algorithm 3.1 shows the flow of K-means.

3.2.3 Semantical Analysis with Wikipedia

Wikipedia is a free online encyclopedia edited collaboratively by large numbers of volunteers. The exponential growth and the reliability of Wikipedia make it a potentially valuable knowledge resource. How to utilize Wikipedia to facilitate IR

Algorithm 3.1 K-Means

```
K-MEANS($\{\vec{x}_1, ..., \vec{x}_N\}, K$)
$(\vec{s}_1, \vec{s}_2, ..., \vec{s}_K) \leftarrow SelectRandomSeeds(\{\vec{x}_1, ..., \vec{x}_N\}, K)$
for $k \leftarrow 1$ to $K$ do
        $\vec{\mu}_k \leftarrow \vec{s}_k$
while stopping criterion has not been met do
        for $k \leftarrow 1$ to $K$ do
                $\omega_k \leftarrow \{\}$
                for $n \leftarrow 1$ to $N$ do
                        $j \leftarrow argmin_{j'} |\vec{\mu}_{j'} - \vec{x}_n|$
                        $\omega_j \leftarrow \omega_j \cup \{\vec{x}_n\} (reassignment\ of\ vectors)$
                for $k \leftarrow 1$ to $K$ do
                        $\vec{\mu}_k \leftarrow \frac{1}{|\omega_k|} \sum_{\vec{x} \in \omega_k} \vec{x} (recomputation\ of\ centroids)$
return $\{\vec{\mu}_1, ..., \vec{\mu}_K\}$
```

has become a hot research topic over the last few years [22; 48; 72]. It has also been shown a good candidate as an external knowledge resource for facilitating biomedical IR [74].

The advantage of Wikipedia is that it not only provides concepts (entities) and lexical variants of a specific term, but also provides abundant contexts. With the help of enriched entity pages, it is possible to identify which concepts and lexical variants are related under a specific context. As Wikipedia articles are constantly being updated and new entries are created everyday [72], we can expect that Wikipedia covers the great majority of medical terms.

Another reason of using Wikipedia is that it contains plenty of linkage information among semantic related entities, which can be seen in the explanatory of figure 3.4. Each link in Wikipedia is associated with an anchor text, which can be regarded as a descriptor of its target article. Anchor texts provide alternative names, morphological variations and related phrases for the target articles. Anchors also encode polysemy, because the same anchor may link to different articles depending on the context in which it is found [34].

There are three steps involved in using Wikipedia for aspect detection:

(1) identifying candidate phrases in the given retrieved document;

(2) mapping them to Wikipedia articles;

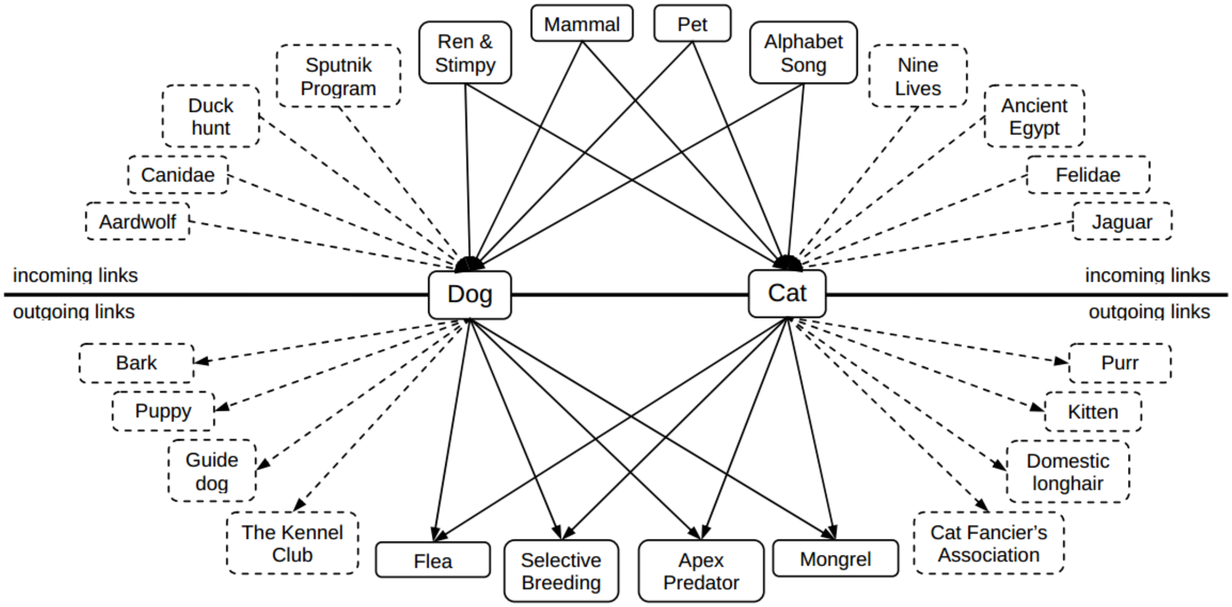


Figure 3.4: A Semantic Relatedness Between Dog and Cat in Wikipedia

(3) selecting the most salient concepts.

The outcome is a set of concepts representing the aspects mentioned in the input documents. The Wikipedia Miner[7] could be used to automatically detect aspects covered by retrieved documents.

[7] http://wikipedia- miner.cms.waikato.ac.nz/

3.2.4 Discussion

In the precious sections, the candidates for detecting topics of documents were introduced. In this research, Wikipedia semantical analysis is adopted for this purpose for the following reasons:

(1) It is the largest online encyclopedia and contains more than 4.3 million entries and includes every domain of human knowledge nearly, and certainly including biomedical or genomics related domains.

(2) The content of Wikipedia is edited by large volumes of volunteers and any inaccuracy or conflicts of editing is shown public, and thus will be revised accordingly. This ensures the acceptable accuracy of entries and relationship between them.

(3) With the extensive coverage of human knowledge, the relationship between entries are existed within. This could be served to explore the semantical connection.

3.3 Diversity Learning to Rank Framework

We propose a learning to rank framework that utilizes both the common features of biomedical text, and the diversity information. More specifically, the novelty and freshness of retrieved results, as well as relevance, will be taken into account. The proposed framework consists of a general ranking model and a diversity-biased ranking model. More specifically, the general ranking model is learned from the training instances represented by the traditional learning to rank features that are common to ad-hoc IR tasks. The diversity-biased model is learned from both general features and diversity-biased features proposed in this research. The final learning to rank model (LTR) is combined linearly as shown in Equation 3.4:

$$LTR(d, Q) = \alpha \cdot gLTR(d, Q) + \beta \cdot dLTR(d, Q) \tag{3.4}$$

where $gLTR(d, Q)$ is the general learning to rank model, $dLTR(d, Q)$ the diversity-biased model, and α and β the parameters that control the weight of two parts and they have the relationship of $\beta = 1 - \alpha$.

To deploy our proposed learning to rank framework in practice, firstly a general ranking model is learned from a set of training queries with their associated relevance assessments. Next for the first pass retrieval results obtained from the general ranking model, we use Wikipedia Miner to extract the related topics of retrieved passages. From this ranking list and the topics information, we generate the diversity-biased features (as shown in Table 3.2) for each query-passage pair. Then the diversity-biased learning to rank model is learned based on all these features.

3.4 General Learning to Rank Model

3.4.1 General Features Extraction

Learning to rank has shown advantage in incorporating various evidences to learn an unified ranking model for enhancing IR [41]. Typical features that will be utilized for constructing a learning to rank model can be categorized as content-based and non-contend-based (e.g. linkage information) features. In this research, due to the following two reasons, only the content-based features are extracted and used for learning a model: (1) the data is from scientific papers, so there is limited linkage structure information that could be extracted, (2) the retrieval task is focusing on using the content to answer the structured questions form biologists, the limited linkage information will contribute little or not to the final goal. The extracted features are summarized in Table 3.1.

Where TF, IDF and query term proximity are the foundamental features used as basis of retrieval models. Term frequency - the number of times a term occurs within a document. Inverse document frequency - inverse of the proportion of documents that contain a given term. Term Proximity - occurrence patterns of terms within a document. The other features are actually widely used coventional ranking models whose ranking functions are defined based on the combination of

Table 3.1: Features for General Learning to Rank Model

Feature	Description
TF-IDF	Term frequency - inverse document frequency.
BM25	Okapi BM25 model [57].
DFR_BM25	The DFR version of BM25 [3].
InL2	An algorithm derived from the divergence from randomness (DFR) framework [3].
DLH13	An DLH hyper-geometric DFR model (parameter free) [3].
DirKL	KL-divergence language model with Dirichlet smoothing [78].
Hiemstra_LM	Hiemstra's language model [32].
ProxQT	Proximity of Query Terms: Intuitively, the more close the query terms occur in a document, the more likely the document would be relevant [64].

the primitive textual features. And these are the state-of-the-art IR models, which are usually used as strong baselines in previous studies. Algorithm 3.2 shows how the features are generated in details.

Algorithm 3.2 The General Features Generating Algorithm

```
Input:
    Q:query set
    D:raw Genomics track dataset
    R:raw relevance judgment for official defined passages(ODP)
    G:aspect judgment for ODPs
    L:maximum legal span for potential passages
Output:
    TR: train dataset with general features, relevance and aspect information
    TE: test dataset with general features, relevance and aspect information
for $q \in Q$ do
        split passages according to L
        using different IR models to get result lists
        generate train features for each passage with each feature a score given by IR models
        for generate relevance score for train dataset for each passage $\in$ R do
                if the passage $\in$ ODP set or has an overlap with some ODP
                the relevance score and aspects information is contributed to the passage
        for generate relevance score for test dataset for each passage retrieved by all the IR models do
                if the passage $\in$ ODP set or has an overlap with some ODP
                the relevance score and aspects information is contributed to the passage
```

3.4.2 Learning to Rank Algorithm

Many learning to rank approaches have been proposed in the literature that could be applied for learning the general ranking model [41]. Among these approaches, we choose to use the Coordinate Ascent algorithm proposed in [46], which has proven to be highly effective for a small number of parameters [5]. Coordinate ascent is a listwise learning method. As discussed in Chapter 2, listwise is more "advanced" than the other two types of learning to rank mothods as it takes the ranking structure of ranking list into account. Coordinate ascent directly optimizes the parameters in the interest of maximizing retrieval metric and it has good empirically verified generalization properties. The ranking function could be obtained by solving the statement shown in Equation 3.5:

$$\begin{aligned} \hat{\Lambda} \quad &= \arg\max\nolimits_{\Lambda} E(\mathcal{R}_{\Lambda}; \mathcal{T}) \\ s.t. \qquad & \mathcal{R}_{\Lambda} \sim S_{\Lambda}(d; q) \\ & \Lambda \in M_{\Lambda} \end{aligned} \tag{3.5}$$

where $S_{\Lambda}(d; q)$ is a scoring function parameterized by a vector of parameters Λ, and it is computed for each query q with each document d in documents set $\mathcal{D}$ ($d \in \mathcal{D}$); $E(\mathcal{R}_{\Lambda}; \mathcal{T})$ is an evaluation matrix, $\mathcal{R}_{\Lambda} \sim S_{\Lambda}(d; q)$ denotes that the orderings in $\mathcal{R}_{\Lambda}$ are induced using scoring function S, and M_{Λ} is the parameter space over Λ.

The optimization is conducted by coordinate ascent, which is a commonly used optimization technique for unconstrained optimization problems. Coordinate ascent iteratively optimizes a multivariate objective function by solving a series of one dimensional searches. It repeatedly cycles through each parameter, holding all other parameters fixed, and optimizes over the free parameter.

3.5 Diversity-Biased Learning to Rank

3.5.1 Diversity Features

We consider the task of promoting diversity as such a scenario that a user would prefer a ranking list of passages so that the top returned passages should be as relevant as possible and meanwhile the passages should cover as many different aspects as possible. Therefore when generating the ranking list, the aspects difference between passages should be taken into consideration to ensure good coverage of different aspects and low redundancy. In such a guildline, we propose the diversity-biased features as shown in Table 3.2.

Table 3.2: Diversity Features for Diversity-Biased Learning to Rank Model

Feature	Description
#RelAsp	Number of relevant aspects the passage contains.
#NonRelAsp	Number of irrelevant aspects the passage contains.
#NewRelAsp	Number of new relevant aspects the passage contains compared with afore ranked passages.
#OldRelAsp	Number of relevant aspects that already existed in afore ranked passages.
NewAspPsg	Ratio of passages that contains new aspects with all afore ranked passages.
%RelAsp	Ratio of number of relevant aspects with allaspects before current rank position.
%UniqRelAsp	Ratio of unique relevant aspects with all aspects before current rank position.

3.5.2 Features Extraction and Model Strategy

Our assumption is that there is a perfect diversified ranking list. Through learning from the general features, which represent the value of each individual query-passage pair, and diversified features, which characterize the novelty and diversity of the whole ranking list, an oracle ranking model could be obtained for further predicting ranking for new dataset.

As can be seen in the previous section, the diversity features aim to reflect the relationship between current document with former ranked documents and therefore the features extraction is related to certain documents ranking and their quality are potentially affected by the ranking list. Actually this simulates the process of generating diversified documents based on former ranked documents in the paradigm of re-ranking for promoting novelty and diversity, where the document for each position is determined in the principle of maximizing the diversity for the whole ranking list. Accordingly these diversity features should be extracted in tandem. We point out that there are different ways to generate diversity features:

- **Once for all**: The diversity features are generated according to the initial ranking given by general learning to rank model, and the oracle model is learned from all features once for all.

- **Dynamic update**: After the diversity features of documents in ith top K subset are determined, the oracle learning to rank model will be re-learned and consequently the general ranking will be updated which results in the re-generating of diversity features.

Heuristically the second strategy might be better; however, we argue that this is much time-consuming and complicated in practice. Therefore in this research, for simplicity, we adopt the first strategy for diversity feature generation.

3.6 Summary

In this chapter, the system used for learning to rank for biomedical IR was presented from the beginning to the end. Firstly the construction of supervised learning to rank dataset was introduced, including the conventional IR process, e.g. text processing, indexing and retrieval as well as labeled training data creation and feature construction, which are unique to learning to rank method. Three types of method for detecting the aspect of retrieved passage was discussed and Wikipedia semantic analysis was selected for this research. A diversity based learning to rank framework was proposed and the general learning to rank method and diversity-biased learning to rank method were given in details. Several diversity-based features were proposed.

4 Experimental Setting

4.1 Data Sets

4.1.1 TREC Data Sets

In order to evaluate the proposed approach, we use the TREC 2006 and 2007 Genomics tracks full-text collection as the test corpus.

It comes from a new full-text biomedical corpus. Permission were btained from a number of publishers who use Highwire Press (www.highwire.org) for electronic distribution of their journals. The document collection is derived from 49 journals and were obtained by a Web crawl of the Highwire site, with post-processing to eliminate as much non-article material as it could be. The full collection contains 162,259 documents. The collection is about 12.3 GB when uncompressed. There are 64 queries in total associated with the collection. Three levels of retrieval metrics were measured in the TREC 2006, namely Passage MAP, Aspect MAP and Document MAP, beyond which 2007 Genomics track utilized an variant called Passage2 MAP.

Golden standard of relevance and aspects judgment for official released legal span of passages are provided. For the sake of generalization, we only utilize the relevance information for generalizing train file for general learning to rank model. We define passage as maximum span of consecutive text within one single document not including any HTML paragraph tag. In this principle we extract passages from the meta data and index. In constructing the train dataset for learning to rank, we compare the extracted passages with the TREC official defined passages that have golden standard of relevance, and assume that whenever there is an overlap, the relevance of official defined passages span will contribute to extracted passage.

Parameters of learning to rank algorithm is optimized using a greedy boosting method on 2-fold cross-validation setting in which the best model is selected according to Document MAP. The parameters α and β in Equation 3.4 are tuned based on 2-fold cross-validation. We also study the effect of parameter setting of α in this thesis.

4.2 Performance Measurement

4.2.1 MAP

Mean average precision (MAP) is a widely used measure in IR. In MAP, binary-notation of relevance is used, i.e., it is assumed that the grades of relevance are at two levels: 1 and 0, representing relevant and irrelevant respectively. Given query q_i , associated documents D_i, ranking list π_i on D_i, and labels y$_i$ of D_i, Average Precision for q_i is defined:

$$AP = \frac{\sum_{j=1}^{n_i} P(j) \cdot y_{i,j}}{\sum_{j=1}^{n_i} y_{i,j}} \tag{4.1}$$

where $y_{i,j}$, is the label (grade) of document $d_{i,j}$ and takes on 1 or 0 as value, $P(j)$ for query q_i is defined:

$$P(j) = \frac{\sum_{k:\pi_i(k) \leq \pi_i(j)} y_{i,k}}{\pi_i(j)} \tag{4.2}$$

where $\pi_i(j)$ is the position of $d_{i,j}$ in π_i. $P(j)$ represents the precision until the position of $d_{i,j}$ for q_i. Note that labels are either 1 or 0, and thus precision (i.e., ratio of label 1) can be defined. Average Precision represents averaged precision over all the positions of documents with label 1 for query q_i.

Average Precision values are further averaged over queries to become MAP.

4.2.2 Passage-Level MAP

This measure uses a variation of MAP, computing individual precision scores for passages based on character-level precision, using a variant of a similar approach used for the TREC 2004 HARD Track [2]. For each nominated passage, the number of characters that overlapped with those deemed relevant by the judges in the gold standard is determined. For each relevant retrieved passage, precision was 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, remaining relevant passages that were not retrieved (no overlap with any nominated passages) were added into the calculation as well, with precision set to 0 for these relevant non-retrieved gold standard passages. Then the mean of these average precisions over all topics was calculated to compute the MAP for passages. Note that this measure is essentially the fraction of retrieved characters that are part of an answer to the topic question.

4.2.3 Passage2 MAP

The original passage retrieval measure for the 2006 track was found to be problematic in that non-content manipulations of passages had substantial effects on

Passage MAP, with one group claiming that breaking passages in half with no other changes doubled their (otherwise low) score. To this end, an alternative measure (Passage2 MAP) was defined that calculates MAP as if each character in each passage were a ranked document. In essence, the output of passages is concatenated, with each character being from a relevant passage or not. Passage2 MAP was used as the primary passage retrieval evaluation measure in 2007.

4.2.4 Aspect-Level MAP

Aspect retrieval is measured using the average precision for the aspects of a topic, averaged across all topics. To compute this, the ranked passages were transformed to two types of values, either the aspect(s) of the gold standard passage that the submitted passage overlapped with or the value "not relevant". This result is a ranked list, for each run and each topic, of lists of aspects per passage. Non-relevant passages had empty lists of aspects. Because it is uncertain of the utility for a user of a repeated aspect (e.g., same aspect occurring again further down the list), these are discarded from the output to be analyzed. For the remaining aspects of a topic, the MAP is calculated similar to how it is calculated for documents, with the additional wrinkle that a single passage may have associated with multiple aspects. Therefore the precision for the retrieval of each aspect is computed as the fraction of relevant passages for the retrieved passages up to the current passage under consideration. These fractions at each point of first aspect retrieval are then averaged together to compute the average aspect precision. Relevant passages that do not contribute any new aspects to the aspects retrieved by higher ranked passages are removed from the ranking. Taking the mean over all topics produces the final aspect-based MAP.

4.2.5 Document-Level MAP

For the purposes of this measure, any PMID (PubMed[8] identifier or PubMed unique identifier) that has a passage associated with a topic ID in the set of gold standard passages is considered a relevant document for that topic. All other documents are considered not relevant for that topic. System run outputs are collapsed by PMID document identifier, with the documents appearing in the same order as the first time the corresponding PMID appeared in the nominated passages for that topic. For a given system run, average precision is measured at each point of correct (relevant) recall for a topic. The MAP is the mean of the average precisions across topics.

[8] http://www.ncbi.nlm.nih.gov/pubmed

5 Experimental Results

The ability to justify the effectiveness of the proposed learning to rank framework could be challenging. The benchmark dataset and various submissions from different teams of TREC Genomics Track 2006 and 2007 provide us objects to compare with. We conduct extensive experiments to evaluate the effectiveness of the proposed learning to rank framework. The principle of designing the experiments is by answering the following questions:

(1) Is learning to rank technique appropriate for using in the field of biomedical information retrieval?

(2) Are the diversity features useful in addressing the diversity of ranking list? And how is the effectiveness of the proposed learning to rank framework?

(3) What is the effect of the parameters in the learning to rank framework?

(4) How effective is the proposed learning to rank framework compared with similar methods?

In order to answer question (1), strong baselines of conventional IR models, BM25 and Language Model, are used as benchmark for comparison. The general learning to rank model is also used to train ranking model and testing. Besides effectiveness comparison in terms of three levels of MAP metrics, the efficiency difference is also discussed. The answer of this could be found in Section 5.1.

To answer question (2), the proposed learning to rank framework will be compared with the general learning to rank method as well as other baseline methods. By doing so, it would be clearly demonstrated that whether the proposed framework has improvement over the general learning to rank method and other conventional models. The result of this could be found in Section 5.2.

For question (3), it is important to know the effect of the parameters in the model since different parameters setting might have impact on the result. Usually tuning parameters is a tedious and time-consuming work. Especially when the final result is largely affected by the parameters, it is a must to obtain the optimal parameters for best performance. This will be studied in Section 5.4.

For question (4), a cost-function based re-ranking method which also utilized Wikipedia as external sources is chosen to compare with the proposed learning to rank framework. The result and discussion are presented in Section 5.5.

5.1 Comparison with Baseline

Following the convention in IR experiments, we use BM25 and Language Model (DirKL) as strong baselines in our experiments. We are concerned with three levels of MAP, namely Document MAP, Passage MAP (and Passage2 MAP on 2007 Collection) and Aspect MAP respectively. We are firstly interested in whether the learning to rank technique will benefit the biomedical information retrieval. So we firstly compare the general learning to rank method (Coordinate Ascent, [46]) to BM25 and Language Model. The comparison results on TREC Genomics Trakck 2006 and 2007 are shown in Table 5.1 and Table 5.2.

Table 5.1: General Learning to Rank Performance Comparison with Baselines on 2006 Collection

MAP	Aspect	Passage	Document
BM25	0.1972	0.0362	0.3449
DirKL	0.1591	0.0360	0.3566
gLTR	0.2292	0.0369	0.3547

From the results, it can be seen that although the general learning to rank model (gLTR) is fairly comparable or even slightly less comparable to BM25 and Language Model in terms of document MAP, it has relatively better performance

Table 5.2: General Learning to Rank Performance Comparison with Baselines on 2007 Collection

MAP	Aspect	Passage	Passage2	Document
BM25	0.1622	0.0651	0.0697	0.2402
DirKL	0.1383	0.0693	0.0637	0.2376
gLTR	0.1878	0.0533	0.0706	0.2179

in terms of Aspect MAP, Passage MAP and Passage2 MAP in 2006 and 2007 collections. For example, on 2006 collection, gLTR outperforms DirKL and BM25 in terms of Aspect MAP over 44% and 16% respectively, on 2007 collection, gLTR outperforms DirKL and BM25 in terms of the improved Passage2 MAP over 1% and 10% respectively. This is desirable because the aim of this research is to promote the diversity of ranking results in biomedical IR. And a diversified ranking result shall cover multiple topics in the top ranking. And the Aspect MAP measures the average precision for the aspects of a topic.

A conclusion could be drawn here that the learning to rank is beneficial to be adopted to biomedical field especially for the sake of promoting the Aspect MAP and Passage MAP.

5.2 Effectiveness of the Proposed Learning to Rank Framework

Secondly, it is of interest that whether the proposed framework of learning to rank could be effective as well. The comparison of our proposed method (LTR) with the baselines and general learning to rank method (gLTR) on 2006 and 2007 collections are presented in Table 5.3 and Table 5.4 respectively. The "+" sign and number in parentheses indicate the statistical significant improvements over gLTR using Student's t-test at alpha level of 0.05. Bold font denotes the best performance on different metric of the four methods.

Table 5.3: Performance Comparison with Baselines on 2006 Collection

MAP	Aspect	Passage	Document
BM25	0.1972	0.0362	0.3449
DirKL	0.1591	0.0360	0.3566
gLTR	0.2292	0.0369	0.3547
LTR	**0.2400** (+4.7%)	**0.0416** (+12.7%)	**0.3910** (+10.23%)

As can be seen from Table 5.3 and Table 5.4, when diversity features are utilized for learning a ranking model, performance improvements over three strong baselines

Table 5.4: Performance Comparison with Baselines on 2007 Collection

MAP	Aspect	Passage	Passage2	Document
BM25	0.1622	0.0651	0.0697	0.2402
DirKL	0.1383	0.0693	0.0637	0.2376
gLTR	0.1878	0.0533	0.0706	0.2179
LTR	**0.1923** (+2.4%)	**0.0784** (+47.1%)	**0.0831** (+17.7%)	**0.2721** (+24.9%)

BM25, DirKL and gLTR can be obtained in terms of all different levels of MAP metrics on both 2006 and 2007 collections. For example, the Aspect MAP improvement of LTR against gLTR, DirKL and BM25 on 2006 collection are 4.7%, 51% and 21.7% respectively; the Passage2 MAP improvement of LTR against gLTR, DirKL and BM25 on 2007 collection are 17.7%, 30.1% and 19% respectively. As to the higher improvement space of Passage MAP than Aspect MAP in general, we attribute it to the paragraph-based indexing of the original data and the way how we generate training dataset for learning to rank: the relevance of passages are contributed by all embedded paragraphs that are relevant while referring to different topics of the query.

It is noticeable that the improvements of Document MAP are also remarkable. This shows that the diversity features are beneficial for promoting not only diversity but also general relevance performance. When the diversity information is used for training model, the passages that are both relevant and have various topics will be favored by the ranking model. This is promising in that when being designed properly, the diversity features are beneficial both in improving general IR metrics and promoting diversity in ranking.

5.3 Comparison with TREC results

We also compare gLTR and LTR with the TREC submission results in Table 5.5 and Table 5.6 respectively.

Table 5.5: Performance Comparison with TREC 2006 Submissions

MAP	Aspect	Passage	Document
Max	**0.4411**	**0.1486**	**0.5439**
Min	0.011	0.0007	0.0198
Median	0.1581	0.0345	0.3083
gLTR	0.2292	0.0369	0.3547
LTR	***0.2400***	***0.0416***	***0.3910***

The italic bold font in Table 5.5 and Table 5.6 denotes the second best result in each matrix. Normally it is not fair to compare with the best TREC result because

Table 5.6: Performance Comparison with TREC 2007 Submissions

MAP	Aspect	Passage	Passage2	Document
Max	**0.2631**	**0.0976**	**0.1148**	**0.3286**
Min	0.0197	0.0029	0.0008	0.0329
Median	0.1311	0.0565	0.0377	0.1897
gLTR	0.1878	0.0533	0.0706	0.2179
LTR	***0.1923***	***0.0784***	***0.0831***	***0.2721***

the submission could comprehensively use many resources, but the median result shows the average level of all submissions. So the outperforming median results at least shows our model is promising.

5.4 Effect of Control Parameter

In this section, we evaluate the parameters α and β in the framework that can affect the retrieval performance. Because $\beta = 1 - \alpha$, so in this section, we present the results under different settings of α, more specifically we sweep over values (0.1, 0.2, ..., 0.9).

In particular, for each dataset we conduct a 2-fold cross validation, where each fold randomly chooses half of the topics for training and the remaining for testing, and vice versa. The overall retrieval performance is averaged over the two test topic sets.

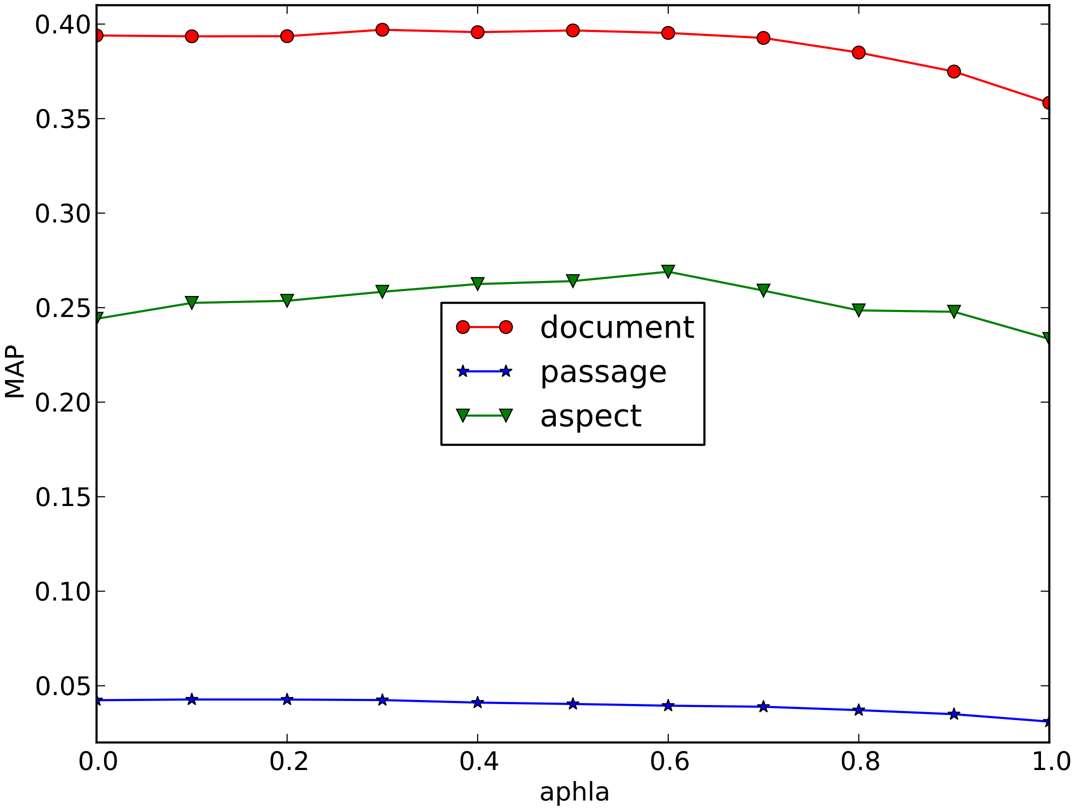


Figure 5.1: Parameter α Against the Retrieval Performance on 2006 Collection

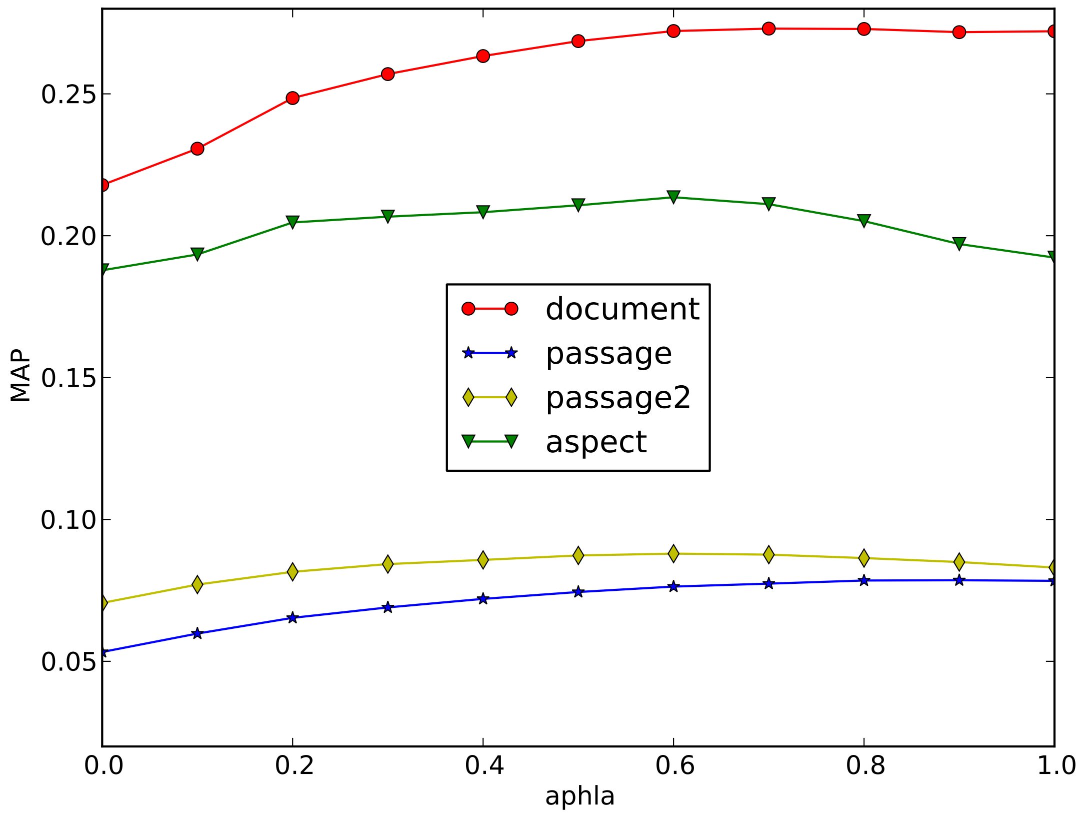


Figure 5.2: Parameter α Against the Retrieval Performance on 2007 Collection

It can be known from Figure 5.1 and Figure 5.2 that the retrieval performance on both 2006 and 2007 data collections are relatively stable under different settings of parameter α, which has significance in practice because the combined model will not be largely affected by different parameter settings and could be free from parameter tuning.

It is also noticed that when α is set to 1, the combined model in Equation 3.4 is equal to gLTR, which is the general model, while it is set to 0, the combined model equals to the diversity-biased model, but neither of them obtains the best result. This shows the necessity and effectiveness of the combination. For some matrices (eg. document MAP on 2007 collection and aspect MAP on both collections), the best result occurs when α is set in the range of $(0.6 \sim 0.8)$. So the empirical setting of parameter α is suggested to be $(0.6 \sim 0.8)$ when no training data is available.

5.5 Comparison with Re-Ranking Method

Yin et al proposed a cost-function re-ranking method based on detected aspects using Wikipedia for promoting diversity in biomedical IR [74]. The re-ranking tactic can be deployed on the basis of arbitrary ranking result. For example, re-ranking on 2007 collection on top of that year's best result receives further improvement. Therefore we compare our performance of combined ranking model with re-ranking method results in Table 5.7 and Table 5.8. No statistical test is conducted because their results for individual queries are not available. In the tables, bold font denotes the better result.

Table 5.7: Comparison with Re-Ranking Method on 2006 Collection

MAP	Aspect	Passage	Document
Re-Rank	0.2374	0.0386	0.3549
LTR	**0.2400**	**0.0416**	**0.3910**

Table 5.8: Comparison with Re-Ranking Method on 2007 Collection

MAP	Aspect	Passage	Passage2	Document
Re-Rank	0.1642	0.0651	0.0679	0.2116
LTR	**0.1923**	**0.0784**	**0.0831**	**0.2721**

As shown in Table 5.7 and Table 5.8, the proposed method achieves performance improvements over the re-ranking method in terms of all metrics on both 2006 and 2007 collections. We attribute this to the diversity-representative features proposed in this thesis and the utilization of learning to rank technology. Learning-to-rank has demonstrated power in integrating multiple sources of features for constructing ranking model. Same as other machine learning methods, features play an important role in learning to rank. As proven usefulness in previous section, diversity-representative features essentially enhance the learning to rank method with greater opportunity to capture novelty and diversity information in ranking list which results in building better ranking model.

5.6 Case Study

In the previous sections, it has shown that the proposed learning to rank framework achieved performance improvement over baselines in terms of different levels of MAP metrics. In this section, we will show how the ranking results perform in reality, more specifically we will show the top 5 ranking results of each different ranking models and analyze the content of the top returned passage.

For a given query "What is the role of PrnP in mad cow disease?", Table 5.9 5.10, and 5.11 show the top 5 passages returned by BM25, Language Model and the proposed learning to rank method, respectively. The content shows the cleaned text of the retrieved passage and aspects shows what aspects the passage mainly cover with respect to the query. The aspects are semantically detected using Wikipedia.

Table 5.9: Case Study: Top 5 Passages in Ranking List Returned by BM25 for Query 160

Rank	Passage ID	Content	Aspects

1	16033959_64234.313	Miller M W amp Williams E S 2004 Chronic wasting disease of cervids In Mad Cow Disease and Related Spongiform Encephalopathies pp 160 193 150 214 Edited by D A Harris New York Springer	New York; Chronic wasting disease; Miller; Deer; Bovine spongiform encephalopathy;
2	10922352_30934.218	9 Booker C Vaccine link to human cases of mad cow disease The Sunday Telegraph 9 May 1999 p 26 cols 1 150 3	Vaccine; Cattle; Bovine spongiform encephalopathy; The Sunday Telegraph; Disease;
3	11980826_0.107	The Mad Cow Talks Back Jo Shapcott	Bovine spongiform encephalopathy; Cattle; Jo Shapcott;

4	10841816_6639.467	Explaining this surprise takes a few steps back to the mid 1990s Clarke and Loo were turning their attention to how P glycoprotein and similar proteins form or fold during their construction The field of protein folding was gaining followers as a host of diseases including cystic fibrosis sickle cell anemia and mad cow disease were found to be mediated by misshapen proteins	Cystic fibrosis; Sickle-cell disease; Protein folding; Glycoprotein; Protein; Sickle; Bovine spongiform encephalopathy; Cattle; Disease; Anemia;
5	15735256_1142.558	We all know that our small polluted violent planet is endangered Since 11 September 2001 we feel that globalization is bringing more than the opening up of markets we now fear terrorist attacks 1 The SARS epidemic and mad cow disease confronted us ...	Endangered species; Cattle; Planet; Globalization; Food industry; Severe acute respiratory syndrome; Disease;

Table 5.10: Case Study: Top 5 Passages in Ranking List Returned by Language Model for Query 160

Rank	Passage ID	Content	Aspects
1	16033959_64234.313	Miller M W amp Williams E S 2004 Chronic wasting disease of cervids In Mad Cow Disease and Related Spongiform Encephalopathies pp 160 193 150 214 Edited by D A Harris New York Springer	New York; Chronic wasting disease; Miller; Deer; Bovine spongiform encephalopathy;
2	10922352_30934.218	9 Booker C Vaccine link to human cases of mad cow disease The Sunday Telegraph 9 May 1999 p 26 cols 1 150 3	Vaccine; Cattle; Bovine spongiform encephalopathy; The Sunday Telegraph; Disease;

3	11980826_0.107	The Mad Cow Talks Back Jo Shapcott	Bovine spongiform encephalopathy; Cattle; Jo Shapcott;
4	10841816_6639.467	Explaining this surprise takes a few steps back to the mid 1990s Clarke and Loo were turning their attention to how P glycoprotein and similar proteins form or fold during their construction The field of protein folding was gaining followers ...	Cystic fibrosis; Sickle-cell disease; Protein folding; Glycoprotein; Protein; Sickle; Bovine spongiform encephalopathy; Cattle; Disease; Anemia;

5	11733532_5909.972	Prions are infectious proteins causing mammalian spongiform encephalopathies such as scrapie mad cow disease and Creutzfeld Jakob disease 1 Prions propagate by converting the normal form of the PrP protein into an altered sheet rich conformation 2 Prion diseases ...	Scrapie; Prion; Bovine spongiform encephalopathy; Disease; Protein; Protein structure; PRNP; Amyloid; ...

Table 5.11: Case Study: Top 5 Passages in Ranking List Returned by Diversity Learning to Rank Model for Query 160

Rank	Passage ID	Content	Aspects

1	15722549_11704.843	In order to determine the individual involvement of the codon 108 and 189 polymorphisms in disease and the mechanism by which they control TSE incubation time in mice 108F and 189V have been introduced separately into the murine Prnp ...	Gene; Gene targeting; Homozygous; Inoculation; Scrapie; Heterozygous; Avian incubation; Genetic code; Allele; Polymorphism (biology);

2	14573822_46115.2249	The effect of the 101L mutation on murine scrapie incubation times largely parallels the effect of the 108 189 polymorphisms in murine PrP as incubation times are extended in 101LL mice compared with the homologous transmission in either Prnp ...These experiments may reveal how mutations in this unstructured N terminal region of PrP can have dramatic effects on disease phenotype	Avian incubation; Scrapie; Homology (biology); Lesion; Inoculation; Phenotype; PRNP; Murinae; Mutation;

3	9300662_13194.1443	Table 1 Human prion diseases Type Clinical syndromes Aetiology Acquired Kuru Cannibalism Iatrogenic CJD Accidental innoculation with human prions Sporadic CJD Somatic PRNP mutation Atypical CJD or spontaneous conversion PrP C to PrP Sc Inherited Familial CJD Germline PRNP mutation GSS FFI Various ...	Fatal familial insomnia; Prion; Inoculation; Insomnia; PRNP; Kuru (disease); Gene; Mutation; Somatic; Germline; Disease; Transmissible spongiform encephalopathy; ...
4	14573822_26237.1667	View larger version 19K in this window in a new window 160 Fig 1 Transmission of murine scrapie strains to Prnp a101L and Prnp a 108F 189V mice Incubation times 177 SEM of six mouse passaged TSE agents in Prnp a mice white bars ...	Scrapie; Avian incubation; Transmissible spongiform encephalopathy; Murinae;

5	15722549_8230.2514	The host PrP gene Prnp has a major influence over the outcome of TSE disease PrP polymorphisms have been shown to alter incubation time and TSE susceptibility in mice Moore et al 1998 sheep Goldmann et al 1994 and man Palmer ...	Scrapie; Avian incubation; Gene; Inbreeding; Genetics; Allele; Genetic analysis; Mouse; Polymorphism (biology); ...

Table 5.12 shows the aspect coverage of the top 5 passages returned by different models. It can be seen that the LTR ranking method has almost double the number of aspects covered than that are covered by baseline methods. This shows that the LTR ranking model provides more aspects of the answer to the query. In this sense, a more diversified ranking result is given by the proposed learning to rank method.

Table 5.12: Aspect Coverage of Top 5 Passage Returned by Different Models

Method	Aspect Number
BM25	17
DirKL	17
LTR	33

5.7 Summary

In this chapter, the principle of the experimental design in this research was first introduced. Then the key questions of this research were experimentally analyzed and answered in different sections. Thorough experimental results have been presented to demonstrate that the learning to rank technique is appropriate and beneficial to applying to biomedical field. Extensive experiments have shown the effectiveness of the proposed diversity-based learning to rank model. From the results and analyses it is safe to draw a conclusion that the proposed diversity features are representative of diversity information of ranking list and useful in advancing ranking model within the combined learning to rank framework proposed in this research. The influence of the parameters in the proposed framework was also studied showing that the proposed method is free of parameter tuning. A case study was given for a given query which demonstrates that the proposed method provides more diversified ranking results.

6 Conclusion and Future Work

6.1 Conclusion

In this thesis, we have applied learning to rank technology to biomedical IR. proposed a combined learning to rank model which integrates a general ranking model and a diversity-biased model. The diversity-biased model is learned from both general features and diversity-favored features to award ranking list with low redundancy and high diversity. The diversity-reflecting features which are defined in the perspective of topics relationship of different passages in ranking order appear to contribute promoting results diversity. Thorough experiments have been conducted on the dataset of TREC 2006 and 2007 Genomics Tracks. Experimental comparison with baselines methods, which are traditional unsupervised IR methods, shows the effectiveness of general learning to rank model. Moreover, within the framework of combined ranking model, with the help of diversity-biased model, the retrieval results are proven to be more promising.

6.2 Future Work

In the future, there are several directions that could be considered for extending this work:

(1) Conduct more experiments of the proposed method, for example, in 10-fold cross-validation setting, on other datasets, with different parameters settings of learning to rank features, comparing with more methods, and assessing significant test.

(2) Explore the usage of the proposed diversity features to other tasks, such as question answering task.

(3) Apply this framework to other IR domains, for example, web search.

(4) Design more features to integrate into this framework, such as biomedical domain specific features. Potential available resources include but are not restricted to: MeSH, ICD-10.

A Topics

A.1 TREC 2006 Genomics Track Topics

⟨160⟩What is the role of PrnP in mad cow disease?

⟨161⟩What is the role of IDE in Alzheimers 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 Parkinsons disease?

⟨165⟩How do Cathepsin D (CTSD) and apolipoprotein E (ApoE) interactions contribute to Alzheimers 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?

⟨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 Huntingtons 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 Alzheimers disease?

⟨187⟩How do mutations in familial hemiplegic migraine type 1 (FHM1) gene affect calcium ion influx in hippocampal neurons?

A.2 TREC 2007 Genomics Track Topics

⟨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 OR 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?

⟨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 [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?

B Related Published Papers

- Wu, J., Huang, J., and Ye, Z. (2013). Expoliting rich features for promoting diversity in biomedical information retrieval. In Bioinformatics and Biomedicine (BIBM), 2013 IEEE International Conference on, pages 624-624.

- Wu, J., Huang, J., and Ye, Z. (2014). Learning to Rank Diversified Results for Biomedical Information Retrieval from Multiple Features. BMC Medical Bioinformatics and Decision Making. (to appear)

- Shang, Y., Hao, H., Lin, H., and Wu, J. (2013). Learning to rank based gene summary extraction. In Bioinformatics and Biomedicine (BIBM), 2013 IEEE International Conference on, pages 618-618.

- Yu, F., Yang, Z., Tang, N., Wu, J., Lin, H., and Wang, J. (2013). Predicting protein complexes in protein interaction networks: A supervised learning based method. In Bioinformatics and Biomedicine (BIBM), 2013 IEEE International Conference on, pages 188-188.

Bibliography

[1] Agrawal, R., Gollapudi, S., Halverson, A., and Ieong, S. (2009). Diversifying search results. In *Proceedings of the Second ACM International Conference on Web Search and Data Mining*, WSDM '09, pages 5–14, New York, NY, USA. ACM.

[2] Allan, J. (2004). Hard track overview in trec 2004 - high accuracy retrieval from documents. In *The Thirteenth Text Retrieval Conference (TREC 2004)*, Gaithersburg. National Institute of Standards and Technology.

[3] Amati, G., Joost, C., and Rijsbergen, V. (2002). Probabilistic models for information retrieval based on divergence from randomness. *TOIS*, 20:357–389.

[4] Auer, P., Cesa-Bianchi, N., and Fischer, P. (2002). Finite-time analysis of the multiarmed bandit problem. *Mach. Learn.*, 47(2-3):235–256.

[5] Bendersky, M., Metzler, D., and Croft, W. B. (2010). Learning concept importance using a weighted dependence model. In *WSDM*, pages 31–40.

[6] Blei, D. M., Ng, A. Y., Jordan, M. I., and Lafferty, J. (2003). Latent dirichlet allocation. *JMLR*, 3:2003.

[7] Burges, C., Shaked, T., Renshaw, E., Deeds, M., Hamilton, N., and Hullender, G. (2005). Learning to rank using gradient descent. In *In ICML*, pages 89–96.

[8] Cao, Z., Qin, T., Liu, T.-Y., Tsai, M.-F., and Li, H. (2007). Learning to rank: from pairwise approach to listwise approach. In *ICML*, pages 129–136.

[9] Carbonell, J. and Goldstein, J. (1998). The Use of MMR, Diversity-based Reranking for Reordering Documents and Producing Summaries. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '98, pages 335–336, New York, NY, USA. ACM.

[10] Chakrabarti, S., Khanna, R., Sawant, U., and Bhattacharyya, C. (2008). Structured learning for non-smooth ranking losses. In *KDD*, pages 88–96.

[11] Chen, S. F. and Goodman, J. (1996). An empirical study of smoothing techniques for language modeling. In *Proceedings of the 34th Annual Meeting on Association for Computational Linguistics*, ACL '96, pages 310–318, Stroudsburg, PA, USA. Association for Computational Linguistics.

[12] Claire Fautsch, J. S. (2007). IR-Specific Searches at TREC 2007: Genomics

& Blog Experiments. In *The Sixteenth Text Retrieval Conference (TREC 2007)*, Gaithersburg. National Institute of Standards and Technology.

[13] Crammer, K. and Singer, Y. (2001). Pranking with ranking. In *Advances in Neural Information Processing Systems 14*, pages 641–647. MIT Press.

[14] Craswell, N., Robertson, S., Zaragoza, H., and Taylor, M. (2005a). Relevance weighting for query independent evidence. In *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '05, pages 416–423, New York, NY, USA. ACM.

[15] Craswell, N., Zaragoza, H., and Robertson, S. (2005b). Microsoft cambridge at trec-14: Enterprise track. In *In Voorhees and Buckland [9*.

[16] Croft, W. B., Turtle, H. R., and Lewis, D. D. (1991). The use of phrases and structured queries in information retrieval. In *Proceedings of the 14th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '91, pages 32–45, New York, NY, USA. ACM.

[17] Demner-Fushman, D., Humphrey, S. M., Ide, N. C., Loane, R. F., Mork, J. G., Ruch, P., Ruiz, M. E., Smith, L. H., Wilbur, W. J., and Aronson, A. R. (2007). Combining resources to find answers to biomedical questions. In *TREC*.

[18] Emanuel, G. (1931). Statistical machine.

[19] Fagan, J. L. (1987). Automatic phrase indexing for document retrieval: An examination of syntactic and non-syntactic methods. In *SIGIR*, pages 91–101. ACM.

[20] Freund, Y., Iyer, R., Schapire, R. E., and Singer, Y. (2003). An efficient boosting algorithm for combining preferences. *J. Mach. Learn. Res.*, 4:933–969.

[21] Freund, Y. and Schapire, R. E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1):119 – 139.

[22] Gabrilovich, E. and Markovitch, S. (2007). Computing semantic relatedness using wikipedia-based explicit semantic analysis. In *IJCAI*, pages 1606–1611.

[23] Garey, M. R. and Johnson, D. S. (1990). *Computers and Intractability; A Guide to the Theory of NP-Completeness*. W. H. Freeman & Co., New York, NY, USA.

[24] Gibson, D., Kleinberg, J., and Raghavan, P. (1998). Clustering categorical data: An approach based on dynamical systems. pages 311–322.

[25] Goldberg, A. B., Andrzejewski, D., Gael, J. V., Settles, B., Zhu, X., and Craven, M. (2006). Ranking biomedical passages for relevance and diversity: University of Wisconsin, Madison at TREC Genomics 2006. In *TREC*.

[26] Gordon, M. D. and Lenk, P. (1991). A utility theoretic examination of the probability ranking principle in information retrieval. *JASIS*, 42(10):703–714.

[27] Guiasu, S. and Shenitzer, A. (1985). The principle of maximum entropy. *The Mathematical Intelligencer*, 7(1):42–48.

[28] He, B., Huang, J. X., and Zhou, X. (2011). Modeling term proximity for probabilistic information retrieval models. *Information Sciences*, 181(14):3017 – 3031.

[29] Herbrich, R., Graepel, T., and Obermayer, K. (1999). Large Margin Rank Boundaries for Ordinal Regression. In *Advances in Large Margin Classifiers*, pages 115–132. The MIT Press.

[30] Hersh, W., Cohen, A., Ruslen, L., and Roberts, P. (2007). TREC 2007 Genomics track overview. In *TREC*. National Institute of Standards and Technology.

[31] Hersh, W., Cohen, A. M., Roberts, P., and Rekapalli, H. K. (2006). TREC 2006 genomics track overview. In *TREC*. National Institute of Standards and Technology.

[32] Hiemstra, D. (2001). *Using language models for information retrieval*. Univ. Twente.

[33] Hofmann, T. (1999). Probabilistic latent semantic analysis. In *UAI*, pages 289–296.

[34] Huang, A., Milne, D., Frank, E., and Witten, I. H. (2008). Clustering documents with active learning using wikipedia. In *ICDM*, pages 839–844.

[35] J. Jiang, X. He, C. Z. (2006). Robust pseudo feedback estimation and HMM passage extraction: UIUC at TREC 2006 genomics track. In *The Fifteenth Text Retrieval Conference (TREC 2006)*, Gaithersburg. National Institute of Standards and Technology.

[36] Jain, A. K. and Dubes, R. C. (1988). *Algorithms for clustering data*. Prentice-Hall, Inc., Upper Saddle River, NJ, USA.

[37] Jelinek, F. and Mercer, R. L. (1980). Interpolated estimation of markov source parameters from sparse data. In *In Proceedings of the Workshop on Pattern Recognition in Practice*, pages 381–397, Amsterdam, The Netherlands: North-Holland.

[38] Joachims, T. (2002). Optimizing search engines using clickthrough data. In *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '02, pages 133–142, New York, NY, USA. ACM.

[39] Jones, K. S., Walker, S., and Robertson, S. E. (2000). A probabilistic model of information retrieval: Development and comparative experiments. *Inf. Process. Manage.*, 36(6):779–808.

[40] Li, P., Burges, C. J. C., and Wu, Q. (2007). Mcrank: Learning to rank using multiple classification and gradient boosting.

[41] Liu, T.-Y. (2009). Learning to rank for information retrieval. *Found. Trends Inf. Retr.*, 3(3):225–331.

[42] Luo Si, J. L. and Callan, J. (2006). Combining Multiple Resources, Evidence and Criteria for Genomic Information Retrieval. In *The Fifteenth Text Retrieval Conference (TREC 2006)*, Gaithersburg. National Institute of Standards and Technology.

[43] MacKay, D. J. and Peto, L. C. B. (1994). A hierarchical dirichlet language model. *Natural Language Engineering*, 1:1–19.

[44] Markowitz, H. (1952). Portfolio selection. *Journal of Finance.*

[45] Maron, M. E. and Kuhns, J. L. (1960). On relevance, probabilistic indexing and information retrieval. *J. ACM*, 7(3):216–244.

[46] Metzler, D. and Bruce Croft, W. (2007). Linear feature-based models for information retrieval. *Inf. Retr.*, 10(3):257–274.

[47] Milne, D. and Witten, I. H. (2013). An open-source toolkit for mining wikipedia. *Artif. Intell.*, 194:222–239.

[48] Milne, D. N., Witten, I. H., and Nichols, D. M. (2007). A knowledge-based search engine powered by wikipedia. In *CIKM*, pages 445–454.

[49] Mooers, C. (1950). *The theory of digital handling of non-numerical information and its implications to machine economics*. Zator technical bulletin. Zator Co.

[50] Nallapati, R. (2004). Discriminative models for information retrieval. In *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '04, pages 64–71, New York, NY, USA. ACM.

[51] Ng, R. T. and Han, J. (1994). Efficient and effective clustering methods for spatial data mining. In Bocca, J. B., Jarke, M., and Zaniolo, C., editors, *VLDB'94, Proceedings of 20th International Conference on Very Large Data Bases, September 12-15, 1994, Santiago de Chile, Chile*, pages 144–155. Morgan Kaufmann.

[52] O. Chapelle, Q. L. and Smola, A. (2007). Large margin optimization of ranking measures. In *In NIPS workshop on Machine Learning for Web Search*.

[53] Page, L., Brin, S., Motwani, R., and Winograd, T. (1999). The pagerank citation ranking: Bringing order to the web.

[54] Ponte, J. M. and Croft, W. B. (1998). A language modeling approach to information retrieval. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '98, pages 275–281, New York, NY, USA. ACM.

[55] Porter, M. F. (1997). Readings in information retrieval. chapter An Algorithm for Suffix Stripping, pages 313–316. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

[56] Radlinski, F., Kleinberg, R., and Joachims, T. (2008). Learning diverse rankings with multi-armed bandits. In *ICML*, pages 784–791.

[57] Robertson, S., Walker, S., and Hancock-Beaulieu, M. (1995). Large test collection experiments on an operational, interactive system: Okapi at TREC. *IPM*, 31(3):345 – 360.

[58] Robertson, S. and Zaragoza, H. (2009). The probabilistic relevance framework: Bm25 and beyond. *Found. Trends Inf. Retr.*, 3(4):333–389.

[59] Robertson, S. E. and Walker, S. (1994). 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*, SIGIR '94, pages 232–241, New York, NY, USA. Springer-Verlag New York, Inc.

[60] Salton, G. and Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. In *INFORMATION PROCESSING AND MANAGEMENT*, pages 513–523.

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

[62] Santos, R. L., Macdonald, C., and Ounis, I. (2010). Exploiting query reformulations for web search result diversification. In *Proceedings of the 19th International Conference on World Wide Web*, WWW '10, pages 881–890, New York, NY, USA. ACM.

[63] Singhal, A., Buckley, C., and Mitra, M. (1996). Pivoted document length normalization. In *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '96, pages 21–29, New York, NY, USA. ACM.

[64] Tao, T. and Zhai, C. (2007). An exploration of proximity measures in information retrieval. In *SIGIR*, pages 295–302.

[65] Vapnik, V. (1998). *Statistical learning theory.* Wiley.

[66] Vapnik, V. N. (1995a). The nature of statistical learning theory.

[67] Vapnik, V. N. (1995b). *The Nature of Statistical Learning Theory*. Springer-Verlag New York, Inc., New York, NY, USA.

[68] Wang, J. and Zhu, J. (2009). Portfolio theory of information retrieval. In *SIGIR*, pages 115–122.

[69] Wei Zhou, C. Y. (2007). TREC Genomics Track at UIC. In *The Sixteenth Text Retrieval Conference (TREC 2007)*, Gaithersburg. National Institute of Standards and Technology.

[70] Xia, F., Liu, T.-Y., Wang, J., Zhang, W., and Li, H. (2008). Listwise approach to learning to rank: Theory and algorithm. In *Proceedings of the 25th International Conference on Machine Learning*, ICML '08, pages 1192–1199, New York, NY, USA. ACM.

[71] Xu, R. and Wunsch, D., I. (2005). Survey of clustering algorithms. *Neural Networks, IEEE Transactions on*, 16(3):645–678.

[72] Xu, Y., Jones, G. J., and Wang, B. (2009). Query dependent pseudo-relevance feedback based on wikipedia. In *SIGIR*, pages 59–66.

[73] Yin, X., Huang, J. X., Zhou, X., and Li, Z. (2010a). A survival modeling approach to biomedical search result diversification using wikipedia. In *SIGIR*, pages 901–902.

[74] Yin, X., Huang, X., and Li, Z. (2010b). Promoting ranking diversity for biomedical information retrieval using wikipedia. In *ECIR*, pages 495–507.

[75] Yue, Y., Finley, T., Radlinski, F., and Joachims, T. (2007). A support vector method for optimizing average precision. In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '07, pages 271–278, New York, NY, USA. ACM.

[76] Zaragoza, H., Craswell, N., Taylor, M., Saria, S., and Robertson, S. (2004). Microsoft cambridge at trec-13: Web and hard tracks. In *IN PROCEEDINGS OF TREC 2004*.

[77] Zhai, C. and Lafferty, J. (2001a). A study of smoothing methods for language models applied to information retrieval.

[78] Zhai, C. and Lafferty, J. D. (2001b). Model-based feedback in the language modeling approach to information retrieval. In *CIKM*, pages 403–410.

[79] Zhai, C. X., Cohen, W. W., and Lafferty, J. (2003). Beyond independent relevance: Methods and evaluation metrics for subtopic retrieval. In *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval*, SIGIR '03, pages 10–17, New York, NY, USA. ACM.

[80] Zhang, B., Li, H., Liu, Y., Ji, L., Xi, W., Fan, W., Chen, Z., and Ma, W.-Y. (2005). Improving web search results using affinity graph. In *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '05, pages 504–511, New York, NY, USA. ACM.

[81] Zhao, J., Huang, J. X., and He, B. (2011). Crter: Using cross terms to enhance probabilistic information retrieval. In *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '11, pages 155–164, New York, NY, USA. ACM.

[82] Zhou, W. and Yu, C. T. (2007). TREC genomics track at UIC. In *TREC*.

[83] Zhou, X., Huang, J. X., and He, B. (2011). Enhancing ad-hoc relevance weighting using probability density estimation. In *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '11, pages 175–184, New York, NY, USA. ACM.