Dual-Space Re-ranking Model for Document Retrieval

Dong Zhou[1], **Seamus Lawless**[1], **Jinming Min**[2], **Vincent Wade**[1]

1. Center for Next Generation Localisation, University of Dublin, Trinity College
2. Center for Next Generation Localisation, Dublin City University

dongzhou1979@hotmail.com, seamus.lawless@scss.tcd.ie,
jinming.min@googlemail.com, Vincent.Wade@sccs.tcd.ie

Abstract

The field of information retrieval still strives to develop models which allow semantic information to be integrated in the ranking process to improve performance in comparison to standard bag-of-words based models. A conceptual model has been adopted in general-purpose retrieval which can comprise a range of concepts, including linguistic terms, latent concepts and explicit knowledge concepts. One of the drawbacks of this model is that the computational cost is significant and often intractable in modern test collections. Therefore, approaches utilising concept-based models for re-ranking initial retrieval results have attracted a considerable amount of study. This method enjoys the benefits of reduced document corpora for semantic space construction and improved ranking results. However, fitting such a model to a smaller collection is less meaningful than fitting it into the whole corpus. This paper proposes a dual-space model which incorporates external knowledge to enhance the space produced by the latent concept method. This model is intended to produce global consistency across the semantic space: similar entries are likely to have the same re-ranking scores with respect to the latent and manifest concepts. To illustrate the effectiveness of the proposed method, experiments were conducted using test collections across different languages. The results demonstrate that the method can comfortably achieve improvements in retrieval performance.

1 Introduction

Information retrieval often suffers from the so called *"vocabulary mismatch"* problem. A document may be semantically relevant to a query despite the fact that the specific query terms used and the terms found in the document completely or partially differ (Furnas et al., 1987). Consequently, overlap with respect to linguistic terms should not be a necessary condition in query-document similarity and methods relying on the bag-of-words model can display poor performance as a result. In order to overcome the vocabulary mismatch problem, several solutions have been suggested which exploit semantic relations between text units. Among these methods, the latent model, the explicit model and the mixed model are commonly employed.

The latent model (Landauer et al., 1998; Blei et al., 2003) tries to directly model the internal structure of "topics" or "concepts" in the text data, thus building meaningful groups beyond single words. Typically some form of dimension reduction (Fodor, 2002) is applied to the data matrix to find such latent dimensions which correspond to concepts. In contrast, the explicit model (Gabrilovich and Markovitch, 2007) indexes texts according to an external knowledge base. Typically the meaning of a piece of text is represented as a weighted vector of knowledge-based concepts derived from ex-

Coling 2010: Poster Volume, pages 1524–1532,
Beijing, August 2010

ternal resources such as ODP[1] or Wikipedia[2] articles. The mixed model (Serban et al., 2005) extends the bag-of-words vector by adding external categories derived from WordNet or similar thesaurus. Based upon these definitions, the explicit model and the mixed model are similar in nature but differ in their use of external knowledge sources.

Models such as those described above, however, have well documented drawbacks. Firstly, these methods are very computationally complex. In the latent model, complexity grows linearly with the number of dimensions and the number of documents. For example, the computational cost of singular value decomposition (SVD) is significant; no successful experiment has been reported with over one million documents (Manning et al., 2008). This has been the biggest obstacle to the widespread adoption of this kind of method. For the explicit and mixed model, the dimensions of projecting documents into the external knowledge space are often limited to ten thousand (Potthast et al., 2008) in order to facilitate the large size of the test collections used. Another problem with the explicit model is that the documents are often distributed over thousands of dimensions in which the semantic relatedness will degrade dramatically. For example, in (Sorg and Cimiano, 2008) when the whole Wikipedia collection is adopted to build the space, one document is mapped to ten thousand dimensions, in which it may only have very few truly semantically related dimensions. The means of identifying these dimensions is not reported and this may significantly influence the retrieval performance.

Therefore, researchers started to consider integrating the aforementioned models into smaller, controlled document collections to overcome these shortcomings and assist the retrieval process. (Zhou and Wade, 2009b) proposed a Latent Dirichlet Allocation (LDA)-based method to model the latent structure of "topics" deduced from the initial retrieval results. The scores obtained from this process are then combined with initial ranking scores to produce a re-ranked list of results that are superior to original ordering. The method also enjoys the benefits of fast and tractable latent semantic computation and successfully avoids the incremental build problem (Landauer et al., 1998) which commonly exists in latent semantic analysis (LSA) techniques.

There is an important factor, however, that needs to be taken into account when applying this method. Due to the smaller corpus size, fitting a latent model into this corpus is less meaningful than fitting the same model into a large, web-scale corpus. This means that some form of justification has to be applied to achieve better performance. A simple approach to address this problem is to directly apply the explicit or mixed model into a controlled corpus to improve ranking performance. A similar problem will arise in the latent model in this single semantic space, resulting in limited improvements.

To address the challenges described above, this paper proposes a dual-space model which incorporates external knowledge to enhance the semantic space produced by the latent concept method. This model is intended to produce global consistency across the semantic space: *similar entries are likely to have the same re-ranking scores with respect to the latent and manifest concepts.* In other words: in this model, if a group of documents deal with the same topic induced from a dual semantic space which shares a strong similarity with a query, the documents will get allocated similar ranking as they are more likely to be relevant to the query.

In the experiments carried out in this paper, the dual-space model is applied to ad-hoc document retrieval and compared with the initial language model-based ranker and single-space model exploiting latent and explicit features. The results show that the explicit model could only bring minor improvements over the initial ranker. The latent model delivered more significant improvements than the explicit model. Both, however, are outperformed by the dual-space model.

The main contribution of this paper is to propose a dual-space semantic model for the re-ranking problem, which aims to improve precision, especially of the most highly ranked results. Other contributions of the paper include proposing a novel way of applying the explicit model to the re-ranking problem, and performing a systematic comparison between different models.

[1] http://www.dmoz.org/
[2] http://www.wikipedia.org/

The rest of this paper is organised as follows. Related work on re-ranking and concept-based methods is briefly summarised in Section 2. Section 3 describes the latent space model and explicit space model used in the framework developed by this research, Section 4 presents details of how to build the dual-space model. In Section 5 a report is provided on a series of experiments performed over three different test collections written in English, French and German. This report includes details of the results obtained. Finally, Section 6 concludes the paper and speculates on future work.

2 Related Work

There exist several strands of related work in the areas of re-ranking and concept-based document retrieval.

A family of work on the structural re-ranking paradigm over different sized document corpora was proposed to refine initial ranking scores. Kurland and Lee performed re-ranking based on measures of centrality in the graph formed by the generation of links induced by language model scores, through a weighted version of the PageRank algorithm (Kurland and Lee, 2005) and a HITS-style cluster-based approach (Kurland and Lee, 2006). Zhang et al. (Zhang et al., 2005) proposed a similar method to improve web search based on a linear combination of results from text search and authority ranking. The graph, which they named an "affinity graph", shares strong similarities with Kurland and Lee's work where the links are induced by a modified version of cosine similarity using the vector space model. Diaz (Diaz, 2005) used score regularisation to adjust document retrieval rankings from an initial retrieval by a semi-supervised learning method. Deng et al. (Deng et al., 2009) further developed this method by building a latent space graph based on content and explicit link information. Unlike their approach this research attempts to model the explicit information directly.

The latent concept retrieval model has a long history in information retrieval. (Dumais, 1993; Dumais, 1995) conducted experiments with latent semantic indexing (LSI) on TREC[3] documents and tasks. These experiments achieved precision at, or above, that of the median TREC participant. On about 20% of TREC topics this system was the top scorer, and reportedly slightly better than average results in comparison to standard vector spaces for LSI at about 350 dimensions. (Hofmann, 1999) provides an initial probabilistic extension of the basic latent semantic indexing technique. A more satisfactory formal basis for a probabilistic latent variable model for dimensionality reduction is the LDA model (Blei et al., 2003), which is generative and assigns probabilities to documents outside of the training set. Wei and Croft (Wei and Croft, 2006) presented the first large-scale evaluation of LDA, finding it to significantly outperform the query likelihood model. (Zhou and Wade, 2009b; Zhou and Wade, 2009a) successfully applied this method to document re-ranking and achieved significant improvement over language model-based ranking and various graph-based re-ranking methods.

The explicit concept model has recently attracted much attention in the information retrieval community. Notably, explicit semantic analysis (ESA) has been proposed as an approach to computing semantic relatedness between words and thus, has a natural application in this field (Gabrilovich and Markovitch, 2007). In essence, ESA indexes documents with respect to the Wikipedia article space, indicating how strongly a given word in the document is associated to a specific Wikipedia article. In this model, each article is regarded as a concept, an analogical unit used in the latent model. As in the latent model, two words or texts can be semantically related in spite of not having any words in common. Specifically, this method has been widely adopted in cross-language information retrieval (CLIR) as an approach to resolving an extreme case of the vocabulary mismatch problem, where queries and documents are written in different languages (Potthast et al., 2008). (Anderka et al., 2009) showed that this approach has comparable performance to linguistic matching methods. (Cimiano et al., 2009) compared this method with a latent concept model based on LSI/LDA and concluded that it will outperform the latent model if trained on Wikipedia articles.

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

3 Latent and Explicit Models

In this section, an overview of the problem addressed by this paper is presented and the latent and explicit document re-ranking models are described in more detail. This section also demonstrates how these models can be used in a re-ranking setting.

3.1 Problem Definition

Let $\mathbb{D} = \{d_1, d_2, \dots, d_n\}$ denote the set of documents to be retrieved. Given a query q, a set of initial results $\mathbb{D}_{init} \in \mathbb{D}$ of top documents are returned by a standard information retrieval model (initial ranker). However, typically the performance of the initial ranker can be improved upon. The purpose of the re-ranking method developed by this research is to re-order a set of documents $\mathbb{D}'_{init}$ so as to improve retrieval accuracy at the most highly ranked results.

3.2 Latent Concept Model

The specific method used here is borrowed from (Zhou and Wade, 2009b), which is based on the LDA model. The topic mixture is drawn from a conjugate Dirichlet prior that remains the same for all documents. The process of generating a document corpus is as follows:

1) Pick a multinomial distribution $\vec{\varphi}_z$ for each topic k from a Dirichlet distribution with hyperparameter $\vec{\beta}$.
2) For each document d, pick a multinomial distribution $\vec{\theta}_d$, from a Dirichlet distribution with hyperparameter $\vec{\alpha}$.
3) For each word token w in document d, pick a topic $z \in \{1 \dots k\}$ from the multinomial distribution $\vec{\theta}_d$.
4) Pick word w from the multinomial distribution $\vec{\varphi}_z$.

LDA possesses fully consistent generative semantics by treating the topic mixture distribution as a k-parameter hidden random variable. LDA offers a new and interesting framework to model a set of documents. The documents and new text sequences (for example, queries) can easily be connected by "mapping" them to the topics in the corpus.

In a re-ranking setting, the probability that a document d generates w is estimated using a mixture model LDA. It uses a convex combination of a set of component distributions to model observations. In this model, a word w is generated from a convex combination of some hidden topics z:

$$LDA_d(w) = \sum_{z=1}^{k} p(w|z)p(z|d)$$

where each mixture model $p(w|z)$ is a multinomial distribution over terms that correspond to one of the latent topics z. This could be generated to give a distribution on a sequence of text:

$$LDA_d(w_1 w_2 \dots w_n) \stackrel{\text{def}}{=} \prod_{j=1}^{n} LDA_d(w_j)$$

Then the distance between a query and a document based on this model can be obtained. The method used here adopts the KL divergence (Baeza-Yates and Ribeiro-Neto, 1999) between the query terms and document terms to compute a Re-Rank score RS_{LDA}^{KL}:

$$RS_{LDA}^{KL} = -D(MLE_q(\cdot)||LDA_d(\cdot))$$

The final score is then obtained through a linear combination of the re-ranking scores based on the initial ranker and the latent document re-ranker, shown as follows:

$$RS_{Latent}^{LDA} = \lambda \cdot OS + (1 - \lambda) \cdot RS_{LDA}^{KL}$$

where OS denotes original scores returned by the initial ranker and λ is a parameter that can be tuned with $\lambda = 1$ meaning no re-ranking is performed.

Another well-known approach to the latent model is the LSI method. It is based on SVD, a technique from linear algebra. This method has not been reported anywhere previously for re-ranking purposes. It has been included here to compare the effectiveness of different latent approaches. As a full SVD is a loss-free decomposition of a matrix M, which is decomposed into two orthogonal matrices U and V and a diagonal matrix Σ. Estimating less singular values and their corresponding singular vectors leads to reduced dimensions resembling latent concepts so that documents are no longer represented by terms but by concepts. New documents (queries) are represented in terms of concepts by folding them into the LSI model. Next, cosine similarities may be used to compute the similarity between a query and a document to obtain RS_{LSI}^{COS} and combine it with the original score to produce the final re-ranking score:

Collection	Contents	Language	Num of docs	Size	Queries
BL (CLEF2009)	British Library Data	English (Main)	1,000,100	1.2 GB	50
BNF (CLEF2009)	Bibliothèque Nationale de France	French (Main)	1,000,100	1.3 GB	50
ONB (CLEF2009)	Austrian National Library	German (Main)	869,353	1.3 GB	50

Table 1. Statistics of test collections

$$RS_{latent}^{LSI} = \lambda' \cdot OS + (1-\lambda') \cdot RS_{LSI}^{COS}$$

3.3 Explicit Concept Model

As an example of explicit concept model (Gabrilovich and Markovitch, 2007), explicit semantic analysis attempts to index or classify a given text t with respect to a set of explicitly given external categories. The basic idea is to take as input a document d and map it to a high-dimensional, real-valued vector space. This space is spanned by a Wikipedia database $W_l = \{a_1, \ldots, a_n\}$.This mapping is given by the following function:

$$\Phi_l : T \to \mathbb{R}^{|W_l|}$$
$$\Phi_l(t) \coloneqq \langle v_1, \ldots, v_{|W_l|} \rangle$$

Where $|W_l|$ is the number of articles in Wikipedia W_l corresponding to language l. The value v_i in the vector t expresses the strength of association between t and the Wikipedia article a_i and is defined as the cosine similarity:

$$RS_{ESA}^{COS} = \frac{\langle t, a_i \rangle}{\| t \| \| a_i \|}$$

As pointed out in section 1, documents are often distributed over thousands of dimensions in which the semantic relatedness will degrade dramatically. The main purpose is to find the most relevant dimensions with respect to queries. To apply this method to re-ranking, W_l is limited to the number of highly relevant documents for a given query. In other words, the entire set of Wikipedia articles in language l is retrieved, and only return a specific number of documents as in W_l. This modification will also lead to fast computation of scores compared to scanning through the whole Wikipedia collection.

Similar to the latent model described above, the final ranking score is defined as:

$$RS_{Explicit}^{ESA} = \mu \cdot OS + (1-\mu) \cdot RS_{ESA}^{COS}$$

4 Dual space model

Armed with the latent and explicit models, the dual-space model proposed by this paper is now described. In order to make a direct connection between the two models, the key point is to make the dimensions comparable across different models. The detail presented on the latent and explicit concept models in the previous section did not describe how to define a specific number of dimensions. A simple assumption is taken here in the dual-space model: the number of dimensions produced by the explicit model has to correspond to the number of dimensions induced by the latent model. As the same group of documents are being mapped into two different semantic spaces, it is assumed that the concepts induced by the latent model reflect the hidden structures in this document collection. Therefore, the same phenomenon should be observed when applying the explicit model and vice-versa. Based on this assumption, the dual-space model could be conducted so as to make a constraint:

$$|W_l| = k$$

and the final ranking score for this dual space is:

$$RS_{dual}^{LDA} = \varsigma \cdot OS + (1-\varsigma-\tau) \cdot RS_{LDA}^{KL} + \tau \cdot RS_{ESA}^{COS}$$

or

$$RS_{dual}^{LSI} = \varsigma \cdot OS + (1-\varsigma-\tau) \cdot RS_{LSI}^{COS} + \tau \cdot RS_{ESA}^{COS}$$

4 Experiments and Results

In this section, an empirical study of the effectiveness of the dual-space model over three data collections written in English, French and German is presented.

4.1 Experimental Setup

The text corpus used in the experiment described below consisted of elements of the CLEF-2008[4] and CLEF-2009 European Library (TEL) collections[5] written in English, French and German. These collections are described in greater detail in Table 1. All of the documents in the experiment were indexed using the Terrier toolkit[6]. Prior to indexing, Porter's stemmer and a stopword list[7] were used for the English documents. A French and German analyser[8] is used to analyse French and German documents.

It is worth noting that the CLEF TEL data is actually multilingual: all collections to a greater or lesser extent contain records pointing to documents in other languages. However this is not a major problem because the majority of documents in the test collection are written in the primary language of those test collections (BL-English, BNF-French, ONB-German). Please refer to (Ferro and Peters, 2009) for a more detailed discussion about this data. These collections were chosen to test the scalability of the proposed method in different settings and over different languages.

The CLEF-2008 and CLEF-2009 query sets were also used. Both query sets consist of 50 topics in each language being tested. The CLEF-2008 queries written in English were used in training the parameters and all of the CLEF-2009 queries were used in the experiment for testing purposes. Each topic is composed of several parts, including: *Title*, *Description* and *Narrative*. *Title+Description* combinations were chosen as queries. The queries are processed similarly to the treatment of the test collections. The relevance judgments are taken from the judged pool of top retrieved documents by various participating retrieval systems from previous CLEF workshops. The initial ranker used in this study is the classic vector space model. This was selected to facilitate the LSI and ESA models used and the main purpose of the experiments is to compare different models in addition to demonstrating the effectiveness of the dual-space model.

A Wikipedia database in English, French and German was used as an explicit concept space. Only those articles that are connected via cross-language links between all three Wikipedia databases were selected. A snapshot was obtained on the 29/11/2009, which contained an aligned collection of 220,086 articles in all three languages.

The following evaluation metrics were chosen to measure the effectiveness of the various approaches: mean average precision (MAP), the precision of the top 5 documents (Prec@5), the precision of the top 10 documents (Prec@10), normalised discounted cumulative gain (NDCG) and Bpref. Statistically-significant differences in performance were determined using a paired t-test at a confidence level of 95%.

4.2 Parameter Tuning

Three primary categories of parameter combinations need to be determined in the experiments. For the latent re-ranking experiments, the parameters λ, λ' must be defined. For the explicit model the parameter μ must be chosen. For both models, the weights ς, τ have to be determined. In addition, the number of dimensions $|W_l|$ and k must be specified. Settings for these parameters were optimised with respect to MAP over the BL collection using CLEF-2008 English queries and were applied to all three collections. This optimisation was not conducted for the other metrics used.

The search ranges for these two parameters were:

$\lambda, \lambda', \mu, \varsigma, \tau$: 0.1, 0.2, …, 0.9

$|W_l|, k$: 5, 10, 15, …, 40

Note that parameters ς and τ are the weights assigned to the latent model and the explicit model in the dual-space model. The choice of one will have direct influence over another. As it turned out, for many instances, the optimal value of λ, λ' with respect to MAP was either 0.3 or 0.4, suggesting the initial retrieval scores still contain valuable information. In contrast, parameter μ shows no obvious difference in performance when the value is above 0.1. With this observation, when setting the parameters ς and τ more weight is assigned to the latent model rather than the explicit model. The optimal

[4] The test collections used in CLEF-2008 and CLEF-2009 are in fact identical.
[5] http://www.clef-campaign.org
[6] http://terrier.org
[7] ftp://ftp.cs.cornell.edu/pub/smart/
[8] http://lucene.apache.org/

	Dual space build upon LDA and ESA				**Dual space build upon LSI and ESA**			
	BL				BL			
	initial ranker	latent space	explicit space	dual space	initial ranker	latent space	explicit space	dual space
Precision@5	0.508	0.528	0.514	0.54*	0.508	0.54*	0.508	**0.556***
Precision@10	0.468	0.498*	0.47	0.508*	0.468	0.51*	0.48	**0.512***
Precision@20	0.408	0.424	0.41	**0.435***	0.408	0.408	0.407	0.409
NDCG	0.4053	0.4137*	0.4053	0.416*	0.4053	0.4145*	0.4055	**0.4213***
MAP	0.2355	0.2433*	0.2358	**0.2499***	0.2355	0.2478*	0.236	**0.2499***
R-Precision	0.316	0.3243	0.3165	**0.3248**	0.316	0.3173	0.3202*	0.3232
bpref	0.271	0.2746	0.2725	0.2812	0.271	0.2836*	0.2714	**0.2879***
	BNF				BNF			
	initial ranker	latent space	explicit space	dual space	initial ranker	latent space	explicit space	dual space
Precision@5	0.376	0.368	0.372	0.376	0.376	0.376	0.376	**0.384***
Precision@10	0.346	0.352*	0.35	0.352	0.346	0.348	0.35	**0.354***
Precision@20	0.297	0.297	0.297	0.3*	0.297	**0.303**	0.299	0.3*
NDCG	0.3162	0.3158	0.3156	0.3163	0.3162	0.317	0.3164	**0.3178**
MAP	0.1621	0.1622	0.162	**0.1634**	0.1621	0.1629	0.1622	0.1624
R-Precision	0.2274	0.2279	0.2211	**0.2285**	0.2274	0.2278	0.2264	0.2277
bpref	0.1897	0.1899	0.1887	0.19	0.1897	0.1914	0.1892	**0.1918**
	ONB				ONB			
	initial ranker	latent space	explicit space	dual space	initial ranker	latent space	explicit space	dual space
Precision@5	0.38	0.388	0.36	0.404*	0.38	0.4	0.364	**0.412***
Precision@10	0.308	0.322	0.302	**0.332***	0.308	0.324	0.302	0.324
Precision@20	0.246	0.252	0.252	**0.259***	0.246	0.247	0.251	0.252
NDCG	0.3042	0.304	0.3059	0.3101	0.3042	0.3152*	0.3062	**0.3154***
MAP	0.1482	0.1524	0.1509	0.1567*	0.1482	0.1567*	0.1494	**0.1578***
R-Precision	0.2115	0.2152	0.2137	**0.2175**	0.2115	0.212	0.2106	0.2128
bpref	0.1778	0.1871	0.1799	**0.1896**	0.1778	0.1833	0.1788	0.1832

Table 2. Experimental Results. For each evaluation setting, statistically significant differences between different methods and the initial ranker are indicated by star. Bold highlights the best results over all algorithms.

value of k was between 25 and 35 for the LDA based model and between 5 and 15 for the LSI based model. Although this demonstrates a relatively large variance, the differences in terms of MAP have remained small and statistically insignificant. $\mathbb{D}_{init}$ is set to 50 in all results reported.

4.3 Results

Primary Evaluation The main experimental results, which describe the performance of the different re-ranking algorithms on the CLEF document collection, are shown in Table 2. The first four rows in each test collection specify the most important measurements because this research is particularly interested in performance over the most highly ranked results. As illus-

trated by the data, the initial ranker was always the lowest performer in terms of nearly all measurements. This indicates the need for re-ranking. Using the method computed by the explicit space always led to an improvement in retrieval effectiveness. But this improvement is only minor in comparison to the other two models and the results are often statistically insignificant. When the re-ranking score was calculated using the latent model, retrieval effectiveness always exceeded initial ranker and the explicit model. There was a noticeable improvement in retrieval effectiveness in the English collection (BL, statistically significant results were often observed), but a modest increase for the other two collections (BNF and ONB).

The empirical results obtained using the dual space model are very promising. Pleasingly, both the LDA+ESA and LSI+ESA models outperformed the basic latent and explicit space model in the majority of retrieval runs, with the best scores relating to the LSI-based models. An important phenomenon is that statistically significant improvements are always recorded in the metrics which measure the most highly ranked results. An even more exciting observation is that in many cases, the dual-space model, even though tuned for MAP, can outperform various baselines and other models for all the evaluation metrics, with statistically significant improvements in many runs.

Another observation that can be drawn from Table 2 is that the relative performance tends to be stable across test collections written in different languages. This indicates a promising future for studying document structure with respect to latent and explicit semantic space for re-ranking purposes.

The Comparison of Latent Methods Table 2 also shows a side-by-side comparison of the various performance measurements between the latent model used in this research on the CLEF-2009 BL test collection. The LSI-based method appeared to outscore the LDA-based method in the latent model in the vast majority of cases, while the difference between the various scorings was fairly marginal as both methods deliver statistically significant results. For the dual-space model, similar results were observed. A possible reason is that the initial ranker used was based on the vector space model and LSI is also vector based. It shows that more research with respect to the latent model selection will be necessary in the future.

Effectiveness of Explicit Methods As part of experimental objectives of this research, it was also necessary to test the newly developed explicit model for re-ranking. In the parameter tuning section, the explicit model displayed no obvious difference in terms of combination effectiveness. However, some variations could be observed when applying different dimensions where statistically significant results often appear in lower dimensions. This confirms the need to find more relevant dimensions, both for performance and efficiency purposes.

5 Conclusion and Future Work

This paper proposed and evaluated a dual-space document re-ranking method for re-ordering the initial retrieval results. The key to refining the results is the global consistency over the semantic space, which leverages latent and explicit semantic information and results in state-of-art performance. This paper also proposed a novel way to apply the explicit model to the re-ranking problem, and performed a systematic comparison between different models.

Further investigation is planned in many research directions. It has been shown that the latent model-based retrieval is a promising method for ranking the whole corpus. There is a desire to call for a direct comparison between ranking and re-ranking using the proposed algorithmic variations. Future work will also include identifying improvements upon linear combination for engineering different models. At the same time, there exist a sufficient number of latent and explicit semantic techniques which will be explored to compare their performance.

Acknowledgments

The authors would like to thank the three anonymous reviewers for their many constructive comments. This research is supported by the Science Foundation Ireland (Grant 07/CE/I1142) as part of the Centre for Next Generation Localisation (www.cngl.ie) at University of Dublin, Trinity College and Dublin City University.

References

Anderka, Maik, Nedim Lipka and Benno Stein. 2009. Evaluating Cross-Language Explicit Semantic Analysis and Cross Querying at TEL@CLEF 2009. In *CLEF 2009 Workshop*, Corfu, Greece.

Baeza-Yates, Ricardo A. and Berthier Ribeiro-Neto. 1999. *Modern Information Retrieval*, Addison-Wesley Longman Publishing Co., Inc.

Blei, David M., Andrew Y. Ng and Michael I. Jordan. 2003. Latent dirichlet allocation. *J. Mach. Learn. Res.* **3**: 993-1022.

Cimiano, Philipp, Antje Schultz, Sergej Sizov, Philipp Sorg and Steffen Staab. 2009. Explicit versus latent concept models for cross-language information retrieval. In *Proceedings of the 21st international jont conference on Artifical intelligence*, Pasadena, California, USA, Morgan Kaufmann Publishers Inc. p. 1513-1518.

Deng, Hongbo, Michael R. Lyu and Irwin King. 2009. Effective latent space graph-based re-ranking model with global consistency. In *Proceedings of the Second ACM WSDM conference*, Barcelona, Spain, ACM. p. 212-221.

Diaz, Fernando. 2005. Regularizing ad hoc retrieval scores. In *Proceedings of the 14th ACM CIKM conference*, Bremen, Germany, ACM. p. 672-679.

Dumais, Susan T. 1993. Latent semantic indexing (LSI) and TREC-2. In *Proceedings of TREC*. p. 105-115.

Dumais, Susan T. 1995. Latent semantic indexing (LSI): TREC-3 report. In *Proceedings of TREC*. p. 219-230.

Ferro, Nicola and Carol Peters. 2009. CLEF 2009 Ad Hoc Track Overview: TEL & Persian Tasks. In *Working notes of CLEF2008*, Corfu, Greece.

Fodor, Imola K. 2002. A Survey of Dimension Reduction Techniques. http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.8.5098. Accessed: 18th April 2010.

Furnas, G. W. , T. K. Landauer, L. M. Gomez and S. T. Dumais. 1987. The vocabulary problem in human-system communication. *Commun. ACM* **30**(11): 964-971.

Gabrilovich, Evgeniy and Shaul Markovitch. 2007. Computing semantic relatedness using Wikipedia-based explicit semantic analysis. In *Proceedings of the 20th international joint conference on Artifical intelligence*, Hyderabad, India, Morgan Kaufmann Publishers Inc. p. 1606-1611.

Hofmann, Thomas. 1999. Probabilistic latent semantic indexing. In *Proceedings of the 22nd annual international ACM SIGIR conference*, Berkeley, California, United States, ACM. p. 50-57.

Kurland, Oren and Lillian Lee. 2005. PageRank without hyperlinks: structural re-ranking using links induced by language models. In *Proceedings of the 28th annual international ACM SIGIR conference,* Salvador, Brazil, ACM. p. 306-313.

Kurland, Oren and Lillian Lee. 2006. Respect my authority!: HITS without hyperlinks, utilizing cluster-based language models. In *Proceedings of the 29th annual international ACM SIGIR conference*, Seattle, Washington, USA, ACM. p. 83-90.

Landauer, Thomas K., Peter W. Foltz and Darrell Laham. 1998. An Introduction to Latent Semantic Analysis. *Discourse Processes* **25**: 259-284.

Manning, Christopher D., Prabhakar Raghavan and Hinrich Schtze. 2008. *Introduction to Information Retrieval*, Cambridge University Press.

Potthast, Martin, Benno Stein and Maik Anderka. 2008. A Wikipedia-Based Multilingual Retrieval Model. In *Proceedings of 30th European Conference on Information Retrieval*, Glasgow, Scotland, Springer. p. 522-530.

Serban, Radu, Annette ten Teije, Frank van Harmelen, Mar Marcos and Cristina Polo. 2005. Ontology-driven extraction of linguistic patterns for modelling clinical guidelines. *Proceedings of the 10th European Conference on Artificial Intelligence in Medicine (AIME-05)*.

Wei, Xing and W. Bruce Croft. 2006. LDA-based document models for ad-hoc retrieval. In *Proceedings of the 29th annual international ACM SIGIR conference*, Seattle, Washington, USA, ACM. p. 178-185.

Zhang, Benyu, Hua Li, Yi Liu, Lei Ji, Wensi Xi, Weiguo Fan, Zheng Chen and Wei-Ying Ma. 2005. Improving web search results using affinity graph. In *Proceedings of the 28th annual international ACM SIGIR conference*, Salvador, Brazil, ACM. p. 504-511.

Zhou, Dong and Vincent Wade. 2009a. Language Modeling and Document Re-Ranking: Trinity Experiments at TEL@CLEF-2009. In *CLEF 2009 Workshop*, Corfu, Greece.

Zhou, Dong and Vincent Wade. 2009b. Latent Document Re-Ranking. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, Singapore, ACL. p. 1571-1580.