# Latent Document Re-Ranking

**Dong Zhou**[1,2] **Vincent Wade**[1]
1. University of Dublin, Trinity College, Dublin 2, Ireland
2. School of Computer and Communication, Hunan University, Changsha, Hunan, China
dongzhou1979@hotmail.com Vincent.Wade@cs.tcd.ie

## Abstract

The problem of re-ranking initial retrieval results exploring the intrinsic structure of documents is widely researched in information retrieval (IR) and has attracted a considerable amount of time and study. However, one of the drawbacks is that those algorithms treat queries and documents separately. Furthermore, most of the approaches are predominantly built upon graph-based methods, which may ignore some hidden information among the retrieval set.

This paper proposes a novel document re-ranking method based on Latent Dirichlet Allocation (LDA) which exploits the implicit structure of the documents with respect to original queries. Rather than relying on graph-based techniques to identify the internal structure, the approach tries to find the latent structure of "topics" or "concepts" in the initial retrieval set. Then we compute the distance between queries and initial retrieval results based on latent semantic information deduced. Empirical results demonstrate that the method can comfortably achieve significant improvement over various baseline systems.

## 1 Introduction

Consider a traditional IR problem, where there exists a set of documents $\mathbb{D}$ in the collection. In response to an information need (as expressed in a query $q$), the system determines a best fit between the query and the documents and returns a list of retrieval results, sorted in a decreasing order of their relevancy. In practice, high precision at the top rankings of the returned results is of particular interest. Generally, there are two ways to automatically assist in achieving this ultimate goal after an initial retrieval process (Baeza-Yates and Ribeiro-Neto, 1999): document re-ranking and query expansion/re-weighting. Since the latter normally need a second round of retrieval process, our method focuses on the document re-ranking approach. We will focus on adjusting the ranking positions directly over initial retrieval results set $\mathbb{D}_{init}$.

Recently, there is a trend of exploring the hidden structure of documents to re-rank results. Some of the approaches represent the document entities as a connected graph $G$. It is usually constructed by links inferred from the content information as a nearest-neighbor graph. For example, Zhang et al. (2005) proposed an affinity ranking graph to re-rank search results by optimizing diversity and information richness. Kurland and Lee (2005) introduced a structural re-ranking approach by exploiting asymmetric relationships between documents induced by language models. Diaz (2005); Deng et al. (2009) use a family of semi-supervised machine learning methods among documents graph constructed by incorporating different evidences. However in this work we are more interested in adopting an automatic approach.

There are two important factors that should be taken into account when designing any re-ranking algorithms: the original queries and initial retrieval scores. One of issues is that previous structural re-ranking algorithms treat the query and the content individually when computing re-ranking scores. Each document is assigned a score independent of other documents without considering of queries. The problem we want to address in this paper is how we can leverage the interconnections between query and documents for the re-ranking purpose.

Another problem with such approaches concerns the fundamental re-ranking strategy they adopted. HITS (Kleinberg, 1999) and PageRank



*Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 1571–1580, Singapore, 6-7 August 2009. 

(Brin and Page, 1998) style algorithms were widely used in the past. However, approaches depend only on the structure of the global graph or sub-graph may ignore important information content of a document entity. As pointed out by Deng et al. (2009), re-ranking algorithms that rely only on the structure of the global graph are likely lead to the problem of topic drift.

Instead, we introduce a new document re-ranking method based on Latent Dirichlet Allocation (LDA) (Blei et al., 2003) which exploits implicit structure of the documents with respect to original queries. Rather than relying on graph-based techniques to identify the internal structure, the approach tries to directly model the latent structure of "topics" or "concepts" in the initial retrieval set. Then we can compute the distance between queries and initial retrieval results based on latent semantic information inferred. To prevent the problem of topic drift, the generative probability of a document is summed over all topics induced. By combining the initial retrieval scores calculated by language models, we are able to gather important information for re-ranking purposes. The intuition behind this method is the hidden structural information among the documents: *similar documents are likely to have the same hidden information with respect to a query*. In other words, if a group of documents are talking about the same topic which shares a strong similarity with a query, in our method they will get allocated similar ranking as they are more likely to be relevant to the query. In addition, the refined ranking scores should be relevant to the initial ranking scores, which, in our method, are combined together with the re-ranking score either using a linear fashion or multiplication process.

To illustrate the effectiveness of the proposed methodology, we apply the framework to ad-hoc document retrieval and compare it with the initial language model-based method and other three PageRank style re-ranking methods. Experimental results show that the improvement brought by our method is consistent and promising.

The rest of the paper is organized as follows. Related work on re-ranking algorithms and LDA based methods is briefly summarized in Section 2. Section 3 describes the re-ranking framework based on latent information induced together with details of how to build generative model. In Section 4 we report on a series of experiments performed over three different test collections in English and French as well as results obtained. Finally, Section 5 concludes the paper and speculates on future work.

## 2 Related Work

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

The first category performs re-ranking by using inter-document relationship (Lee et al., 2001), evidences obtained from external resources (Kamps, 2004), or through local context analysis (Xu and Croft, 2000). In the past, document distances (Balinski and Daniowicz, 2005), manually built external thesaurus (Qu et al., 2001), and structural information (such as document title) (Luk and Wong, 2004), etc have been used extensively for this very purpose.

A second category of work is related to recent advances in structural re-ranking paradigm over graphs. Kurland and Lee performed re-ranking based on measures of centrality in the graph formed by generation links induced by language model scores, through a weighted version of PageRank algorithm (Kurland and Lee, 2005) and HITS-style cluster-based approach (Kurland and Lee, 2006). 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 affinity graph, shares strong similarities with Kurland and Lee's work with the links induced by a modified version of cosine similarity using the vector space model. Diaz (2005) used score regularization to adjust document retrieval rankings from an initial retrieval by a semi-supervised learning method. Deng et al. (2009) further developed this method. They built a latent space graph based on content and explicit links information. Unlike their approach we are trying to model the latent information directly.

This work is also related to a family of methods so called latent semantic analysis (LSA) (Landauer et al., 1998), especially topic models used for document representation. Latent Dirichlet Allocation (LDA), after it was first introduced by Blei et al. (2003), has quickly become one of the most popular probabilistic text modeling techniques and has inspired research ranging from text classification and clustering (Phan et al., 2008), information discovery (Mei et al., 2007; Titov and McDonald, 2008) to information retrieval (Wei and Croft, 2006). In this model, each topic is represented by a set of words and each word corresponds with a weight to measure its contribution to the topic. Wei and Croft

(2006) described large-scale information retrieval experiments by using LDA. In their work, LDA-based document model and language model-based document model were linearly combined to rank the entire corpus. However, unlike this approach we only apply LDA to a small set of documents. There are two reasons by doing so. One is the concern of computational cost. LDA is a very complex model and the complexity will grow linearly with the number of topics and the number of documents. Only running it through a document set significantly smaller than the whole corpus has obvious advantages. Secondly, it is well known that LSA-based method suffers from an incremental build problem. Normally adding new documents to the corpus needs to "be folded in" to the latent representation. Such incremental addition fails to capture the co-occurrences of the newly added documents (and even ignores all new terms they contain). As such, the quality of the LSA representation will degrade as more documents are added and will eventually require a re-computation of the LSA representation. Because our method only requires running LDA once for a small number of documents, this problems could be easily avoided. In addition, we also introduce two new measures to calculate the distance between a query and a document.

## 3 Latent Re-Ranking Framework

In this section, we describe a novel document re-ranking method based on extracting the latent structure among the initial retrieval set and measuring the distance between queries and documents.

### 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, the initial ranker tends to be imperfect. The purpose of our re-ranking method is to re-order a set of documents $\mathbb{D}'_{init}$ so as to improve retrieval accuracy at the very top ranks of the final results.

### 3.2 Latent Dirichlet Allocation

We will first introduce Latent Dirichlet Allocation model which forms the basis of the re-ranking framework that will be detailed in the next subsection. It was previously shown that co-occurrence structure of terms in text documents can be used to recover some latent topic structures without any usage of background information (Landauer et al., 1998). This means that latent-topic representations of text allow modeling of linguistic phenomena such as synonymy and polysemy. By doing so, information retrieval systems can match the information needs with content items on a meaning level rather than by just lexical congruence.

The basic generative process of LDA closely resembles PLSA (Hofmann, 1999). LDA extends PLSA method by defining a complete generative model of text. 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 $\vec{\varphi}_z$.

Thus, the likelihood of generating a corpus is:

$$p(d_1, \dots, d_n | \vec{\alpha}, \vec{\beta}) = \iint \prod_{d=1}^{n} p(\vec{\theta}_d | \vec{\alpha}) \cdot \prod_{z=1}^{k} p(\vec{\varphi}_z | \vec{\beta}) \cdot \prod_{i=1}^{N_d} \sum_{z_i=1}^{k} p(z_i | \vec{\theta}_d) p(w_i | z, \vec{\varphi}_z) \, d\vec{\theta}_d \, d\vec{\varphi}_z$$

Unlike PLSA model, 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) could be easily connected by "mapping" them to the topics in the corpus. In the next subsection we will introduce how to achieve this goal and apply it to document re-ranking.

LDA is a complex model and cannot be solved by exact inference. There are a few approximate inference techniques available in the literature: variational methods (Blei et al., 2003), expectation propagation (Griffiths and Steyvers, 2004)

and Gibbs sampling (Griffiths and Steyvers, 2004). Gibbs sampling is a special case of Markov-Chain Monte Carlo (MCMC) simulation and often yields relatively simple algorithms. For this reason, we choose to use Gibbs sampling to estimate LDA.

According to Gibbs sampling, we need to compute the conditional probability $p(z_i|\vec{z}_{\neg i}, \vec{w})$, where $\vec{w}$ denotes the vector of all words and $\vec{z}_{\neg i}$ denotes the vector of topic assignment except the considered word at position $i$. This probability distribution can be derived as:

$$p(z_i|\vec{z}_{\neg i}, \vec{w}) = \frac{n^{w_i}_{z,\neg i} + \beta_{w_i}}{(\sum_{v=1}^{V} n^{v}_{z} + \beta_v) - 1} \cdot \frac{n^{z}_{d_i,\neg i} + \alpha_k}{(\sum_{z=1}^{k} n^{z}_{d_i} + \alpha_z) - 1}$$

where $n^{t}_{z,\neg i}$ indicates the number of instances of word $w_i$ assigned to topic $z = k$, not including the current token and $n^{z}_{d_i,\neg i}$ denotes the number of words in document $d_i$ assigned to topic $z = k$, not including the current token.

Then we can obtain the multinomial parameter sets:

$$\theta_{d_i,k} = \frac{n^{k}_{d_i} + \alpha_k}{\sum_{z=1}^{k} n^{z}_{d_i} + \alpha_z}$$

$$\varphi_{k,w_i} = \frac{n^{w_i}_{k} + \beta_{w_i}}{\sum_{v=1}^{V} n^{v}_{z} + \beta_v}$$

The Gibbs sampling algorithm runs over three periods: initialization, burn-in and sampling. We do not tune to optimize these parameters because in our experiments the markov chain turns out to converge very quickly.

### 3.3 LDA-based Re-Ranking

Armed with this LDA methodology, we now describe the main idea of our re-ranking method. Given a set of initial results $\mathbb{D}_{init}$, we are trying to re-measure the distance between the query and a document. In the vector space model, this distance is normally the cosine or inner product measure between two vectors. Under the probabilistic model framework, this distance can be obtained from a non-commutative measure of the difference between two probability distributions. The distance used in our approach is the Kullback-Leibler (KL) divergence (Kullback and Leibler, 1951). Given two probability mass function $p(x)$ and $q(x)$, the KL divergence (or relative entropy) between $p$ and $q$ is defined as:

$$D(p||q) = \sum_{x} p(x) log \frac{p(x)}{q(x)}$$

In terms of text sequences (either queries or documents), the probability distribution can be regarded as a probabilistic language model $M_d$ or $M_q$ from each document $d$ or each query $q$. In other words, it assumes that there is an underlying language model which "generates" a term (sequence) (Ponte and Croft, 1998). The unigram language model is utilized here. There are several ways to estimate the probabilities. Let $g(w \in d)$ denotes the number of times the term $w$ occurs in a document $d$ (same idea can be used on a query). The Maximum-likelihood estimation (MLE) of $w$ with respect to $d$ is defined as:

$$MLE_d w \stackrel{\text{def}}{=} \frac{g(w \in d)}{\sum_{w'} g(w' \in d)}$$

Previous work in language-model-based information retrieval (Zhai and Lafferty, 2004) advocates the use of a Dirichlet-smoothed estimation:

$$DIR_d w \stackrel{\text{def}}{=} \frac{g(w \in d) + \mu \cdot MLE_w \mathbb{D}}{\sum_{w'} g(w' \in d) + \mu}$$

where smoothing parameter $\mu$ controls the degree of reliance on relative frequencies in the document corpus rather than on the counts in $d$. The initial ranker that we choose to use later in the experiment computes the KL divergence between the $MLE_q w$ and a modified version of $DIR_d w$ (Zhai and Lafferty, 2001).

Both estimations can be easily extended to distributions over text sequences by assuming that the terms are independent:

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

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

In the re-ranking setting, we estimate that the probability of a document $d$ generates $w$, 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 $z$. Similar to MLE and DIR estimations, this could be generated to give a distribution on a sequence of text:

| Collection | Contents | Language | Num of docs | Size | Queries |
|---|---|---|---|---|---|
| BL (CLEF2008) | British Library Data | English (Main) | 1,000,100 | 1.2 GB | 50 |
| BNF (CLEF2008) | Bibliothèque Nationale de France | French (Main) | 1,000,100 | 1.3 GB | 50 |
| LAT (CLEF2007) | Los Angeles Times 2002 | English | 135,153 | 434 MB | 50 |

Table 1. Statistics of test collections

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

Then the distance between a query and a document based on this model can be obtained. The first method we propose here adopts the KL divergence between the query terms and document terms to compute a Re-Rank score $RS_{LDA}^{KL1}$:

$$RS_{LDA}^{KL1} = -D(MLE_q(\cdot)||LDA_d(\cdot))$$

This method also has the property of length-normalization to ameliorate long document bias problems (Kurland and Lee, 2005).

The second method also measures a KL divergence between a query and a document, however, in a different way. As in the original LDA model, the multinomial parameter $\vec{\theta}_d$ indicates the topic distribution of a document $d$. Query $q$ can be considered as topic estimation of a unknown document $\tilde{\vec{w}}$. Thus by first randomly assigning topics to words and then performing a number of loops through the Gibbs sampling update, we have:

$$p(z_i|\tilde{\vec{z}}_{\neg i}, \tilde{\vec{w}}; \vec{z}_{\neg i}, \vec{w}) = \frac{n_{z,\neg i}^{w_i} + \tilde{n}_{z,\neg i}^{w_i} + \beta_{w_i}}{(\sum_{v=1}^{V} n_z^v + \tilde{n}_z^v + \beta_v) - 1} \cdot \frac{n_{\tilde{d}_i,\neg i}^{z} + \alpha_k}{(\sum_{z=1}^{k} n_{\tilde{d}_i}^z + \alpha_z) - 1}$$

where $\tilde{n}_{z,\neg i}^{w_i}$ counts the observations of word $w_i$ and topic $k$ in unseen document. Then the topic distribution for the query (just the unseen document $\tilde{d}_i$) is:

$$\tilde{\theta}_{\tilde{d}_i,k} = \frac{n_{\tilde{d}_i}^k + \alpha_k}{\sum_{z=1}^{k} n_{\tilde{d}_i}^z + \alpha_z}$$

so that the distance between a query $q$ and a document $d$ is defined as the KL divergence between the topic distributions of $q$ and $d$. Then the re-ranking score is calculated as:

$$RS_{LDA}^{KL2} = -D(\vec{\theta}_q||\vec{\theta}_d)$$

Thus we can re-rank the initial retrieved documents according to the scores acquired. However, as in other topic models, a topic in the LDA model represents a combination of words, and it may not be as precise a representation as words in language model. Hence we need to further consider how to combine initial retrieval scores with the re-ranking scores calculated. Two combination methods will be presented in the next subsection.

### 3.4 Combining Initial Retrieval Scores

Motivated by the significant improvement obtained by (Wei and Croft, 2006) and (Zhang et al., 2005), we formulate our method through a linear combination of the re-ranking scores based on initial ranker and the latent document re-ranker, shown as follow:

$$RS1 = (1-\lambda) \cdot OS + \lambda \cdot RS_{LDA}^{KL};$$

where $OS$ denotes original scores returned by the initial ranker and $\lambda$ is a parameter that can be tuned with $\lambda = 0$ meaning no re-ranking is performed.

Another scheme considers a multiplication combination to incorporate the original score. It does not need to tune any parameters:

$$RS2 = OS \cdot RS_{LDA}^{KL};$$

This concludes our overview of the proposed latent re-ranking method.

## 4 Evaluation

In this section, we will empirically study the effectiveness of the latent document re-ranking method over three different data collections.

### 4.1 Experimental Setup

**Data** The text corpus used in our experiment was made up from elements of the CLEF-2007 and CLEF-2008 the European Library (TEL) collections[1] written in English and French. These collections are described in greater detail in Table 1. All of the documents in the experiment were indexed using the Lemur toolkit[2]. Prior to

[1] http://www.clef-campaign.org
[2] http://www.lemurproject.org

indexing, Porter's stemmer and a stopword list[3] were used for the English documents. We use a French analyzer[4] to analyze French documents.

It is worth noting that the CLEF-2008 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 main languages of those test collections (BL-English, BNF-French). Furthermore, documents written in different languages tend not to match the queries in main languages. Also the data is very different from the newspaper articles and news agency dispatches previously used in the CLEF as well as TREC[5]. The data tends to be very sparse. Many records contain only title, author and subject heading information; other records provide more detail. The average document lengths are 14.66 for BL and 24.19 for BNF collections after pre-processing, respectively. Please refer to (Agirre et al., 2008) for a more detailed discussion about this data. The reason we choose these data collections is that we wanted to test the scalability of the proposed method in different settings and over different guages. In addition we also select a more tional collection (LAT from CLEF2007) as a test base.

We also used the CLEF-2007 and CLEF-2008 query sets. The query sets consist of 50 topics in English for LAT, BL and in French for BNF, all of which were used in the experiment. Each topic is composed of several parts such as: *Title*, *Description*, *Narrative*. We chose to conduct *Title+Description* runs as queries. The queries are processed similarly to the treatment in 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.

We compare the proposed latent re-ranking method with four other approaches: the initial ranker, mentioned above, is a KL-divergence retrieval function using the language models. Three other baseline systems are: Kurland and Lee's structural re-ranking approach (Recursive Weighted Influx + Language Model), chosen as it demonstrates the best performance in their paper (Kurland and Lee, 2005), Zhang et al.'s affinity graph-based approach (Zhang et al., 2005) and a variant of Kurland and Lee's work with links in the graph calculated by the vector-space model (cosine similarity as mentioned in (Kurland and Lee, 2005)). We denote these four systems as InR, RWILM, AFF, and VEC respectively. Furthermore, we denote the permutations of our methods as follows: LDA1: $RS2\ with\ RS_{LDA}^{KL1}$, LDA2: $RS1\ with\ RS_{LDA}^{KL1}$, LDA3: $RS2\ with\ RS_{LDA}^{KL2}$, LDA4: $RS1\ with\ RS_{LDA}^{KL2}$.

Because the inconsistency of the evaluation metrics employed in the past work, we choose to employ all of them to measure the effectiveness of various approaches. These include: mean average precision (MAP), the precision of the top 5 documents (Prec@5), the precision of the top 10 documents (Prec@10), normalized discounted cumulative gain (NDCG) (Jarvelin and Kekalainen, 2002) and Bpref (Buckley and Voorhees, 2004). Statistical-significant differences in performance were determined using a paired t-test at a confidence level of 95%.

It is worth pointing out that the above measurements are not directly comparable with those of the CLEF participants because we restricted our initial pool to a smaller number of documents and the main purpose in the paper is to compare the proposed method with different baseline systems.

**Parameter** Two primary parameters need to be determined in our experiments. For the re-ranking experiments, the combination parameter $\lambda$ must be defined. For the LDA estimation, the number of topics $k$ must be specified. We optimized settings for these parameters with respect to MAP, not with all other metrics over the BL collection and apply them to all three collections directly.

The search ranges for these two parameters were:

$\lambda$ : 0.1, 0.2, …, 0.9
$k$ : 5, 10, 15, …, 45

As it turned out, for many instances, the optimal value of $\lambda$ with respect to MAP was either 0.1 or 0.2, suggesting the initial retrieval scores have valuable information inside them. In contrast, the optimal value of $k$ was between 20 and 40. Although this demonstrates a relatively large variance, the differences in terms of MAP have remained small and statistically insignificant. We set $\mathbb{D}_{init}$ to 50 in all results reported, as in Kurland and Lee's paper (Kurland and Lee, 2005) and we later show that the performance turns out to be very stable when this set enlarged.

[3] ftp://ftp.cs.cornell.edu/pub/smart/
[4] http://lucene.apache.org/
[5] http://trec.nist.gov/

| | BL | | | | |
|---|---|---|---|---|---|
| | MAP | Prec@5 | Prec@10 | NDCG | Bpref |
| InR | 0.1913 | 0.52 | 0.452 | 0.3489 | 0.2287 |
| RWILM | **0.2152** | 0.532 | 0.468 | 0.3663 | 0.2242 |
| AFF | 0.1737 | 0.444 | 0.434 | 0.3273 | 0.22 |
| VEC | 0.1756 | 0.448 | 0.434 | 0.3258 | 0.2216 |
| LDA1 | 0.21 o, a, v | *0.544 a, v* | *0.47* | *0.3679 o, a, v* | *0.2429 a, v* |
| LDA2 | 0.2148 o, a, v | ***0.58 o, a, v*** | ***0.5 o, a, v*** | ***0.3726 o, a, v*** | ***0.2491 o, l, a, v*** |
| LDA3 | 0.1673 | 0.452 | 0.402 | 0.3297 | 0.2 |
| LDA4 | 0.2035 o, a, v | *0.548 a, v* | 0.468 a, v | 0.3626 o, a, v | *0.2326 a* |
| | BNF | | | | |
| | MAP | Prec@5 | Prec@10 | NDCG | bpref |
| InR | 0.1266 | 0.268 | 0.216 | 0.2456 | 0.1482 |
| RWILM | 0.1274 | 0.264 | 0.218 | 0.2495 | 0.1498 |
| AFF | 0.108 | 0.248 | 0.21 | 0.2221 | 0.1404 |
| VEC | 0.1126 | 0.252 | 0.214 | 0.2262 | 0.1463 |
| LDA1 | *0.1374 a, v* | *0.292 a* | *0.242* | *0.2544 a, v* | *0.1617* |
| LDA2 | ***0.1452 o, a, v*** | ***0.292 a, v*** | *0.244 a* | ***0.2608 o, a, v*** | ***0.1697 o, l, a, v*** |
| LDA3 | 0.1062 | 0.232 | 0.202 | 0.2226 | 0.1439 |
| LDA4 | *0.1377 a,v* | *0.28 a* | ***0.246 o, a, v*** | *0.2507 a, v* | *0.1672 o, a, v* |
| | LAT02 | | | | |
| | MAP | Prec@5 | Prec@10 | NDCG | bpref |
| InR | 0.3119 | 0.568 | 0.48 | 0.5093 | 0.3105 |
| RWILM | 0.3097 | 0.556 | 0.478 | 0.5096 | 0.3064 |
| AFF | 0.3065 | 0.572 | 0.492 | 0.5037 | 0.312 |
| VEC | 0.301 | 0.536 | 0.474 | 0.4975 | 0.3087 |
| LDA1 | *0.3253 v* | ***0.584 v*** | ***0.502 v*** | *0.5158 v* | *0.3339 o, l, v* |
| LDA2 | *0.3271 a, v* | ***0.584 o, v*** | *0.496* | *0.518 o, v* | *0.3351 o, l, a, v* |
| LDA3 | 0.2848 | 0.444 | 0.398 | 0.486 | 0.2879 |
| LDA4 | ***0.3274 o*** | 0.552 | 0.478 | ***0.5202 o, v*** | ***0.3396 o, l, v*** |

Table 2. Experimental Results. For each evaluation setting, improvements over the RWILM baseline are given in italics (because it has highest performance); statistically significant differences between our methods and InR, RWILM, AFF, VEC are indicated by o, l, a, v, respectively. Bold highlights the best results over all algorithms.

Lastly, the parameters in the baseline systems are set according to the tuning procedures in their original papers[6].

[6] More specifically, the combination parameter was set to 0.5 for AFF, the number of links was set to 4 for RWILM.

### 4.2 Results

**Primary Evaluation** The main experimental results are presented in Table 2. The first four rows in each collection specify reference-comparison data. The first question we are interested in is how our latent re-ranking methods

perform (taken as a whole). It is shown that our methods bring improvements upon the various baselines in 75% of the 48 relevant comparisons (4 latent re-ranking methods × 4 corpora × 4 baselines). Only the algorithm permutation LDA3 performs less well. Furthermore, our methods are able to achieve the highest performance across all the evaluation metrics over three test collections except in one case (MAP in BL collection). An even more exciting observation is that in many cases, our methods, even though tuned for MAP, can outperform various baselines for all the evaluation metrics, with statistically significant improvements in many runs.

A closer examination of the results in Table 2 reveals some interesting properties. As expected, the RWILM method bought improvements in many cases in CLEF-2008 test collections. However, the performance over CLEF-2007 collection was somewhat disappointing. This seems to indicate that the language model induced graph method tends to perform better in sparse data rather than longer documents. Also Language Modeling requires large set training data to be effective, while the complexity of our method is only linear with number of topics and the number of documents for each iteration. The affinity and vector graph based methods demonstrated poor performance across all the collections. This may be due to the fact that the approach Zhang et al. (Zhang et al., 2005) developed focuses more on diversity and information richness and cares less about the precision of the retrieval results while asymmetric graph as constructed by the vector space model fails in capturing important relationship between the documents.

Another observation we can draw from Table 2 is that the relative performance tends to be stable during test collections written in different languages. This shows a promising future for studying structure of the documents with respect to queries for re-ranking purpose. At the same time, efficiency is always an issue in all re-ranking methods. Although this is not a primary concern in the current work, it would definitely worth thinking in the future.

We also conducted some experiments over queries constructed by using *Title* field only. This forms some more realistic short queries. The experiments showed very similar results compared to longer queries. This demonstrates that the query length is a trivial issue in our methods (as in other graph-based structural re-ranking). We examined the best and worse performed queries, their performance are generally consistent across all the methods. This phenomenon should be investigated further in the follow up evaluation.

**Comparison of Different Methods** In comparison of performance between four permutations of our methods, LDA2 is the clear winner over CLEF-2008 test collections. The results obtained by LDA2 and LDA4 over CLEF-2007 test collection were mixed. LDA2 performed better in precision at top $n$ documents while LDA4 showed promising results in terms of more general evaluation metrics. On the other hand, the linear combination approach performed much better than multiplication based combination. The situation is even worse when we adopted the $RS_{LDA}^{KL2}$ method, which was inferior in several cases. Thus the linear combination should be highly recommended.

**Scalability** We have shown that our latent document re-ranking method is successful at accomplishing the goal of improving the results returned by an initial retrieval engine. But one may raise a question of whether it is necessary to restrict our attention to an initial pool $\mathbb{D}_{init}$ at such a small size. As it happens, preliminary experiments with LDA2 on larger size of the initial pool are presented in Figure 1. As we can see, our method can bring consistently stable improvements.

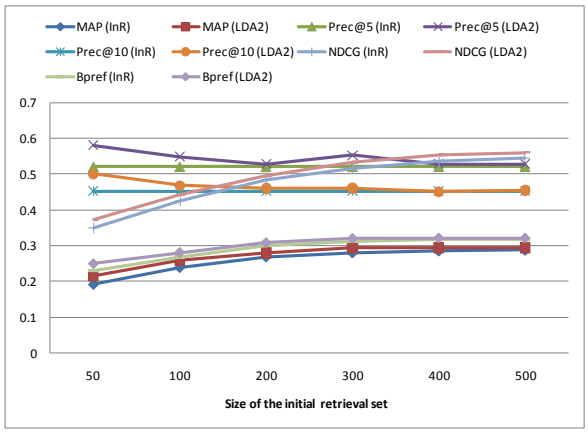


Figure 1. Experiments with larger initial pools

## 5 Conclusion and Future Work

In this paper we proposed and evaluated a latent document re-ranking method for re-ordering the initial retrieval results. The key to refine the results is finding the latent structure of "topics" or "concepts" in the document set, which leve-

rages the latent Dirichlet allocation technique for the query-dependent ranking problem and results in state-of-art performance.

There are many research directions we are planning to investigate. It has been shown that LDA-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 the comparison between our methods with other related approaches, such as Kurland and Lee's cluster-based approach (Kurland and Lee, 2006).

There exist a sufficient number of latent semantic techniques such as singular vector decomposition, non-negative matrix factorization, PLSA, etc. We are planning to explore these methods to compare their performance. Also direct re-ranking can be used to improve automatic query expansion since better ranking in top retrieved documents can be expected to improve the quality of the augmented query. We believe this is another fruitful line for future research.

**Acknowledgments**

The authors would like to thank three anonymous reviewers for 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**.**

## References

Eneko Agirre, Giorgio M. Di Nunzio, Nicola Ferro, Thomas Mandl and Carol Peters (2008). CLEF 2008: Ad Hoc Track Overview. In *Working notes of CLEF2008*.

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

Jaroslaw Balinski and Czeslaw Daniowicz (2005). "Re-ranking method based on inter-document distances." *Inf. Process. Manage.* **41**(4): 759-775.

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

Sergey Brin and Lawrence Page (1998). "The anatomy of a large-scale hypertextual Web search engine." *Comput. Netw. ISDN Syst.* **30**(1-7): 107-117.

Chris Buckley and Ellen M. Voorhees (2004). Retrieval evaluation with incomplete information. In *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, Sheffield, United Kingdom, ACM. p. 25-32.

Hongbo Deng, Michael R. Lyu and Irwin King (2009). Effective latent space graph-based re-ranking model with global consistency. In *Proceedings of the Second ACM International Conference on Web Search and Data Mining*, Barcelona, Spain, ACM. p. 212-221.

Fernando Diaz (2005). Regularizing ad hoc retrieval scores. In *Proceedings of the 14th ACM international conference on Information and knowledge management*, Bremen, Germany, ACM. p. 672-679.

Thomas L. Griffiths and Mark Steyvers (2004). Finding scientific topics. In *Proceeding of the National Academy of Sciences*. p. 5228-5235.

Thomas Hofmann (1999). Probabilistic latent semantic indexing. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, Berkeley, California, United States, ACM. p. 50-57.

Kalervo Jarvelin and Jaana Kekalainen (2002). "Cumulated gain-based evaluation of IR techniques." *ACM Trans. Inf. Syst.* **20**(4): 422-446.

Jaap Kamps (2004). Improving Retrieval Effectiveness by Reranking Documents Based on Controlled Vocabulary In *Proceedings of 26th European Conference on IR Research, ECIR 2004*, Sunderland, UK. p. 283-295.

Jon M. Kleinberg (1999). "Authoritative sources in a hyperlinked environment." *J. ACM* **46**(5): 604-632.

S. Kullback and R. A. Leibler (1951). "On Information and Sufficiency." *The Annals of Mathematical Statistics* **22**(1): 79-86.

Oren Kurland 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 on Research and development in information retrieval*, Salvador, Brazil, ACM. p. 306-313.

Oren Kurland 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 on Research and development in information retrieval*, Seattle, Washington, USA, ACM. p. 83-90.

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

Kyung-Soon Lee, Young-Chan Park and Key-Sun Choi (2001). "Re-ranking model based on document clusters." *Inf. Process. Manage.* **37**(1): 1-14.

Robert W. P. Luk and K.F. Wong (2004). Pseudo-Relevance Feedback and Title Re-ranking for Chinese Information Retrieval. In *Working Notes of the Fourth NTCIR Workshop Meeting*, Tokyo, Japan, National Institute of Informatics.

Qiaozhu Mei, Xu Ling, Matthew Wondra, Hang Su and ChengXiang Zhai (2007). Topic sentiment mixture: modeling facets and opinions in weblogs. In *Proceedings of the 16th international conference on World Wide Web*, Banff, Alberta, Canada, ACM. p. 171-180.

Xuan-Hieu Phan, Le-Minh Nguyen and Susumu Horiguchi (2008). Learning to classify short and sparse text \& web with hidden topics from large-scale data collections. In *Proceeding of the 17th international conference on World Wide Web*, Beijing, China, ACM. p. 91-100.

Jay M. Ponte and W. Bruce Croft (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*, Melbourne, Australia, ACM. p. 275-281.

Youli Qu, Guowei Xu and Jun Wang (2001). Rerank Method Based on Individual Thesaurus. In *Proceedings of the Second NTCIR Workshop on Research in Chinese & Japanese Text Retrieval and Text Summarization*, Tokyo, Japan, National Institute of Informatics.

Ivan Titov and Ryan McDonald (2008). Modeling online reviews with multi-grain topic models. In *Proceeding of the 17th international conference on World Wide Web*, Beijing, China, ACM. p. 111-120.

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

Jinxi Xu and W. Bruce Croft (2000). "Improving the effectiveness of information retrieval with local context analysis." *ACM Trans. Inf. Syst.* **18**(1): 79-112.

Chengxiang Zhai and John Lafferty (2001). Model-based feedback in the language modeling approach to information retrieval. In *Proceedings of the tenth international conference on Information and knowledge management*, Atlanta, Georgia, USA, ACM. p. 403-410.

Chengxiang Zhai and John Lafferty (2004). "A study of smoothing methods for language models applied to information retrieval." *ACM Trans. Inf. Syst.* **22**(2): 179-214.

Benyu Zhang, 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 on Research and development in information retrieval*, Salvador, Brazil, ACM. p. 504-511.