Promoting Ranking Diversity for Biomedical Information Retrieval based on LDA

Yan Chen*†, Xiaoshi Yin*†, Zhoujun Li*†, Xiaohua Hu§ and Jimmy Xiangji Huang¶
*State Key Laboratory of Software Development Environment, Beihang University, Beijing 100191, China
†School of Computer Science and Engineering, Beihang University, Beijing, China
§College of Information Science and Technology, Drexel University, Philadelphia, PA, USA
¶School of Information Technology, York University, Canada
chenyan@cse.buaa.edu.cn, xiaoshiyin@cse.buaa.edu.cn, lizj@buaa.edu.cn, xiaohua.hu@ischool.drexel.edu, jhuang@yorku.ca

***Abstract*—In this paper, we propose an approach based on a topic generative model called Latent Dirichlet Allocation (LDA) to promoting ranking diversity for biomedical information retrieval. Different from other approaches or models which consider aspects on word level, our approach assumes that aspects should be identified by the topics of retrieved documents. We present LDA model to discover topic distribution of retrieval passages and word distribution of each topic dimension, and then re-rank retrieval results with topic distribution similarity between passages based on N-size slide window. Experiments on TREC 2007 Genomics collection and two distinctive IR baseline runs demonstrate the effectiveness of our method in promoting ranking diversity for biomedical information retrieval. Evaluation results show that our approach can achieve 8% improvement over the highest Aspect MAP reported in TREC 2007 Genomics track.**

***Index Terms*—ranking diversity, biomedical IR, LDA**

I. Introduction

For biomedical information retrieval, there are immense data and tremendous increase of genomics and biomedical relevant publications. The wealth of information has led to an increasing amount of interest in and need for applying information retrieval techniques to access the scientific literature in genomics and related biomedical disciplines. In many cases, the desired information of a question (query) asked by biologists is a list of a certain type of entities covering different aspects that are related to the question [1], such as cells, genes, diseases, proteins, mutations, etc. Hence, it is important of a biomedical IR system to be able to provide relevant and diverse answers to fulfill biologists' information needs. In recent years, the "aspect retrieval" was proposed in TREC Genomics tracks. Aspects of a retrieved passages could be a list of named entities or MeSH terms [2], representing answers that cover different portions of a full answer to the query. Aspect Mean Average Precision (MAP) [2] was defined in the Genomics tracks. Its purpose is to study how a biomedical retrieval system can support a user to gather information about different aspects of a query. Biomedical retrieval system should return relevant information at the passage level. Relevant passages that do not contribute any new aspects will not be used to accumulate Aspect MAP. Therefore, Aspect MAP is a measurement for both relevance and diversity of an IR ranked list.

There has been several research focused on promoting ranking diversity in recent years. Perhaps the most representative method is maximum marginal relevance (MMR) [3], as well as mixture models [4], subtopic diversity [5], and others. The basic idea of above three methods is to penalize redundancy by lowering an item's rank if it is similar to the items already ranked. However, these methods often treat relevance ranking and diversity ranking separately, and sometimes with heuristic procedures. Rianne Kaptein *et al.* [6] employed a top down sliding window to diversify ranked list of retrieved documents. A recent study concerning on the Genomics aspect retrieval was conducted by Huang *et al.* [7] and Yin *et al.* [8]. A side effect of these three re-ranking strategies is that they favor long documents, as the long documents tend to contain more distinct terms. Zhu *et al.* [9] proposed a clustering-based ranking algorithm called GRASSHOPPER to promote ranking diversity in biomedical retrieval domain. Unfortunately, this re-ranking method would reduce their system's performance and decrease the Aspect MAP of the original results for the genomics aspect retrieval [10].

However, the previous work considers the aspects of user query and retrieved documents mainly on word level. In other words, one word or more co-occurrence words are used to identify a specific aspect. This assumption could cause two problems: firstly, one or more co-occurrence words in a passage are used to identify the aspect. However, it is common sense for us that a specific word can express more than one latent topics according to different contexts in a passage; secondly, words in a passage are considered as independent to each other. However, some potential relationships between words might exist. Therefore, it is insufficient to identify aspect on word level.

In this paper, we aim at addressing both above problems. We propose an approach which employs Latent Dirichlet Allocation (LDA) [11], a topic generative model, to promote diversity in the ranked list for biomedical information retrieval. Experiments conducted on TREC 2007 Genomics track collection and two very different IR baseline runs demonstrate the effectiveness of our approach. The evaluation results show that our approach can achieve 8% improvement over the highest Aspect MAP reported in TREC 2007 Genomics track.

The rest of this paper is organized as follows. In Section 2,

978-0-7695-4574-5/11 $26.00 © 2011 IEEE
DOI 10.1109/BIBM.2011.28

we present our motivation of using LDA model and introduce two re-ranking strategies. The experimental results of applying two algorithms to TREC 2007 Genomics track collections are presented in Section 3. We conclude our work and discuss the possible further research directions in Section 4.

II. Reranking Strategies based on LDA

We assume that aspects of retrieved passages should be identified by latent topics [1] hidden in passages which are considered to be more abstract, and latent topics can be identified by word distribution. In this section, we will expound how we use a particular generative model called LDA to discover the topics covered by retrieved passage collection, and illustrate how these topics can be used to improve ranking diversity.

A. Aspect Discovery and Transformation

1) Aspect Discovery using LDA: Discovering aspects covered by each retrieved passage is the first step for re-ranking and we employ LDA for aspect discovery [2]. Its basic idea is that documents are represented as random mixtures over latent topics, where each topic is characterized by a distribution over words.

The LDA model is represented (using plate notation) as a probabilistic graphical model in Figure 1. It can be seen clearly

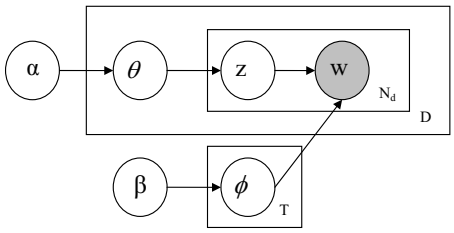


Fig. 1: LDA Model

from the figure that the LDA representation has three levels and the generation of a document collection is modeled as a three-step process. First, for each document, a distribution over topics is sampled from a Dirichlet distribution. Second, for each word in the document, a single topic is chosen according to this distribution. Finally, each word is sampled from a multinomial distribution over words specific to the sampled topic. In this model, ϕ denotes the matrix of topic distributions, with a multinomial distribution over N_d word items for each of T topics being drawn independently from a symmetric Dirichlet(β) prior. θ is the matrix of document-specific mixture weights for these T topics, each being drawn independently from a symmetric Dirichlet(β) prior. For each word, z denotes the topic responsible for generating that word, drawn from the θ distribution for that document, and w is the word itself, drawn from the topic distribution ϕ corresponding to z. N_d stands for the number of words in the document. D stands for the size of document collection. Estimating ϕ and θ provides information about the topics in a collection and the weights of those topics in each document. A host of algorithms have been used to estimate these parameters, ranging from Mean field variational methods [11], Expectation propagation [15], Gibbs sampling [16], Collapsed variational inference [17] to Fast Collapsed Gibbs Sampling [18].

2) Aspect Distribution Transformation: We construct θ matrix of Eq.(1) in light of LDA model discussed in above subsection, which is the matrix of passage-specific mixture weights for these T aspects discovered. θ provides the information about the aspects in the retrieved passage collection and the weights of those aspects in each retrieved passage. θ_i denotes the aspects distribution for each passage P_i. a_{ij} stands for the weight of the aspect A_j given the passage P_i such that $\sum_{j=1}^{T} a_{ij}$ equals one.

$$\begin{aligned}\theta &= (\theta_1, \theta_2, \cdots, \theta_D)^T \\ &= \begin{pmatrix} a_{11} & \cdots & a_{1j} & \cdots & a_{1T} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ a_{i1} & \cdots & a_{ij} & \cdots & a_{iT} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ a_{D1} & \cdots & a_{Dj} & \cdots & a_{DT} \end{pmatrix} \\ &\quad (1 \le i \le D, 1 \le j \le T)\end{aligned} \tag{1}$$

We have observed the following two interesting phenomena from the column of matrix θ. First, for some specific aspects, majority passages of the retrieved collection get large weight values; however, for some other specific aspects, a few passages get large weight values. Second, even the same weight value in different columns of θ matrix would have a different importance for different aspects. Therefore, we tend to make transformation of θ matrix to represent the importance of each passage in each aspect.

Given such a hypothesis that for each aspect, the importance of retrieved passages is an normal distribution, we can have T normal distributions, denoting by $N = (N_1, N_2, \cdots, N_T)$. Given an normal distribution N_i $(1 \le i \le T)$, mean μ_i and variance σ_i are referred to Eqs.(2) and (3) respectively.

$$\mu_i = \frac{\sum_{j=1}^{D} a_{ji}}{D} \tag{2}$$

$$\sigma_i = \frac{\sum_{j=1}^{D} (a_{ji} - \mu_i)^2}{D} \tag{3}$$

where a_{ji} stands for the weight of the aspect A_i for passage P_j, and D denotes the number of retrieved passages. In addition, we get a new matrix Θ shown as Eq.(4) to measure the passage importance for each aspect.

$$\begin{aligned}\Theta &= (\Theta_1, \Theta_2, \cdots, \Theta_T)^T \\ &= \begin{pmatrix} N_1(a_{11}) & \cdots & N_1(a_{j1}) & \cdots & N_1(a_{D1}) \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ N_i(a_{1i}) & \cdots & N_i(a_{ji}) & \cdots & N_i(a_{Di}) \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ N_T(a_{1T}) & \cdots & N_T(a_{jT}) & \cdots & N_T(a_{DT}) \end{pmatrix} \\ &\quad (1 \le i \le T, 1 \le j \le D)\end{aligned} \tag{4}$$

where $N_i(a_{ji})$ denotes the importance of passage P_j for the aspect A_i. Θ_i denotes the importance distribution of passage collection for the aspect A_i.

[1] We use "topic" and "aspect" interchangeable in the rest of this paper.

[2] Apart from text modeling, LDA has been used in many other applications such as computer vision [12], image modeling [13], social tagging system [14], etc.

3) Re-ranking with N-Size Slide Window: We define the re-ranking problem as this: Given a query q and an initial ranking R produced for this query only with respect to relevance, we build a new ranking S taking account of both relevance and diversity. In this section, we introduce two re-ranking algorithms based on a slide window with size N, and put top N passages from R as candidate passages into the slide window when re-ranking. As we commonly set N with a small number, we suppose that there is no distinctive difference between passages in a slide window with respect to their query-relevance.

First, we choose a passage from the slide window as the first passage in ranking S, which contains the largest aspect coverage as show in Eq.(5).

$$MaxAspCoverg = \arg\max_{q\in[1,N]} \sum_{t=1}^{T} N_t(a_{tq}) \tag{5}$$

where $N_t(a_{tq})$ denotes the importance of passage P_q for the aspect A_t and $\sum_{t=1}^{T} N_t(a_{tq})$ stands for the aspect coverage of passage P_q. After adding this passage into ranking S, we remove it from ranking R. For the rest of passages in R, if the number of passages in R is not less than N, we will put the top N passages in R into the slide window, or else we will put all the passages in R into slide window. Then we choose a passage from the slide window, which contains the most distinctive aspects compared with the observed passages in ranking S, add it into S, and remove it from R. The working scheme of this ranking method based on N size slide window is described in Algorithm 1, named $rank\_NWin$.

Algorithm 1 rank_NWin Algorithm

Input: An initial passage ranking R produced for current user query only with respect to relevance, and the size N of the slide window
Output: A reranked passage list S
Process:
Given top N passages in R, we find a passage $pass_1$ containing the most aspect coverage value using Eq.(5);
$R \leftarrow R \setminus \{pass_1\}$;
$S \leftarrow \emptyset \cup \{pass_1\}$;
while $R.length \neq 0$ **do**
  Choose top N passages in R as candidate passages and if the length of rank R is less than N, take all passages in R as candidate passages;
  for each passage i of candidate passages **do**
    $distance\_R_i = 0$;
    for each passage j in S **do**
      $distance\_R_i = distance\_R_i + Distance(R_i, S_j)$;
    end for
    $distance\_R_i = distance\_R_i / S.length$;
  end for
  Find the max $distance_R$ passage $pass_{rest}$ in candidate passages;
  $R \leftarrow R \setminus \{pass_{rest}\}$;
  $S \leftarrow S \cup \{pass_{rest}\}$;
end while
return S.

The advantage of the Algorithm 1 is that it considers aspect distinctions between candidate passages in the slide window and observed passages ranked in S.

However, considering original query-relevance ranking R, it is not appropriate for Algorithm 1 to change in a wide range of R. Therefore, another algorithm named $rank\_NWin\_Group$ is proposed to ensure that the new ranking S is just the original ranking R with slight adjustments. The key idea of this algorithm is described below. For the first passage in S, we still choose a passage containing the largest aspect coverage from the slide window, add it into S and remove it from R. Different from $rank\_NWin$ Algorithm, we first group rank R into several N size groups, and the size of last group may be less than N. We put each group into the slide window in turn, re-rank the passages in current group, and add them into S finally. The process of re-ranking in groups is similar to algorithm $rank\_NWin$. Algorithm 2 describes the process of re-ranking by using N-size slide window to group ranking R.

$Distance(i, j)$ in algorithms $rank\_NWin$ and

Algorithm 2 rank_NWin_Group Algorithm

Input: An initial passage ranking R produced for current user query only with respect to relevance, and the size N of the slide window
Output: A reranked passage list S
Process:
Given top N passages in R, we find a passage $pass_1$ containing the most aspect coverage value using Eq.(5);
$R \leftarrow R \setminus \{pass_1\}$;
$S \leftarrow \emptyset \cup \{pass_1\}$;
Group passages in R into $\lceil R.length/N \rceil$ groups;
for each group i **do**
  for each passage j in group$i$ **do**
    $distance\_R_j = 0$;
    for each passage k in S **do**
      $distance\_R_j = distance\_R_j + Distance(R_j, S_k)$;
    end for
    $distance\_R_j = distance\_R_j / S.length$;
  end for
  Rank passages in group i according to $distance_R$ in a descend order.
  $R \leftarrow R \setminus \{pass\ in\ group\ i\}$;
  $S \leftarrow S \cup \{pass\ in\ group\ i\}$;
end for
return S.

$rank\_NWin\_Group$ is the measurement of the aspect distinction between two passages. In our work, we use two slightly different ways to evaluate it. The first one can be seen as the original Euclidean distance as shown in Eq.(6).

$$Distance(i, j) = \sqrt{\sum_{t=1}^{T} \left(N_t(a_{ti}) - N_t(a_{tj})\right)^2} \quad (i \neq j) \tag{6}$$

Furthermore, we assume that the importance of each aspect is different, and regard $\mu_t (1 \leq t \leq T)$ defined in the last subsection as the weight value and get another equation for Euclidean distance as shown in Eq.(7).

$$Distance(i, j)^* = \sqrt{\sum_{t=1}^{T} \mu_t (N_t(a_{ti}) - N_t(a_{tj}))^2} \quad (i \neq j) \tag{7}$$

III. Experimental Results

A. Dataset and Evaluation Metrics

We employ TREC 2007 Genomics track collection as the test data set, which is a full-text biomedical collection consisting of 162,259 documents from about 49 genomics-related journals. There are 36 official topics [3] for the track in 2007, which are in the form of questions. These questions usually cover one or more aspects contained in the literature collection (i.e. one or more answers to each question). The followings are examples of queries from the 2007 Genomics Track:

- Query 200: What serum [PROTEINS] change expression in association with high disease activity in lupus?
- Query 221: Which [PATHWAYS] are mediated by CD44?
- Query 231: What [TUMOR TYPES] are found in zebrafish?

For TREC 2007 Genomics track, there are three levels of retrieval performance measured: passage retrieval, aspect retrieval, and document retrieval. Passage MAP, Passage2 MAP, Aspect MAP and Document MAP, defined in [1] and [2], are four evaluation metrics corresponding to the three levels of retrieval performance. In this paper, we mainly focus on two evaluation metrics, Aspect MAP and Passage2 MAP, since our objective is to promote diversity in the ranked list of retrieved passages. Furthermore, aspect retrieval and passage retrieval are also the major tasks in TREC 2007 Genomics tracks.

B. Retrieval Baselines

We employ two retrieval baseline runs, NLMinter [19] and UniNE2 [20]. NLMinter developed by U.S. National Library of Medicine achieved the best performance in TREC 2007 Genomics track in terms of Aspect MAP, Passage2 MAP and Document MAP. UniNE2 which is developed by University of Neuchatel Rue Emile-Argand combined different search strategies. The performance of UniNE2 was above average among all results reported in TREC 2007 Genomics track.

C. Re-ranking Performance

Re-ranking results of the proposed methods on TREC 2007 Genomics collection are shown in Table I. The values in the parentheses are the relative rates of improvement over the original results. It can be seen from the table that our approaches can make improvements over both baseline runs. For the efficiency reason, we re-ranked only the top 100 passages. Distinctive improvements over all baseline runs in terms of Aspect MAP can be observed.

Re-ranking performance is effected by the parameters chosen from LDA model. Comparison with NLMinter baseline run, we only show the re-ranking results with parameters of β whose values are equal to 0.04 in algorithm 1 and 0.06 in algorithm 2, respectively. For UniNE2 baseline run, we show re-ranking results with parameters of β whose values are equal to 0.004 in algorithm 1 and 0.008 in algorithm 2, respectively. The choosing of the parameters in LDA will be discussed in the next subsection.

[3]In TREC Genomics tracks, "topic" means "query" [1] [2]. In the rest of the paper, we use "query" instead of "topic" to avoid the confusion with "aspect".

D. Analysis

1) Impact of Parameter β: The statistical model LDA we have described is conditioned on three parameters, the Dirichlet hyper-parameters α and β, and the number of topics T. The value of β affects the granularity of the model. Retrieved passages can be sensibly factorized into a set of topics at several different scales, and the particular scale of the topics assessed by the model will be set by β. Since we focus on biomedical domain, we tend to employ smaller values of β, which will result in more topics that address specific fields.

We should choose the value of β for each specific user query. In order to improve experimental efficiency, we choose β according to the retrieval passages instead. For NLMinter, we set the values of $\beta \in [0.01, 0.08]$ in steps of 0.01. However, for baseline UniNE2, we set the values of $\beta \in [0.002, 0.009]$ in steps of 0.001. The reason why we give two different settings of β is that the words set size of NLMinter is comparative larger than UniNE2 [4]. These values of β are relatively small and can be expected to give rise to a fine-grained decomposition of the collection into topics that address specific research fields.

2) Impact of Parameters α and T: Given values of β, the problem of choosing appropriate values for α and T thus is a problem of model selection. We let $\alpha T = constant$ to keep constant the sum of the Dirichlet hyper-parameters, which can be interpreted as the number of virtual samples contributing to the smoothing of θ [16]. Moreover, because our strategy in this article is to fix $\alpha T = constant$ and β, and explore the consequences of varying T, for each fixed β value we set the values of T from 10 to 100 in steps of 10 consecutively [5].

Next, we need to choose an appropriate value of T for each specific query. In our case, the data are the words in the retrieved passages, w, and the model is specified by the number of topics, T, thus we wish to compute the likelihood $p(w|T)$. However, this requires to sum over all possible assignments of words to topics z. We can approximate $p(w|T)$ by the harmonic mean of a set of values of $p(w|z, T)$ when z is sampled from the posterior $p(z|w, T)$ [16]. In all cases, $p(w|T)$ increases at the beginning, and decreases after reaching a peak.

Figure 2 shows the log-likelihood of the data for different settings of the number of topics T for query 200, 221 and 231 in our data collection with β being equal to 0.06. For example, for query 200, the results suggest that the data are best accounted for by a model incorporating 50 topics. $p(w|T)$ initially increases as a function of T, reaches a peak at $T = 50$, and then decreases thereafter.

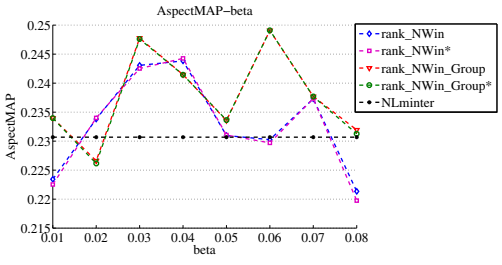
[4]Preprocess give us two word sets of 10,222 words and 2,387 words for retrieved passages by NLMinter and UniNE2 baseline runs, respectively.

[5]Here, we set $constant = 10$ in order to make α not larger than 1. [16].

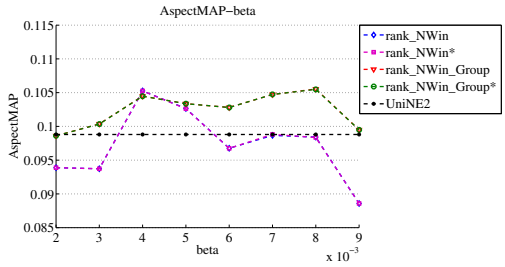
TABLE I: Re-ranking performance with two baseline runs.

NLMinter model					UniNE2 model				
MAP	Aspect	Passage2	Passage	Document	MAP	Aspect	Passage2	Passage	Document
NLMinter	0.23068962	0.07335484	0.05971977	0.20962491	UniNE2	0.09880169	0.01777397	0.05236709	0.13771527
$rank\_NWin$	0.2438342	0.07368625	0.05868155	0.20790886	$rank\_NWin$	0.1052544	0.01946295	0.05459447	0.13969831
	(+5.70%)	(+0.45%)	(-1.74%)	(-0.82%)		(+6.53%)	(+9.50%)	(+4.25%)	(+1.44%)
$rank\_NWin^*$	0.24426998	0.07372402	0.05849706	0.20744464	$rank\_NWin^*$	0.1052544	0.01946007	0.05459788	0.13964510
	(+5.89%)	(+0.50%)	(-2.05%)	(-1.04%)		(+6.53%)	(+9.49%)	(+4.26%)	(+1.40%)
$rank\_NWin\_Group$	0.24908569	0.07792334	0.06151813	0.20976964	$rank\_NWin\_Group$	**0.10554020**	0.01902429	0.05490502	0.14035642
	(+7.97%)	(+6.23%)	(+3.01%)	(+0.07%)		**(+6.82%)**	(+7.03%)	(+4.85%)	(+1.92%)
$rank\_NWin\_Group^*$	**0.24910669**	0.07793161	0.06152586	0.20977025	$rank\_NWin\_Group^*$	0.10549095	0.01902427	0.05490508	0.14035350
	(+7.98%)	(+6.24%)	(+3.02%)	(+0.07%)		(+6.77%)	(+7.03%)	(+4.85%)	(+1.92%)

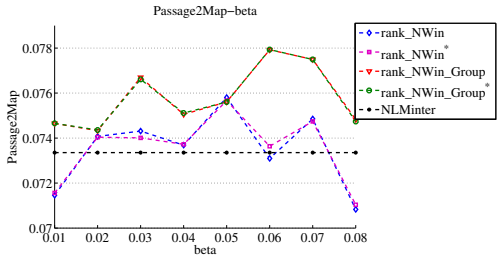
Note: $rank\_NWin$ and $rank\_NWin\_Group$ stand for two reranking approaches with $Distance(i,j)$, $Rank\_NWin^*$ and $rank\_NWin\_Group^*$ denote two reranking approaches with $Distance^*(i,j)$.



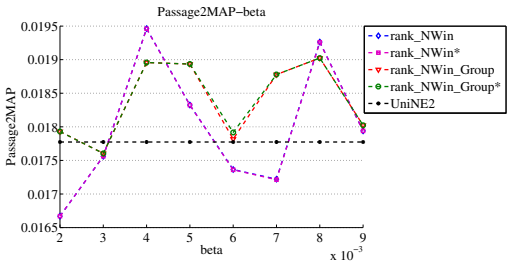
(a) Aspect MAP for NLMinter Run



(b) Aspect MAP for UniNE2 Run



(c) Passage2 MAP for NLMinter Run



(d) Passage2 MAP for UniNE2 Run

Fig. 3: Impact of β Parameter

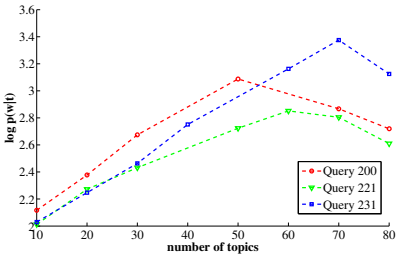


Fig. 2: Model Selection Results

3) Comparison of the Two Reranking Strategies: The parameter β indicates the scale of topics for the retrieved passages. Given different β, retrieved passages can be factorized into a series of topics at different scales. We propose two re-ranking algorithms and two distance metrics, and therefore have four re-ranking algorithms, whose re-raking performance can be also shown in Figure 3. As the aspect level retrieval and the passage level retrieval were two major tasks in the TREC 2007 Genomics tracks, system performances at these two levels with different β are also shown in Figure 3.

Figures 3(a) and 3(b) respectively show NLMinter and UniNE2 system performances at aspect level with different β. It can be seen from Figure 3(a) that when β's value is between 0.03 and 0.07, performance improvements on aspect level can be achieved for all re-ranking strategies. For ranking strategies $rank\_NWin$ and $rank\_NWin^*$, the Aspect MAP increases with the increase of β, reaches at a peak for $\beta = 0.04$, then decreases, and reaches at a local peak when $\beta = 0.07$, and finally it plummets. For $rank\_NWin\_Group$ and $rank\_NWin\_Group^*$, the Aspect MAP increases from $\beta = 0.02$, reaches at a local peak when $\beta = 0.03$, then drops down and jumps to a peak for $\beta = 0.06$, and thereafter falls down. Figure 3(b) shows that retrieval results at aspect level are better than the baseline runs with all βs for $rank\_NWin\_Group$ and $rank\_NWin\_Group^*$. Aspect Map increases as β increases, reaches a local peak when $\beta = 0.004$, and then decreases slightly, after that grows when $\beta = 0.006$, reaches the maximum, and then

drops down quickly. For $rank\_NWin$ and $rank\_NWin^*$, the performance improvements on aspect level are achieved when $\beta = 0.004$ and 0.005.

NLMinter and UniNE2 system performances at passage level with different β are shown in Figures 3(c) and 3(d). Comparing Figures 3(a) and 3(b) with Figures 3(c) and 3(d) respectively, we could observe that the trends of performances on aspect level and passage level are generally in agreement with $rank\_NWin\_Group$ and $rank\_NWin\_Group^*$. The observation illustrates that there are a clear correlation between Aspect MAP and Passage MAP. However, for $rank\_NWin$ and $rank\_NWin^*$, the trends of performances on aspect level is different from passage level. This could be caused by the reason that $rank\_NWin$ and $rank\_NWin^*$ algorithms change original passage ranking within a large ranges. Furthermore, we demonstrate that two different distance metrics, with or without weight, do not influence re-ranking performance significantly.

The comparison results shown in Figure 3 indicate that both of the two proposed re-ranking methods are effective in promoting diversity for biomedical information retrieval, and $rank\_NWin\_Group$ outperforms $rank\_NWin$ in most cases.

IV. Conclusion and Future work

In this paper, we propose an approach which employs LDA, a topic generative model, to promoting ranking diversity for biomedical information retrieval. Our contribution is three-fold. First, to the best of our knowledge, this is the first study of adopting topic model to biomedical IR. Different from other approaches considering aspects on word level, our approach assumes that aspects should be identified by the topics of retrieved documents. We employ LDA model to discover topic distribution of retrieval passages. Second, since retrieved passages' distribution for each aspect is different, even the same weight value in different aspects would be of different importance, we made transformations with topic distribution. Third, two re-ranking algorithms based on "N-size slide window" are proposed, which take both passage novelty and relevance into account. Experiments conducted on TREC 2007 Genomics track collection demonstrate the effectiveness of our approach. The evaluation results show that our approach can achieve 8% improvement over the highest Aspect MAP reported in TREC 2007 Genomics track.

In future research, we intend to extend our work by exploring both more complex models and more sophisticated algorithms and to apply our approach to other test collections, such as ClueWeb09 collection, to investigate whether the approach is still effective for improving ranking diversity in the Web search. Furthermore, ranking diversity plays an important role in a range of tasks or applications, such as social network analysis and recommendation system, etc. We thus plan to further improve our approach to solve the diversification in the above mentioned fields.

Acknowledgment

The authors would like thank anonymous reviewers for their valuable comments and suggestions. This research is supported by the Fund of State Key Laboratory of Software Development Environment (under grant no. SKLSDE-2011ZX-03) and the National Natural Science Foundation of China (under grant no. 61170189).

References

[1] W. Hersh, A. Cohen, L. Ruslen, and P. Roberts, "TREC 2007 Genomics track overview," in *Proc. of TREC-16*, 2007.
[2] W. Hersh, A. Cohen, P. Roberts, and H. Rekapalli, "TREC 2006 Genomics track overview," in *Proc. of TREC-15*, 2006.
[3] J. Carbonell and J. Goldstein, "The use of MMR, diversity-based reranking for reordering documents and producing summaries," in *Proc. of the 21st ACM SIGIR*, 1998.
[4] Y. Zhang, J. Callan, and T. Minka, "Novelty and redundancy detection in adaptive filtering," in *Proc. of the 25th ACM SIGIR*, 2002.
[5] C. Zhai, W. Cohen, and J. Lafferty, "Beyond independent relevance: methods and evaluation metrics for subtopic retrieval," in *Proc. of the 26th ACM SIGIR*, 2003.
[6] R. Kaptein, M. Koolen, and J. Kamps, "Experiments with result diversity and entity ranking: Text, anchors, links, and wikipedia," in *Proc. of TREC-18*, 2009.
[7] X. Huang and Q. Hu, "A bayesian learning approach to promoting diversity in ranking for biomedical information retrieval," in *Proc. of the 32nd ACM SIGIR*, 2009.
[8] X. Yin, X. Huang, and Z. Li, "Promoting ranking diversity for biomedical information retrieval using wikipedia," in *Proc. of the 32nd European Conference on Information Retrieval*, 2010.
[9] X. Zhu, A. Goldberg, J. V. Gael, and D. Andrzejewski, "Improving diversity in ranking using absorbing random walks," in *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proc. of the Main Conference*, 2007.
[10] A. Goldbery, D. A. J. Gael, B. Settles, X. Zhu, and M. Craven, "Ranking biomedical passages for relevance and diversity," in *University of Wisconsin, Madison at TREC Genomics 2006;Proc. of TREC-15*, 2006.
[11] D. Blei, A. Ng, and M. Jordan, "Latent dirichlet allocation," *Journal of Machine Learning Research*, vol. 3, pp. 993–1022, 2003.
[12] F. Li and P. Pietro, "A bayesian hierarchical model for learning natural scene categories," in *Proc. of the 10th IEEE CVPR*, 2005.
[13] C. Lu, X. Hu, X. Chen, J. Park, T. He, and Z. Li, "Probabilistic models for topic learning from images and captions in online biomedical literatures," in *Proc. of the 18th ACM CIKM*, 2009.
[14] X. Chen, C. Lu, Y. An, and P. Achananuparp, "The topic-perspective model for social tagging system," in *Proc. of the 16th KDD*, 2010.
[15] T. Minka and J. Lafferty, "Expectation-propagation for the generative aspect model," in *Proc. of 18th Conference on Uncertainty in Artificial Intelligence*, 2002.
[16] T. Griffiths and M. Steyvers, "Finding scientific topics," in *Proc. of the National Academy of Science*, 2004.
[17] Y. The and D. Newman, "A collapsed variational bayesian inference algorithm for latent dirichlet allocation," in *Proc. of 20th NIPS*, 2006.
[18] I. Porteous, D. Newman, A. Ihler, A. Asuncion, P. Smyth, and M. Welling, "Fast collapsed gibbs sampling for latent dirichlet allocation," in *Proc. of the 14th KDD*, 2008.
[19] D. Demner-Fushman, S. Humphrey, N. Ide, R. Loane, J. Mork, P. Ruch, M. Ruiz, L. Smith, W. Wilbur, and A. Aronsona, "Combining resources to find answers to biomedical questions," in *Proc. of TREC-16*, 2007.
[20] A. Abdou and J. Savoy, "Report on the trec 2006 genomics experiment," in *Proc. of TREC-15*, 2006.