Boosting Novelty for Biomedical Information Retrieval through Probabilistic Latent Semantic Analysis

Xiangdong An, Jimmy Xiangji Huang
Information Retrieval and Knowledge Management Research Lab
School of Information Technology
York University, Toronto, ON M3J 1P3, Canada
{xan, jhuang}@yorku.ca

ABSTRACT

In information retrieval, we are interested in the information that is not only relevant but also novel. In this paper, we study how to boost novelty for biomedical information retrieval through probabilistic latent semantic analysis. We conduct the study based on TREC Genomics Track data. In TREC Genomics Track, each topic is considered to have an arbitrary number of aspects, and the novelty of a piece of information retrieved, called a *passage*, is assessed based on the amount of new aspects it contains. In particular, the aspect performance of a ranked list is rewarded by the number of new aspects reached at each rank and penalized by the amount of irrelevant passages that are rated higher than the novel ones. Therefore, to improve aspect performance, we should reach as many aspects as possible and as early as possible. In this paper, we make a preliminary study on how probabilistic latent semantic analysis can help capture different aspects of a ranked list, and improve its performance by re-ranking. Experiments indicate that the proposed approach can greatly improve the aspect-level performance over baseline algorithm Okapi BM25.

Categories and Subject Descriptors

H.3.3 [**Information Systems**]: Information Search and Retrieval

General Terms

Algorithms, Experimentation

Keywords

Genomics IR, passage retrieval, aspect search

1. INTRODUCTION

Information retrieval (IR) in the context of biomedical databases is characterized by the frequent use of abundant acronyms, homonyms and synonyms. How to deal with the tremendous variants of the same term has been a challenging task in biomedical IR. The Genomics track of Text REtrieval Conference (TREC) provided a common platform to evaluate the methods and techniques proposed by various research groups for biomedical IR. In its last two years (2006 & 2007), the Genomics track focused on the passage retrieval for question answering, where a passage is a piece of continuous text ranging from a phrase up to a paragraph of a document. One of the performances concerned for passage retrieval was the aspect-based mean average precision (MAP) [8]. To evaluate the performance of a ranked list, in 2006, the judges of the competition first identified all relevant passages for each topic from all submissions, and then, based on the content of such relevant passages, assigned a set of Medical Subject Headings (MeSH) terms to each topic as their representative "aspects". In 2007, instead of MeSH terms, the judges picked and assigned terms from the pool of nominated passages deemed relevant to each topic as their "aspects". That is, the "aspects" of a topic in Genomics Track are represented by a set of terms. A passage for a topic is *novel* if it contains aspect terms assigned to the topic which has not appeared in the passages ranked higher. The *novelty* of a ranked list is rewarded by the amount of relevant aspects reached at each rank and penalized by the amount of irrelevant passages ranked higher than novel ones. The aspect-based MAP is an average reflection of novelty retrieval performance on all topics.

For the aspect-level evaluation, the search should reach as many relevant aspects as possible and rank their containing passages as high as possible. "Aspects" are assigned to each topic by the judges only after the submission by all groups, and such aspects are picked only from the nominated passages. At competition, nobody knows how many aspects there exist for each topic in the literature, and what they are. Therefore, "aspects" of each topic in this problem are latent, and it is also not an easy problem to figure out the aspects covered by a passage from its "bag-of-words" representation. However, it is well known that a topic model can represent a document as a mixture of latent aspects. That is, a topic model can convert a document from its "bag-of-words" space to its latent semantic space of a reduced dimensionality. In this paper, we study whether the latent semantic representation would help capture different "aspects" of a passage and further improve the performance of a ranked list by re-ranking. There exist a list of topic models such as Latent Semantic Analysis (LSA)[5], Probabilistic Latent Semantic Analysis (PLSA) [9], and Latent Dirichlet Allocation (LDA) [2]. In this preliminary study, we focus on PLSA. In the future, we would study the problem with both LSA and LDA included.

To the best of our knowledge, this is the first investigation about how well a topic model such as PLSA can help capture hidden aspects in novelty information retrieval. In the investigation, we also examine the hyperparameter settings for PLSA such as initial conditional probabilities and zero estimate smoothing in the context of our problem. Besides standard PLSA model [9], we also examine its variants, e.g. instead of word frequencies, tf-idf weighting is used.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
SIGIR'13, July 28 – August 1, 2013, Dublin, Ireland.
Copyright 2013 ACM 978-1-4503-2034-4/13/07 ...$15.00.

2. RELATED WORK

In information retrieval, ranking based on pure relevance may not be sufficient when the potential relevant documents are huge and highly redundant with each other. In [3, 16, 14, 15, 18], different ways representing and optimizing novelty and diversity of the retrieved documents are studied. The objective is to find the documents that cover as many different aspects (subtopics) as possible while maintaining minimal redundancy. One problem with the novelty (diversity, aspect, subtopic)-based retrieval is how to evaluate the ranking quality. In [14], 3 metrics are introduced: the *subtopic recall* measures the percentage of subtopics covered as a function of rank, the *subtopic precision* measures the precision of the retrieved documents as a function of the minimal rank for a certain subtopic recall, and the *weighted subtopic precision* measures the precision with redundancy penalized based on ranking cost. In [4], a cumulative gain-based metric is proposed to measure the novelty and redundancy, which is also a function of rank.

Most existing methods [3, 16, 14, 15] improve novelty in IR by penalizing redundancy, but they seem not to work well in Genomics aspect search. In the 2006 TREC Genomics track, University of Wisconsin at Madison failed to promote novelty by penalizing redundancy based on a clustering-based approach [7]. In the 2007 TREC Genomics track, most teams simply submitted their relevant passage retrieval results for aspect evaluation such as National Library of Medicine (NLM) [6] and University of Illinois at Chicago [17]. In [10] and [12], a Bayesian learning approach is proposed to find potential aspects for different topics. In [13], a survival model approach is applied to biomedical search diversification with Wikipedia.

3. HIDDEN ASPECT-BASED RE-RANKING

It is well known that PLSA can help to reveal semantic relations between entities of interest in a principled way [9]. In this paper, we consider each retrieved passage d_i ($1 \leq i \leq N$) in $D = \{d_1, ..., d_N\}$ as being generated under the influence of a number of hidden aspect factors $Z = \{z_1, ..., z_K\}$ with words from a vocabulary $W = \{w_1, ..., w_M\}$. Therefore, all passages retrieved initially can be described as an $N \times M$ matrix $T = ((c(d_i, w_j))_{ij}$, where $c(d_i, w_j)$ is the number of times w_j appears in passage d_i. Each row in D is then a frequency vector that corresponds to a passage. Assume given a hidden aspect factor z, a passage d is independent of the word w. Then by Bayes' rule, the joint probability $P(d, w)$ can be obtained as follows:

$$P(d, w) = \sum_{z \in Z} P(z)P(d|z)P(w|z).$$

To explain the observed frequencies in matrix T, we need to find $P(z)$, $P(d|z)$, and $P(w|z)$ that maximize the following likelihood function:

$$L(D, W) = \sum_{d \in D} \sum_{w \in W} c(d, w) log P(d, w).$$

It can be shown that the solution can be achieved by EM algorithm iteratively through the following two alternating steps.

1. By E-step, we calculate the posterior probabilities of the hidden aspect factors:

$$P(z|d, w) = \frac{P(z)P(d|z)P(w|z)}{P(z')P(d|z')P(w|z')}.$$

2. By M-step, we update parameters to maximize the complete data likelihood:

$$P(w|z) = \frac{\sum_{d \in D} c(d, w)P(z|d, w)}{\sum_{d \in D} \sum_{w' \in W} c(d, w')P(z|d, w')},$$

$$P(d|z) = \frac{\sum_{w \in W} c(d, w)P(z|d, w)}{\sum_{d' \in D} \sum_{w \in W} c(d', w)P(z|d', w)},$$

$$P(z) = \frac{\sum_{d \in D} \sum_{w \in W} c(d, w)P(z|d, w)}{\sum_{d \in D} \sum_{w \in W} c(d, w)}.$$

After its convergence, we can calculate the probability of hidden aspect factor z given passage d by

$$P(z|d) = \frac{P(d|z)P(z)}{\sum_{z \in Z} P(d|z)P(z)} \quad \propto \quad P(d|z)P(z) = P(d, z).$$

Hence, we can summarize the aspect trend of each passage d by a normalized factor vector $(P(z_i|d))_{i=1}^{K}$. By this way, we transform the passage representation from the "bag-of-words" space to a lower latent semantic space. We expect this representation would capture the aspect trend of each passage in a better way. All passages can then be clustered based on this vector representation or simply based on their most probable hidden aspect factor

$$z_d = \underset{z_i \in Z}{\operatorname{argmax}} P(z_i|d).$$

With latter, we may sort all passages in each group based on the probability $P(z_d|d)$ in descending order. By either way, we can always re-rank retrieved passages by repetitively picking one passage from the top of each group until none is left.

4. EXPERIMENTAL RESULTS

We test our method on a set of runs obtained by the improved Okapi retrieval system [11] for TREC Genomics Track 2007 topics. The set of runs are acquired under different conditions as shown in Tables 1 to 4, where *k1* and *b* are tuning constants of the weighting function BM25. Indexing on database could be paragraph-based (where each piece of indexed information is a paragraph from documents) or word-based (where each piece of indexed information has a limited number of words), and topic expansion is applied once based on unified medical language system (UMLS). To enhance the performance of these runs, feedback analysis is performed by the Okapi retrieval system. In feedback analysis, the system retrieves ten passages that are deemed most relevant for a particular topic, and forms a list of the most recurring words from those passages. Each topic is expanded by these words, and then relevant passages for the extended topic is retrieved. Each feedback term is assigned a weight by Okapi. In our experiments, feedback weight is set to 0.25.

To get their vector representation, we apply both Porter stemming and a stoplist with general stopwords to passages. After Porter stemming and stoplist application, around 4000 words are left for each topic. All passages nominated for each topic are then represented with these words weighted by tf-idf (Better performance is observed with tf-idf instead of frequency used in the standard PLSA model as described in Section 3). We try to use principal component analysis (PCA) to reduce vector dimensionality. It seems PCA is not very helpful in reducing vector dimensionality without hurting performance in this problem. It might be because of the sparsity of data, no obvious dimensions are much more important than others, and every word has some contribution in representing passages nominated for a topic.

Topic models like PLSA typically operate in extremely high dimensional spaces. As a consequence, the "curse of dimensionality" is lurking around the corner, and thus the hyperparameters (such as initial conditional probabilities and smoothing parameters) settings have the potential to significantly affect the results [1]. In the experiments, we find that we cannot start PLSA model with a

Table 1: Run1: k1=1.4, b=0.55, word-based indexing, no topic expansion, aspect-level MAP 0.1017.

# of aspects (K)	1	2	3	4	5	6	7	8	9	10
Rerank	0.1017	0.1124	0.1157	0.1430	0.1263	0.1295	0.1243	0.1355	0.1373	0.1279
Improvement	0.00%	10.48%	13.78%	**40.67%**	24.18%	27.38%	22.24%	33.25%	35.02%	25.75%

Table 2: Run2: k1=1.4, b=0.55, word-based indexing, with topic expansion, aspect-level MAP 0.0611.

# of aspects (K)	1	2	3	4	5	6	7	8	9	10
Rerank	0.0611	0.0639	0.0721	0.0779	0.0886	0.0627	0.0726	0.0739	0.0815	0.0852
Improvement	0.00%	4.50%	17.91%	27.45%	**44.90%**	2.62%	18.72%	20.92%	33.41%	39.32%

Table 3: Run3: k1=2.0, b=0.4, paragraph-based indexing, no topic expansion, aspect-level MAP 0.0596.

# of aspects (K)	1	2	3	4	5	6	7	8	9	10
Rerank	0.0596	0.0672	0.0650	0.0875	0.0774	0.0832	0.0726	0.0616	0.0660	0.0723
Improvement	0.00%	12.74%	9.10%	**46.97%**	30.05%	39.76%	21.83%	3.43%	10.88%	21.46%

Table 4: Run4: k1=2.0, b=0.4, word-based indexing, no topic expansion, aspect-level MAP 0.08237957.

# of aspects (K)	1	2	3	4	5	6	7	8	9	10
Rerank	0.0824	0.0886	0.0942	0.0919	0.0846	0.0888	0.0836	0.0930	0.0953	0.0902
Improvement	0.00%	7.54%	14.34%	11.56%	2.57%	7.83%	1.47%	12.85%	**15.68%**	9.45%

uniform distribution for $P(z)$, $P(d|z)$, and $P(w|z)$; otherwise, the convergence will happen immediately in the first iteration due to the sparsity of data. Instead, we start with a normalized random distribution for all these conditional probabilities (the results reported in this paper are the average of a few runs). Due to the large dimensionality, there are a lot of zero probabilities in each passage vector representation. Zero estimates could cause significant problems such as zeroing-out the impact of some other useful parameters in multiplication. Zero estimates could also cause computation problems such as "division by zero". In our experiments, we apply Laplace smoothing to avoid zero probability estimates. We add a small value 2^{-52} to all probabilities before normalization. In the future, more smoothing techniques would be studied.

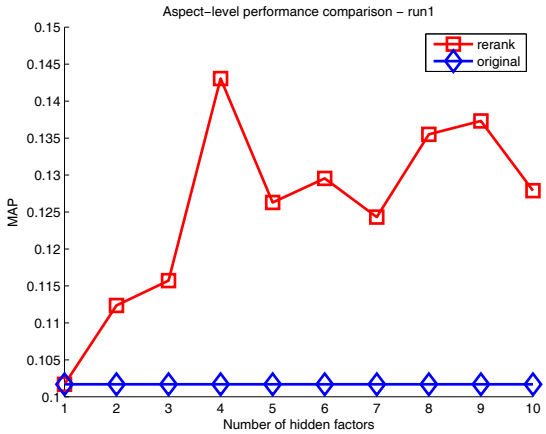

Figure 1: Performance improvement for run1.

In the experiment, we examine two ways of clustering passages in latent semantic space: one is centroid-based clustering with different distance functions (squared Euclidean, cosine, and cityblock) and the other is based on their most probable aspect factor. It is found that our problem is not so sensitive to either way of clustering, and for the former, not so much sensitive to the change of distance functions. We believe that this is also caused by the sparsity of data. Our experiment results reported here are from centroid-based clustering with cityblock distance function. In the future, we would explore other clustering algorithms that might be more suitable to our problem such as hierarchical clustering and density-based clustering.

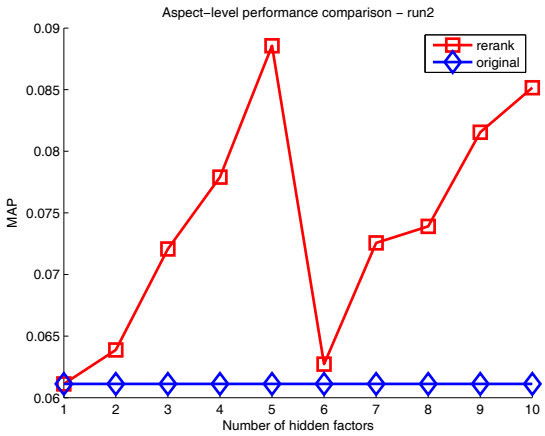

Figure 2: Performance improvement for run2.

In the experiments, we change the number of hidden aspects K from 1 to 10 continuously for all runs. When the number of hidden aspects is set to 1, there is no re-ranking and hence the performances are the same as the original runs. It turns out for all other 9 different hidden aspect numbers, all runs get positive performance improvements by re-ranking as shown in Tables 1 to 4. To illustrate the re-ranking performance graphically, we plot the data in Figures 1 to 4, respectively, where y-axis stands for the aspect-level performance MAP. It can be observed that on all 9 different number of hidden factors, the re-ranked results are all better than the original ones. Over all runs, the maximum improvement is 46.97% when $K = 5$ for run2, the minimum improvement is 1.47% when $K = 7$ for run4, and the average improvement is 20.06%. This is illustrated in Figure 5.

It should be noted that the hidden aspect factors in PLSA mod-

els are not necessarily the same as the aspects of Genomics Track. In PLSA models, the number of hidden aspect factors is a tuning variable, while the aspects of Genomics Track topics are constants once the corpus and topics are determined. The hidden aspect factors in PLSA models are statistically identified from data while the aspects of Genomics Track topics are assigned by the judges but not results of statistical analyses. Since PLSA models are good in semantic analysis and synonym and concept recognition [9], we use the hidden aspect factors identified by PLSA models to classify passages and then use this classification information to re-rank ranked lists in the hope that the hidden aspect factors do have some correlation with topic aspects in some way. Our experiment results highly support the hope.

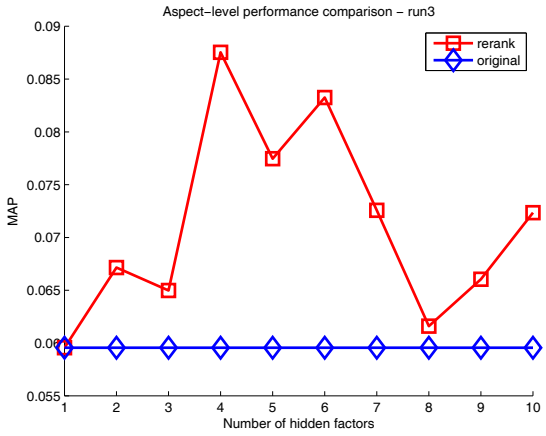

Figure 3: Performance improvement for run3.

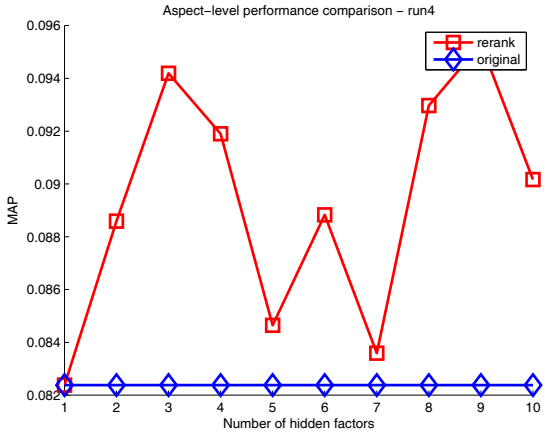

Figure 4: Performance improvement for run4.

5. CONCLUSIONS AND FUTURE WORK

In this paper, we conducted a preliminary study on using PLSA models to capture hidden aspects of retrieved passages. The hidden aspects caught are used to improve the performance of a ranked list by re-ranking. It turned out all runs on all 9 continuous hidden aspect numbers got positive improvements. This indicates PLSA models are very promising in finding diverse aspects in retrieved passages. By contrast, it was indicated [7] a clustering-based method always failed to improve the aspect performance over baseline algorithms.

In the future, more experiments will be conducted to further investigate the proposed method. We will extend the method to more runs, and will study whether there exist a range of hidden aspect numbers that can always be safely used in re-ranking to improve performance. In addition, we will investigate how to set different hidden aspect numbers for different topics. We will also examine other topic models such as LDA and LSA on this matter.

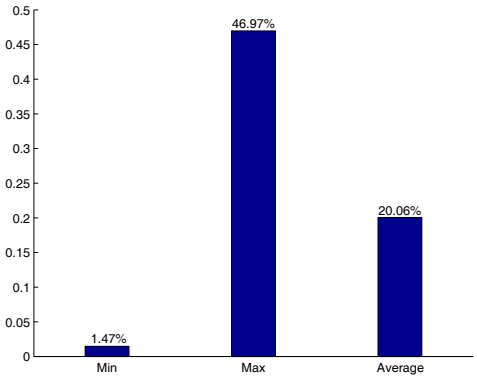

Figure 5: Performance improvement summary.

6. ACKNOWLEDGMENT

This research is supported by the research grant from the Natural Sciences & Engineering Research Council (NSERC) of Canada. We thank anonymous reviewers for their thorough review comments on this paper.

7. REFERENCES

[1] A. Asuncion and et al. On smoothing and inference for topic models. In *UAI'09*, pages 27–34, 2009.
[2] D. M. Blei and et al. Latent dirichlet allocation. *JMLR*, 3(4-5):993–1022, 2003.
[3] J. Carbonell and J. Goldstein. The use of MMR, diversity-based reranking for reordering documents and producing summaries. In *SIGIR'98*, pages 335–336.
[4] C. Clarke and et al. Novelty and diversity in information retrieval evaluation. In *SIGIR'08*, pages 659–666.
[5] S. Deerwester and et al. Indexing by latent semantic analysis. *JASIST*, 41, 1990.
[6] D. Demner-Fushman and et al. Combining resources to find answers to biomedical questions. In *TREC-2007*, pages 205–214.
[7] A. B. Goldberg and et al. Ranking biomedical passages for relevance and diversity: University of Wisconsin, Madison at TREC genomics 2006. In *TREC-2006*, pages 129–136.
[8] W. Hersh, A. Cohen, and P. Roberts. TREC 2007 genomics track overview. In *TREC-2007*, pages 98–115.
[9] T. Hofmann. Probabilistic latent semantic analysis. In *UAI'99*, pages 289–296.
[10] Q. Hu and X. Huang. A reranking model for genomics aspect search. In *SIGIR'08*, pages 783–784.
[11] X. Huang and et al. A platform for okapi-based contextual information retrieval. In *SIGIR'06*, pages 728–728, 2006.
[12] X. Huang and Q. Hu. A bayesian learning approach to promoting diversity in ranking for biomedical informaiton retrieval. In *SIGIR'09*, pages 307–314.
[13] X. Yin and et al. Survival modeling approach to biomedical search result diversification using wikipedia. *TKDE*, 25(6):1201–1212, 2013.
[14] C. Zhai, W. W. Cohen, and J. Lafferty. Beyond independent relevance: methods and evaluation metrics for subtopic retrieval. In *SIGIR'03*, pages 10–17.
[15] B. Zhang and et al. Improving web search results using affinity graph. In *SIGIR'05*, pages 504–511.
[16] Y. Zhang, J. Callan, and T. Minka. Novelty and redundancy detection in adaptive filtering. In *SIGIR'02*, pages 81–88.
[17] W. Zhou and C. Yu. TREC genomics track at UIC. In *TREC-2007*, pages 221–226.
[18] X. Zhu and et al. Improving diversity in ranking using absorbing random walks. In *NAACL-HLT 2007*, pages 97–104.