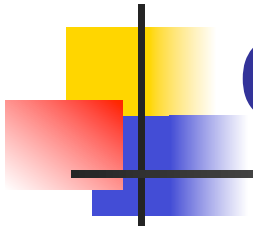
Promoting Ranking Diversity for Biomedical Information Retrieval based on LDA

Yan Chen, Xiaoshi Yin, Zhoujun Li, Xiaohua Hu and Jimmy Huang

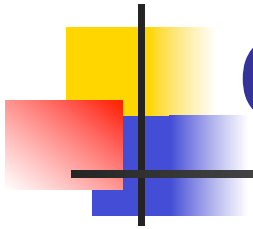
State Key Laboratory of Software Development Environment, Beihang University, China
School of Computer Science and Engineering, Beihang University, China
College of Information Science and Technology, Drexel University, Philadephia, PA, USA
School of Information Technology, York University, Canada

IEEE BIBM 2011
Atlanta, Georgia, USA, 15th Nov. 2011



Outline

- Background and Motivation
- Related Work and Contributions
- Reranking Strategies Based on LDA
  - Aspect Discovery and Transformation
  - Reranking with N-size Slide Window
- Experiments
  - Test Collections, Evaluation Measures and Baseline Runs
  - Experimental Results and Analyses
- Conclusion and Future Work



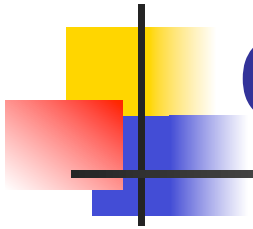
Outline

- Background and Motivation
- Related Work and Contributions
- Reranking Strategies Based on LDA
  - Aspect Discovery and Transformation
  - Reranking with N-size Slide Window
- Experiments
  - Test Collections, Evaluation Measures and Baseline Runs
  - Experimental Results and Analyses
- Conclusion and Future Work

Background and Motivation

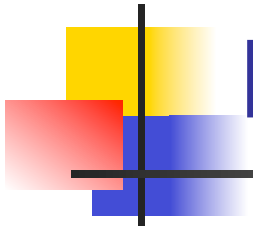
- Background
  - Traditional IR models assume that the relevance of a document is independent of the relevance of other documents.

    ⇒ High redundancy and low diversity.
  - Aspect search in biomedical IR
    - 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, such as genes, proteins, diseases, mutations, etc.
    - TREC 2007 Genomics tracks' "aspect retrieval" : to study how a biomedical retrieval system can support a user gather information about the different aspects of a topic.
    - Diversity evaluation: Aspect Mean Average Precision (Aspect MAP).
- Motivation: promoting ranking diversity for biomedical IR



Outline

- Background and Motivation
- Related Work and Contributions
- Reranking Strategies Based on LDA
  - Aspect Discovery and Transformation
  - Reranking with N-size Slide Window
- Experiments
  - Test Collections, Evaluation Measures and Baseline Runs
  - Experimental Results and Analyses
- Conclusion and Future Work



Related Work

- Carbonell *et al.* introduced the maximal marginal relevance (MMR) method, which attempts to maximize relevance while minimizing similarity to higher ranked documents.
- Zhang *et al.* presented four redundancy measures. They modeled relevance and redundancy separately. Since they focused on redundant document filtering, experiments in their study were only conducted on a set of relevant documents.
- Zhai *et al.* validated a subtopic retrieval method based on a risk minimization framework. Their subtopic retrieval method combines the mixture model novelty measure with the query likelihood relevance ranking.

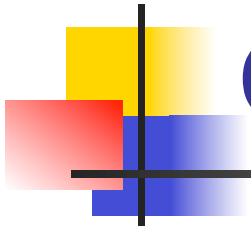
Related Work

- Rianne Kaptein *et al.* employed a top down sliding window to diversify ranked list of retrieved documents and diversity according to some diversity indicators.
- Genomics aspect retrieval conducted by Huang *et al.* demonstrated that the hidden property based re-ranking method can achieve promising and stable performance improvements.
- Yin *et al.* proposed a cost-based re-ranking method to promote ranking diversity. This method concerns with finding the passages that cover more different aspects of a query topic.
- University of Wisconsin re-ranked the passages using a clustering-based approach named GRASSHOPPER to promote ranking diversity.

Related Work

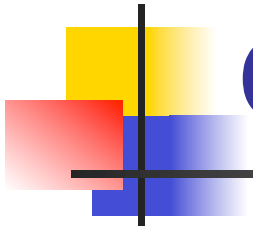
- Consider the aspects of user query and retrieved documents mainly on word level.
- For example, given two retrieved passages:
  - the first one is related to some disease research, in which **kidneys** of white rats are used as experimental materials;
  - the second one is relevant to subject of **kidney** transplantation.
- Two Reasons:
  - Firstly, one or more co-occurrence words in a passage are used to identify the aspect.
  - Secondly, words in a passage are considered as independent to each other.

→ It is insufficient to identify aspect on word level.



Contribution

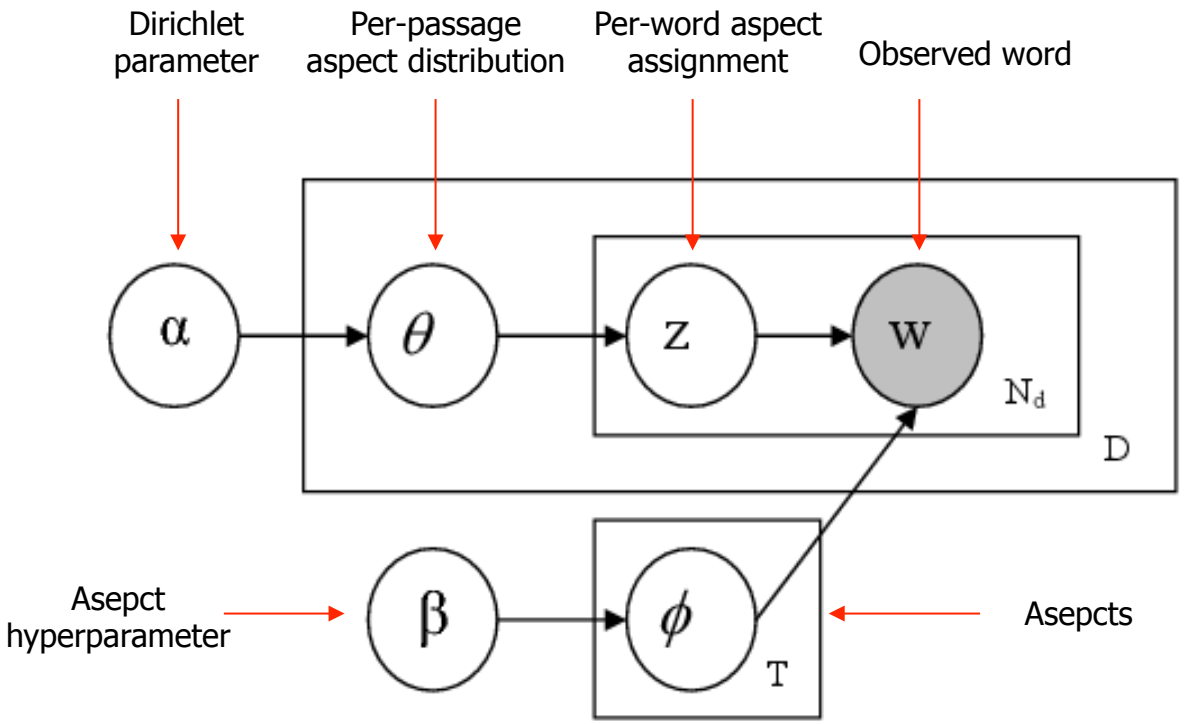
- Our contribution is three-fold.
  - First, to the best of our knowledge, this is the first study of adopting topic model to biomedical IR.
  - Second, some transformations with topic distribution for retrieved passages are made.
  - Third, two re-ranking algorithms based on "N-size slide window" are proposed, which take both passage novelty and relevance into account.



Outline

- Background and Motivation
- Related Work and Contributions
- Reranking Strategies Based on LDA
  - Aspect Discovery and Transformation
  - Reranking with N-size Slide Window
- Experiments
  - Test Collections, Evaluation Measures and Baseline Runs
  - Experimental Results and Analyses
- Conclusion and Future Work

Aspect Discovery



LDA Model

Aspect Distribution Transformation

Aspect distribution matrix

$$\theta = (\theta_1, \theta_2, \cdots, \theta_D)^T$$

$$= \begin{pmatrix} a_{11} & \cdots & a_{1j} & \cdots & a_{1T} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ a_{i1} & \cdots & a_{ij} & \cdots & a_{iT} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ a_{D1} & \cdots & a_{Dj} & \cdots & a_{DT} \end{pmatrix}$$

A new matrix $\Theta = (\Theta_1, \Theta_2, \cdots, \Theta_T)^T$

Measuring the passage importance for each aspect

$$\begin{pmatrix} N_1(a_{11}) & \cdots & N_1(a_{j1}) & \cdots & N_1(a_{D1}) \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ N_i(a_{1i}) & \cdots & N_i(a_{ji}) & \cdots & N_i(a_{Di}) \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ N_T(a_{1T}) & \cdots & N_T(a_{jT}) & \cdots & N_T(a_{DT}) \end{pmatrix}$$

$(1 \le i \le T, 1 \le j \le D)$

$$\mu_i = \frac{\sum_{j=1} \cdots}{D}$$

Re-ranking with N-size Slide Window

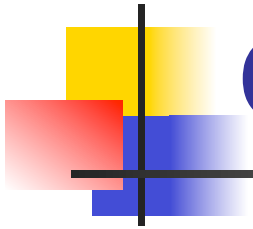
Algorithm 2 rank_NWin_Group Algorithm

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

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

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

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

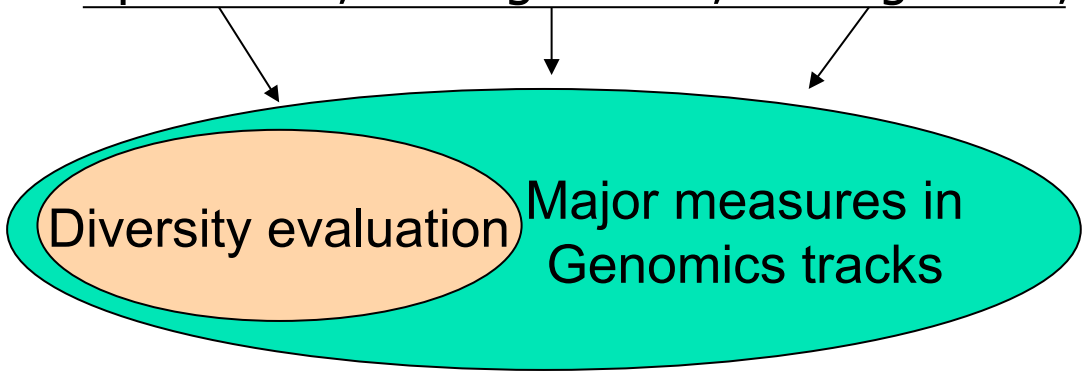


Outline

- Background and Motivation
- Related Work and Contributions
- Reranking Strategies Based on LDA
  - Aspect Discovery and Transformation
  - Reranking with N-size Slide Window
- Experiments
  - Test Collections, Evaluation Measures and Baseline Runs
  - Experimental Results and Analyses
- Conclusion and Future Work

Test Collection and Evaluation Measures

- TREC 2007 Genomics Track Collections
  - Full-text biomedical literature corpus.
  - 36 topics from the 2007 Genomics track;
  - Topics are in the form of questions asking for lists of specific entities that cover different portions of full answers to the topics.
- Evaluations Measures
  - Aspect MAP; Passage2 MAP; Passage MAP; Document MAP



IR Baseline Runs

- NLMinter
  - It achieved the highest Aspect MAP, Passage2 MAP and Document MAP in 2007 Genomics track.
- UniNE2
  - Its performance was above average among all results reported in 2007 Genomics track.

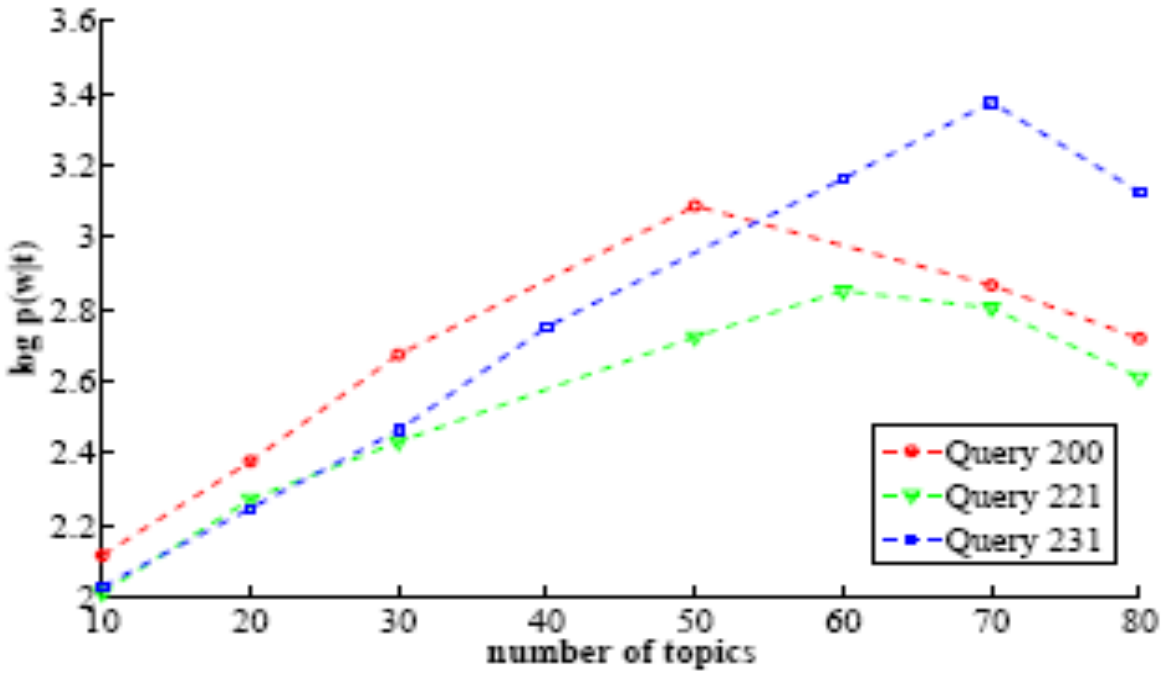
Experimental Results

NLMinter model				
MAP	Aspect	Passage2	Passage	Document
NLMinter	0.23068962	0.07335484	0.05971977	0.20962491
$rank\_NWin$	0.2438342 (+5.70%)	0.07368625 (+0.45%)	0.05868155 (-1.74%)	0.20790886 (-0.82%)
$rank\_NWin^*$	0.24426998 (+5.89%)	0.07372402 (+0.50%)	0.05849706 (-2.05%)	0.20744464 (-1.04%)
$rank\_NWin\_Group$	0.24908569 (+7.97%)	0.07792334 (+6.23%)	0.06151813 (+3.01%)	0.20976964 (+0.07%)
$rank\_NWin\_Group^*$	**0.24910669 (+7.98%)**	0.07793161 (+6.24%)	0.06152586 (+3.02%)	0.20977025 (+0.07%)

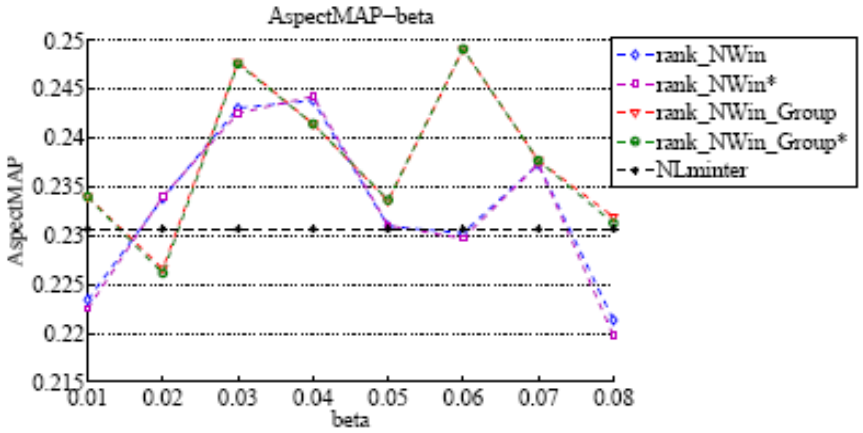
UniNE2 model				
MAP	Aspect	Passage2	Passage	Document
UniNE2	0.09880169	0.01777397	0.05236709	0.13771527
$rank\_NWin$	0.1052544 (+6.53%)	0.01946295 (+9.50%)	0.05459447 (+4.25%)	0.13969831 (+1.44%)
$rank\_NWin^*$	0.1052544 (+6.53%)	0.01946007 (+9.49%)	0.05459788 (+4.26%)	0.13964510 (+1.40%)
$rank\_NWin\_Group$	**0.10554020 (+6.82%)**	0.01902429 (+7.03%)	0.05490502 (+4.85%)	0.14035642 (+1.92%)
$rank\_NWin\_Group^*$	0.10549095 (+6.77%)	0.01902427 (+7.03%)	0.05490508 (+4.85%)	0.14035350 (+1.92%)

Results Analysis

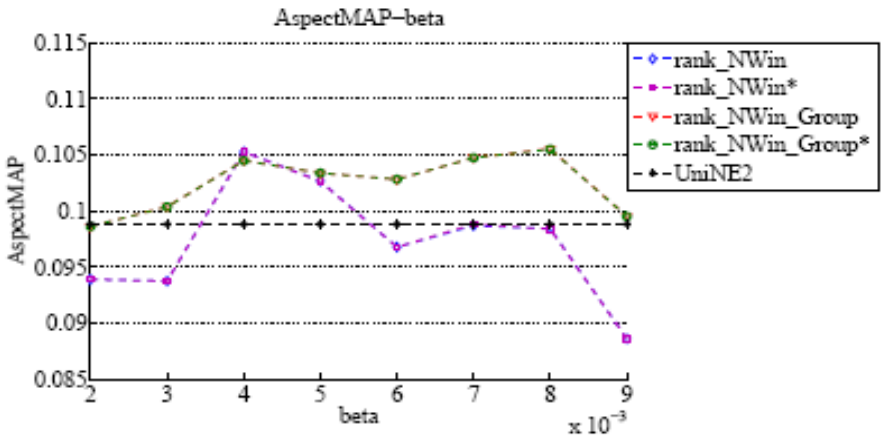
- Impact of Parameter β
- Impact of Parameter α and T



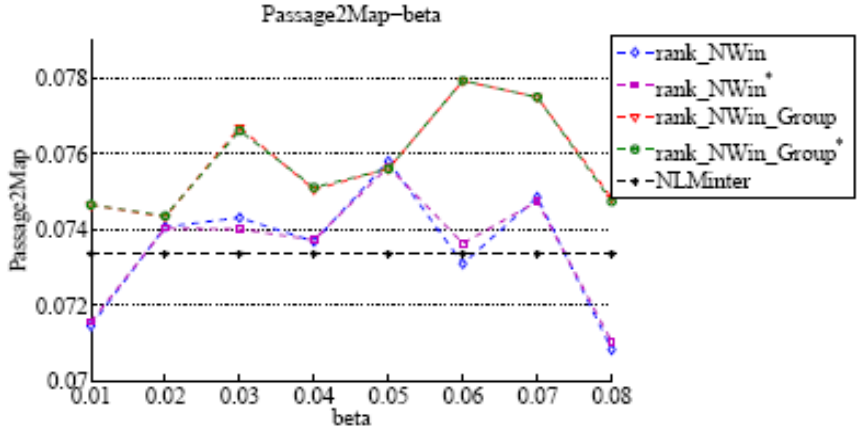
Results Analysises



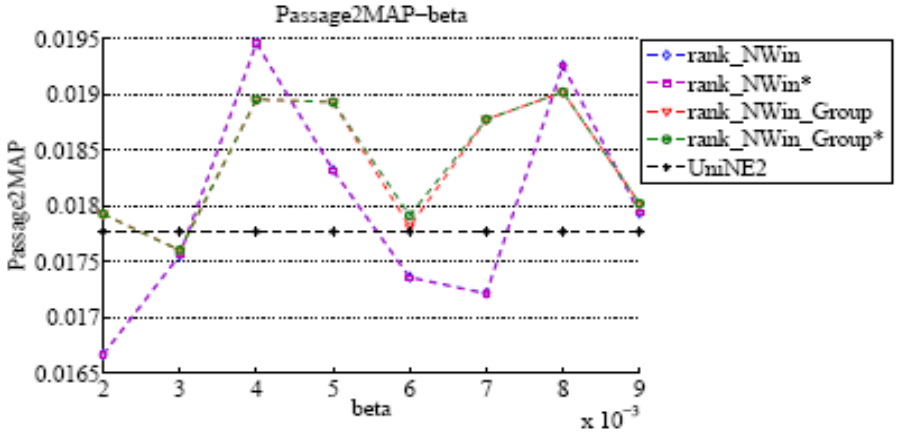
(a) Aspect MAP for NLMinter Run



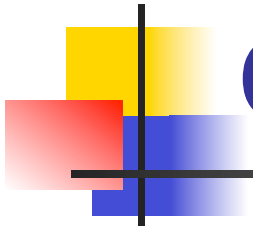
(b) Aspect MAP for UniNE2 Run



(c) Passage2 MAP for NLMinter Run



(d) Passage2 MAP for UniNE2 Run



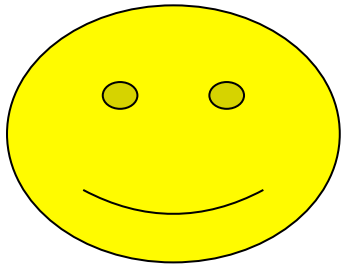
Outline

- Background and Motivation
- Related Work and Contributions
- Reranking Strategies Based on LDA
  - Aspect Discovery and Transformation
  - Reranking with N-size Slide Window
- Experiments
  - Test Collections, Evaluation Measures and Baseline Runs
  - Experimental Results and Analyses
- Conclusion and Future Work

Conclusion and Future Work

- We propose an approach which employs LDA to promoting ranking diversity for biomedical IR.
  - The first study of adopting topic model to biomedical IR.
  - Transformations with topic distribution for retrieved passages are made.
  - Two re-ranking algorithms based on "N-size slide window" are proposed.
- We intend to extend this work by exploring both more complex models and more sophisticated algorithms.
- We also plan to further improve our approach to solve the diversification in the other application fields, such as SNS, recommendation system, etc.

Thank you!



Questions?

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.

References

[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.