SNUMedinfo at TREC Web track 2014

Sungbin Choi, Jinwook Choi

Medical Informatics Laboratory, Seoul National University, Seoul, Republic of Korea

wakeup06@empas.com, jinchoi@snu.ac.kr

Abstract. This paper describes the participation of the SNUMedinfo team at the TREC Web track 2014. This is the first time we participate in the Web track. Rather than applying more sophisticated retrieval method such as learning to rank models, this year we used only baseline retrieval models with spam filtering and pagerank prior.

Keywords: Web search, Information retrieval, Sequential dependence model, Spam filtering

1. Introduction

In this paper, we describe the methods in participation of the SNUMedinfo team at the TREC Web track 2014. For a detailed task introduction, please see the overview paper of this track.

2. Methods

We used sequential dependence model (SDM) [1] as a baseline retrieval model. For the experiment, we used batch query service offered by lemur project website [2]. ClueWeb12-Full dataset is our test corpus. Waterloo spam filter [3] is used to filter out spam documents. Details of our submitted runs can be summarized as following table.

Table 1. Submitted runs

RunID	Method description
SNUMedinfo11	SDM
SNUMedinfo12	SDM + Spam filtering (threshold: 50)
SNUMedinfo13	SDM + Spam filtering (threshold: 50) + Pagerank Prior score

SDM : Sequential dependence model

Regarding SNUMedinfo13, we used Pagerank Prior [4] scores offered by lemur project website.

3. Results

Table 2. Evaluation results

RunID	ndcg@20	err@20
SNUMedinfo11	0.2436	0.1386
SNUMedinfo12	0.2698	0.1759
SNUMedinfo13	0.1927	0.1230

4. Conclusion

This year, we submitted baseline retrieval model with spam filtering and pagerank prior score. We plan to experiment with more advanced retrieval methods in the next year's participation.

5. Acknowledgements

This study was supported by a grant of the Korean Health Technology R&D Project, Ministry of Health & Welfare, Republic of Korea. (No. HI11C1947)

6. References

1. Metzler, D. and W.B. Croft, *A Markov random field model for term dependencies*, in *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*. 2005, ACM: Salvador, Brazil. p. 472-479.
2. *The Lemur Project*. [cited 2014 Oct 28]; Available from: http://www.lemurproject.org.
3. Cormack, G., M. Smucker, and C.A. Clarke, *Efficient and effective spam filtering and re-ranking for large web datasets.* Information Retrieval, 2011. **14**(5): p. 441-465.
4. Page, L., et al., *The PageRank Citation Ranking: Bringing Order to the Web*. 1999, Stanford InfoLab.