IR Project Proposal

Manuela Bergau
s4543645

Shima Yousefi Roudbordeh
s1039415

Evgeniia Martynova
s1038931

October 2019

1 Introduction

Modern search systems have reached the impressive results in terms of document relevance for a query, however, their architecture and the current state of web documents collection lead to usability problems. One of them is the redundancy of search results when the content of the top-ranked documents overlaps to a large extent. Thus the utility of all documents from search results except the first one is little for the users. This problem is relevant to most search queries but especially noticeable in the news domain.

Getting all similar content can sometimes not be a user intention. In order to resolve this we want to test a re-ranking method that combines the Affinity ranking [2] and the redundancy measure [3]. We want to compare the results based on diversity and information richness.

2 Resources

2.1 Data

We will use data from the TREC 2005 Robust data sets.

2.2 Software

We want to use Lucene as our main search engine and conduct our information retrieval experiments using Anserini. Anserini is an open-source information retrieval toolkit built on Lucene. Experiments verify that Anserini is both efficient and effective, providing a solid foundation to support our research [1].

3 Experimental design

We are going to develop the re-ranking method on the top of Anserini open source IR tool. The idea of our re-ranking algorithm is to use both information richness [2] and redundancy measures to re-rank the original results of the search system. We want to try a different approach to the calculation of information richness of a document. Instead of building a graph for the whole document collection we will try to calculate information richness only for the documents returned by the original search system. We might need to modify the algorithm for this purpose.

To analyze our implementation, we will compare original Anserini search results with the search results which re-ranked by our algorithm and analyse the trade-off between the improvements in information richness and diversity and decrease of precision and recall. In addition to that, we are using the baseline from [2] to compare our performance.

4 Expected outcome

We think that an IR tool with adopted re-ranking will demonstrate better performance in terms of diversity of search results and information richness than the original search engine.

References

[1] Peilin Yang, Hui Fang, and Jimmy Lin. Anserini: Enabling the use of lucene for information retrieval research. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '17, pages 1253–1256, New York, NY, USA, 2017. ACM.

[2] Benyu Zhang, Hua Li, Yi Liu, Lei Ji, Wensi Xi, Weiguo Fan, Zheng Chen, and Wei-Ying Ma. Improving web search results using affinity graph. In *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '05, pages 504–511, New York, NY, USA, 2005. ACM.

[3] Yi Zhang, Jamie Callan, Jamie Callan, and Thomas Minka. Novelty and redundancy detection in adaptive filtering. In *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '02, pages 81–88, New York, NY, USA, 2002. ACM.