Information Retreival - Mini-Project Ideas

Venkatesh Vinayakarao

November 2020

*Please scope your mini-project such that the work is approximately equivalent to the effort put in an assignment.

Following are some mini-project ideas for you to consider:

1. **Cross-Document Order Independent LTR**: Relevance annotations also show that judges look at several documents before deciding on the relevance. However, most Probability Ranking Principle (PRP) based Learning to Rank (LTR) models work on a document at a time. Capturing cross-document interactions in an LTR model is a challenge. Specifically, some of the proposed state of the art cross-document LTR models depend on the order in which these documents are presented. Can we take a set of documents as input and learn an order independent scoring function that maps it to its permutation?

2. **A Domain Specific Ranker**: Microsoft Learning to Rank[1] Dataset (MSLR-WEB10K) is a large scale dataset often cited in Information Retrieval research. Each query-document pair was labeled by human annotators with 5-level relevance judgments ranging from 0 (irrelevant) to 4 (perfectly relevant). For a specific information need, can you implement a simple ranker in Solr/Lucene and compute the NDCG of your ranker?

3. **Phrase-Level Query Segmentation**: Several query reformulation algorithms depend on query segmentation. We have seen a simple query segmentation idea that uses dictionary terms. Can we improve this idea to propose a query segmentation algorithm that can work on phrases as well?

4. **Role of Entities in Improving Search**: Queries and documents may contain entities. Entity detection can these days be easily performed using tools like the Stanford Named Entity Recognizer[2] (NER). Can you use the NER to prepare an auxiliary index with entities found in documents? Further the same entity extraction can be used on queries to improve search in Solr.

[1] http://research.microsoft.com/en-us/projects/mslr/

[2] https://nlp.stanford.edu/software/CRF-NER.html

5. **Custom Analyzer for Hindi**: Tokenizing and filtering can be quite challenging to be implemented for a non-English language. Can you implement one for Hindi?

6. **Inferring User Intent from Interactions with the Search Results** Eugene Agichtein mentions that “Extracting meaningful signals from this data would enable a search engine to accurately infer user intent for tasks such as real-time result reranking”. Can you write a small software component that can take few queries from the same user session and re-rank the results?