CS6370 Natural Language Processing
Project Proposal

Shubham Patel, Bagul Amit Sunil (ME19B170, MM19B023)

IIT Madras, Chennai, 600036, Tamil Nadu, India .

1 Proposal

1.1 Limitations of the Vector Space Model

1. The current search engines suffers from Dimensionality curse. As the number of documents and queries increase, it takes more storage and time to compute. It becomes computationally expensive to process data with current model. Also the longer documents will get more relevance score due to more words and likewise, the smaller ones will get less score.
2. Vector space model is such that all important quantities pertaining to documents need to be recalculated when a new document or term is added.
3. As current model only uses the presence of words without considering adjacent words or sentences (context indifferent) for similarity calculation, Order in which term and sentences occurs is lost.
4. Polysemy. The current model does not take into consideration meaning of particular word in the given context. Query with polysemous words may get high similarity with irrelevant documents (having same words but different meaning) and thus may suffer from precision.
5. Synonymy. The current model does not identify different words that have same meaning as related words in similarity calculation. Thus it may give less similarity even for relevant documents if words are synonyms.

1.2 Proposed Hypotheses

- **LSA:** Latent Semantic Analysis can be used to solve above mentioned limitations namely, synonymy and to some extent polysemy. Instead of thinking documents as "bag of words" with each word occurring separately, topic model (word or text clustering) can be used to improve document representation The reason that it can overcome above mentioned limitations is because it is a process that finds unstructured data to get hidden relationships between terms and concepts using singular value decomposition. The term matrix is used for dimensionality reduction (Latent semantic space). It can help reduce noise and data is placed according to correlation which aids retrieval.

- **BM25:** Using different ranking than TF-IDF cosine similarity. BM25 improves TF-IDF by using two parameters to do the following: *Term saturation Parameter k* is used to detect saturation of TF after which, no additional relevance is provided. For example if two document contains many occurrences of 'dog', one with frequency 200, other with 100. Both will have different relevance to query 'dog' with TF-IDF but relevance likelihood can be controlled in BM25 using k. *Document Length normalization Parameter b* is used to penalised the documents with lengths larger than average document length and smaller ones are rewarded. This is to normalize the size of documents and have fair judgement unlike TF-IDF.
- **Query expansion:** Using only keywords is not very efficient way. Generally, query is small and results in poor document retrieval due to lack of sufficient information. Query expansion is process to improve the quality of user query. If Query expansion is used along with the above mentioned models, improvement in retrieval performance is possible. By Adding additional relevant terms to the terms of query vocabulary mismatch can be addressed increasing efficiency of retrieval system. [1] Hybrid approach that combines WordNet and DBpedia gives better result than using only WordNet.
- A combination of above mentioned models, with weight w_m for model m such that $\sum_i w_m = 1$. This can potentially help in improving recall, precision and overall performance of our search engine.

1.3 Flow of modelling

- Preprocessing the documents from carnfield dataset and query.
- Expand query using WordNet and DBpedia
- Fit he LSA model using SVD
- LSA model Vector representation for query.
- Building BM25
- Use of combination of models
- Rank the documents
- Retrive top k 1,10 documents
- Evaluate and compare the result with previous search engine

1.4 Evaluation Measures

- Hypothesis testing methods such as, Precision@k, Recall@k, F-score, MAP@k, nDCG@k for different values of $k \in (1, 10)$.
- χ^2-testing with a reasonable null hypothesis

References

[1] Dahir Sarah, Khalifi Hamid, El Qadi, Abderrahim: Query Expansion using DBpedia and WordNet