Indian Institute of Technology, Madras

CS6370: Natural Language Processing - Report
Project Proposal

Nithin Uppalapati

EE18B035
ee18b035@smail.iitm.ac.in

D Krithin Chowdary
EE18B047
ee18b047@smail.iitm.ac.in

19-04-2021

Contents

1. Aim

The goal of this project is to improve the search engine by addressing its current limitations, and compare it's performance to the search engine which is updated. Both the search engines will be operated on the Cranfield Data-Set and the results are compared against each other.

2. Proposal

2.1 Limitations of the current Vector Space Model (VSM) we are trying to address:

Below is the list of limitations of the current model:

- The VSM model of search engine doesn't check for the spelling errors in the query, as a result when a spelling error is encountered, we don't get relevant results. Hence any document in the same context as that of the query, but using different vocabulary is lost.
- The VSM model doesn't retrieve the results / docs which are also related to the query (synonyms), which does not contain the query terms in the documents. So, even if the synonym of the query is present in documents, they are treated with low importance (ranks).
- The VSM model does not take into account of the order of terms in the query. It just considers the query as a bag of words.
- The VSM model doesn't ask the user for any kind of satisfaction with the results it retrieved.

2.2 Hypotheses for addressing some of the above limitations

- In the VSM model, the query is taken as it is, other than stopword removal and tokenization, no other processing is done. Any documents containing words semantically/lexically related to the query terms weren't retrieved. To address this we perform query expansion, which maps the terms of the input query to the relevant terms in form of similarity arcs. In this way we can also retrieve the documents which are relevant to the query, but do not contain the exact query terms.
- In VSM model, the results were vulnerable to the spell errors created by the users. So, by implementation of spell check and spell correction, we expect the results to be better (immune) even when there are errors in spelling of the query terms.
- In the VSM model, as there was no notion of relatedness between the documents, we were not able to retrieve the other documents which are indirectly related to the query. Where

as in LSA, we expect that the retrieved results are improved significantly better as both documents and queries are mapped to the lower dimensional latent concept space, which also retrieves the documents which are latently entangled with each other.

2.3 Methods for realizing the above hypotheses in current search engine

- Implementing LSA (Latent Semantic Analysis) to also retrieve the documents which are contextually related to the query.
- Implementing Spell check and spelling correction to correct the mistypes in queries.
- Implementation of WSD (Word Sense Disambiguation) using SEMCOR as background resource.
- Also implementing ESA (Explicit Semantic Analysis) to retrieve the documents which are of relevant to the query, based on the document titles / Wikipedia articles' titles.

2.4 Evaluation of updated search engine.

- Evaluation of the metrics which are already performed on the VSM model.
- Comparison of P vs R graphs for the old and new models.
- Using Kendall's tau measure to compare the similarity between old rankings and new rankings. And hence learning the degree of variation in the new model when compared to the old model.
- Also calculate Spearman's rho to compare the similarity between old rankings and new rankings.