High Level Project Description

In this project, I have developed my own Search Engine. Search engines are used in many applications like Facebook, LinkedIn, Google and many more. A search engine enables a user to submit a query and get a list of documents as result for the query just like popular search engine "Google" does.

In this project I have applied my "Information Retrieval" concepts to develop a search engine. My search engines accepts queries as input, applies Information Retrieval concepts on these queries and returns a list of documents as a result to the user. I have used CACM corpus and queries for the development of this project and for evaluation purposes.

The resultant documents are arranged in the order of relevance, that is, top most results are most relevant and the relevance decreases as the user goes down the documents list. Google also follows the concept of arranging the results based on relevance. That is why the top results are most relevant.

In order to get the best results, I have applied a strategy to expand the queries by adding more words to it using thesauri and ontology and applied a retrieval model to get the top results from any given query.

Query expansion technique is helpful when the input queries are understated, that is, if the queries are very brief. For example, query "Italian food" is very brief thus the query expansion technique uses modern thesauri to add more relevant terms to the query like "dishes", "recipe", "meals", "cuisine" etc. Another example, query "wooden furniture" is also very brief thus the technique will first add terms like "chairs", "couch", "home", "design", "tables" etc. to the query in order to get more relevant documents.

A retrieval model is used to calculate a score for every document. A higher score of the document suggests that the document is relevant to the query.

In addition to returning results, I also return snippet for each document in the result and also highlight important terms in the snippets so that the user can pick up the best document by just looking at the snippet of the document. A snippet is representative of a document that shows how relevant a document may be.

For evaluation purposes, I calculated the precision that is the measure of relevance of the documents with the respective queries, recall that is the measure of how many relevant documents were returned by search engine out of a bunch of relevant documents. Evaluation of the search engine suggested that the query expansion is the best technique to get a large number of relevant documents.