Query Expansion

Team: 6
Team Members
201001095 - Akshat Khandelwal
201101011 - Mahaver Chopra
201305514 - Uma K.L
201307674 - Veggalam Spandan

Problem Statement

Query expansion aims to extend the user query with related terms so as to improve the effectiveness of search. We need to overcome the problems that existing query expansion techniques has like query drift, the performance robustness problem through clustering.

Motivation

Query expansion is a process of reformulating the query in order to increase the relevance of documents retrieved by IR system. Information provided by the user may not be sufficient to retrieve or it may miss some relevant documents, to avoid this query is augmented to improve the search results. This is because often a user may not be able to represent his information need using the correct query which retrieves all the relevant documents.

There could be problems with vocabulary, spellings, words mismatch, which can be handled during query expansion. Hence query expansion is crucial to improve the precision and recall of any IR system.

Approach

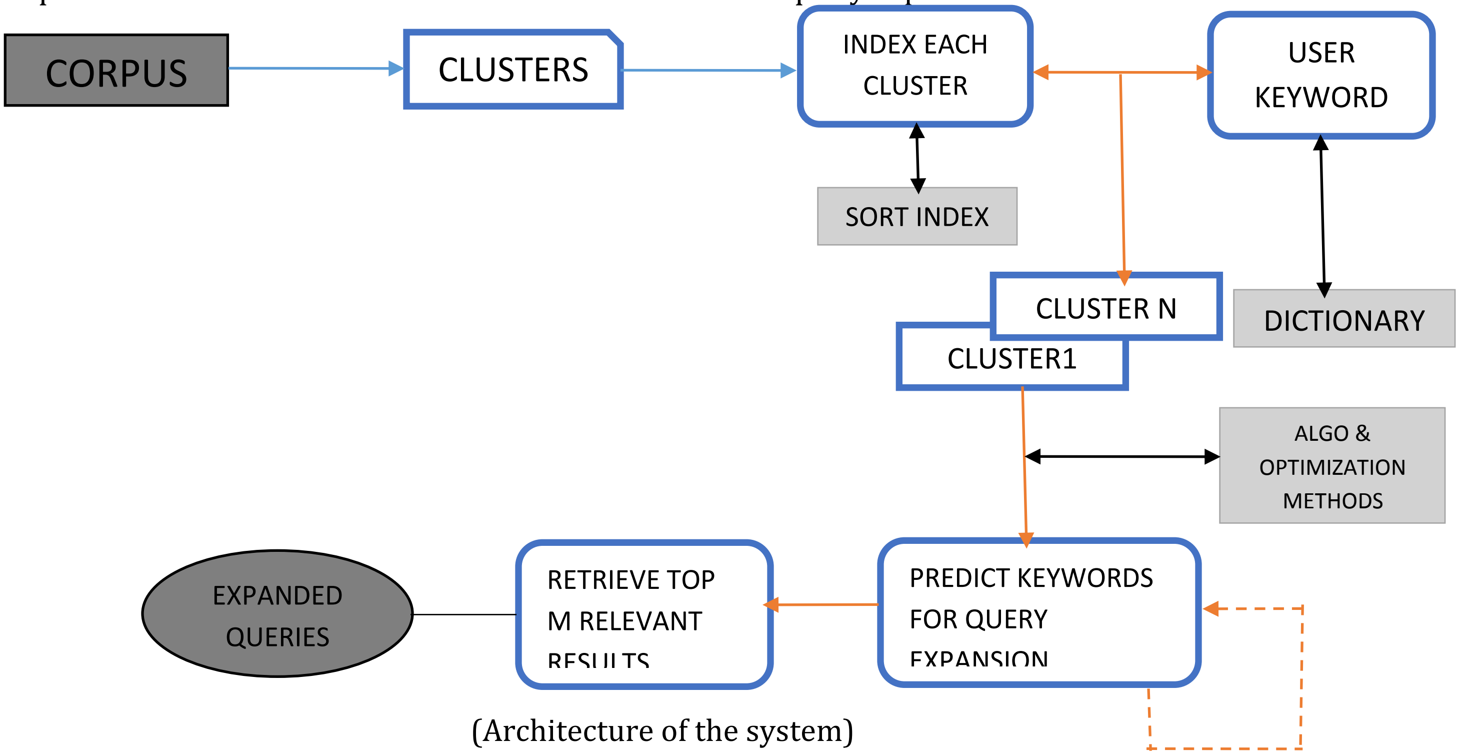
In this project we aim to implement a system that does Cluster based query expansion. We also show that the results that this system gives have better recall and precision when compared to the existing systems that use only Pseudo-feedback (PF)-based query expansion methods.

The existing system uses query expansion based on Pseudo feedback which expands the query with terms from documents that are present in an initially retrieved list. The list will contain documents which are ranked high based on document-query similarities.

The problem with this system is that some or even many of the documents in the initial list may not be relevant. The initial list based on which the query expansion is done may not reflect all the aspects on the information need. Hence the expanded query may exhibit query drift. That is, it may represent an information need different from that of the underlying original query. The expanded queries may even correspond to a subset of possible query semantics and miss many relevant results. Indeed, there are many queries for which state-of-the-art PF expansion methods yield retrieval performance that is substantially inferior to that of using the original query with no expansion, the performance robustness problem.

The potentially degraded quality of the initial list is often caused by the virtue of the way it is created, that is, using surface-level document-query similarities. Thus we perform query-expansion based on clusters of similar documents that are created before accepting user input.

Clusters are a collection of documents which are created based on the similarities that exist between the documents and they reflect the corpus-context better than individual documents. The document clusters are created before the execution of the query. Clusters can even contain relevant documents that do not exhibit high surface-level query similarity. Having such clusters that are created offline can improve the overall effectiveness and performance robustness of a state-of-the-art PF-based query expansion method.



(Architecture of the system)

We use Cluto clustering tool for creating clusters, and Apache Lucene for tokenizing and indexing.

Clusters and un-clustered documents containing input key words are retrieved, attributes required for query expansion are predicted using pseudo relevance feedback approach and we also use external sources to find words related to user input key words.

Recall and precision values are calculated for top ranked expanded queries and are compared with existing approach.

Dataset

- Dataset contains different stories from different pages published in "The Telegraph – Calcutta" newspaper.
- Newspaper include following pages: front page, sports, business, opinion, bengal, calcutta, foreign, nation, etc.
- Each story is saved to a document in following structure.

  ```
  <DOC>
          <DOCNO>
          <TEXT>
  </DOC>
  ```
- Dataset contains 8000+ story documents. DocNo gives the information about the document and is of following format.

  <ID1>_<PageName>_story_<StoryNum>.utf8

  ID1 gives information about the date, PageName gives name of page in which this story is published. Each story is assigned with a unique StoryNum.

Final Deliverable

Deliverable 1: Clustering and Indexing clusters. Define cluster attributes to tokens

Deliverable 2: Implementing query expansion techniques

Deliverable 3: Retrieve top query expansion results and compare the results from other techniques.

Deliverable 4: Optimize both clustering and query expansion modules. Document the approach and present final evaluation report.

References

1. Improving search engine Query Expansion techniques with ILP
   Jos_e Carlos Almeida Santos and Manuel Fonseca de Sam Bento Ribeiro

2. Query Expansion Based on Clustered Results
   Ziyang Liu Sivaramakrishnan Natarajan Yi Chen
   Arizona State University

3. Cluster-Based Query Expansion
   Inna Gelfer Kalmanovich and Oren Kurland
   Faculty of Industrial Engineering and Management Technion
   Israel Institute of Technology

4. An Empirical Study of Query Expansion and Cluster-Based Retrieval in Language Modeling Approach
   Seung-Hoon Na, In-Su Kang, Ji-Eun Roh, and Jong-Hyeok Lee
   Division of Electrical and Computer Engineering,
   POSTECH, AITrc, Republic of Korea