Enhancing Document Retrieval using Enriched Representation of Query Terms

A thesis submitted by

Rajendra Prasath. R

[09CS9404] - Institute Research Scholar

Supervisor: **Prof. Sudeshna Sarkar**

Dept. of Computer Science and Engineering

Searching for information has become an important part of us in our day to day life. Modern Information Retrieval (IR) systems attempt to satisfy the user's information needs given in the form of a query and output a ranked list of documents. A document presented by an IR system is said to match a user's information needs only if the user perceives it to be relevant and useful. User queries are often short and ambiguous. An important challenge for IR systems is how to find documents relevant to such queries. The occurrence of query terms alone is not often a sufficient indicator of the relevance of a document to the query. In many situations, a query may be reformulated, enhanced with added terms so as to be able to retrieve more relevant documents. It may also be possible to have a representation of the terms that capture the meaning of the terms and then relation of association with other terms.

In this thesis, we examine some approaches to query understanding, query enhancement and query representation to improve the efficiency of document retrieval. Given a specific collection of documents and an user information need, our task is to retrieve documents that are relevant to the user information needs in terms of various aspects like comprehensiveness and diversity. We also look at some domain specific aspects and cross language information retrieval.

In order to improve document retrieval efficiency, we have first studied the use of pseudo relevance feedback based clustering approach. This approach identifies additional terms for query expansion from the initial set of documents by applying a clustering approach. Each cluster is assumed to represent an aspect and the additional terms are chosen from a set of diverse clusters. Using these additional terms, we perform query expansion and document retrieval as well.

In some cases user may be interested to know everything about the query topic. In such cases, it may be necessary to identify documents covering different aspects of the query topic. We present an approach for finding comprehensive documents with respect to the given query. The proposed approach analyzes the content of a document to find whether it contains information on various aspects of the query.

Using query terms as atomic units, one may not be able to retrieve all relevant

documents. A representation of the terms that capture the relation of the term with other terms is useful to understand and capture the semantics of the terms. We have developed an approach based on random indexing that encoded the context of the terms in query and documents and used them for improving document retrieval.

Understanding the intent of the query may be difficult from short and ambiguous query. By being able to identify the type of the user query, it may be possible to match the query to the desired type of the documents. This allows to organize the information in the documents under major topics. These topics could be exploited to perform retrieval of specific type of information. In this thesis, we have studied the tourism domain and explored the use of topics in document retrieval.

Finally we addressed the query translation problem in cross language document retrieval. The task is to translate the query from the source language to the target language and getting the right translation of the given query is of great importance. The dictionary may have multiple translations of a given word. Some terms may be absent in the dictionary. In such cases, one has to use other resources to formulate appropriate query in the target language. We addressed this problem by proposing a method based on corpus driven query suggestion approach. In this approach, co-occurrence based term associations are captured using a bi-lingual dictionary and bipartite network has been created using the terms co-occurring with source query and the translated query. Based on the term importance of the terms in the target language, top ranking terms are selected for query formulation. Based on this probable query terms in the target language, we perform document retrieval.

We have used the standard corpus released by Forum for Information Retrieval and Evaluation (FIRE) in many of our experiments and also a portion of web corpus having tourism related documents.

Keywords: Information retrieval, topic identification, comprehensiveness, pseudo relevance feedback, segment clustering, domain specific search, random indexing, term contexts, query translation, dictionary-based approach, cross language information retrieval, document retrieval and ranking.