ABSTRACT

The number of Web Users accessing information over Internet is increasing day by day. A huge amount of information on Internet is available in different languages that can be accessed by anybody at any time. Information Retrieval (IR) deals with finding useful information from a large collection of unstructured, structured and semi-structured data. Information Retrieval can be classified into different classes such as monolingual information retrieval (MIR), cross lingual information retrieval (CLIR) and multilingual information retrieval (MLIR) etc. In the current scenario, the diversity of information and language barriers are the serious issues for communication and cultural exchange across the world. To solve such barriers, cross language information retrieval (CLIR) systems, are nowadays in strong demand. CLIR refers to the information retrieval activity when the queries and the documents (to be retrieved) are in different languages.

In CLIR for searching relevant documents, three types of translation are possible: query or document or both. Translation is an important and difficult activity in CLIR. So, for effective retrieval of results through CLIR correct translation is needed. Back translation is an effective way to automatically evaluate the accuracy of query translation. Our analysis using back translation to evaluate the translation accuracy of queries has also revealed the same. The analysis using three popular translators shows that the query translation and back-translation is more effective with Google translator.

As compared to monolingual IR, the relevancy of retrieved documents in CLIR is often poor due to language barrier where highly occurring problems are query mismatching, multiple representations of query terms and un-translated query terms. Query Expansion (Q.E.) is the process of adding related and relevant terms to the original query to enhance its indexing capability in order to improve the relevancy of retrieved documents in CLIR.

Purpose of query expansion is to improve the performance and quality of retrieved information in CLIR. In this research work, Q.E. has been explored for a Hindi-English CLIR in which Hindi queries are used to search English documents. After query translation, retrieved

results are ranked using OkapiBM25 to arrange the most relevant document at the top for increasing the relevancy of retrieved documents using QE. Term Selection Value (TSV) is used to select the most appropriate term for Q.E.

We proposed architecture for Hindi-English CLIR using Q.E. to improve the relevancy of retrieved documents. In the first experiment, Q.E. is performed with and without OkapiBM25 ranking. The results shows that the relevancy (in term of mean average precision (MAP)) of retrieved documents is higher with OKapiBM25 as compare to the one without ranking. After Q.E using FIRE test collection, the relevancy of retrieved results improved further significantly.

Addition of term(s) in a query at appropriate location plays an important role in Q.E. Integration of words at proper location in a query also resolves the issue of drift query. It can change the meaning of required user's information if the Q.E. is performed by adding term(s) at inappropriate location. For this a new algorithm has been proposed and implemented which finds the location of expansion term(s) in the query with the help retrieved documents. In our experiment, using the proposed algorithm for Q.E., we obtained further improvement of results when compared with results without using any such algorithm for Q.E.

FIRE queries have been used throughout the research as it is a standard forum for IR/CLIR. In real time environment however, FIRE test collection could not be appropriate, so we created two more test collections, one using Snippets of retrieved documents of each query and another using the Nearest-Neighborhood word against each query words among the ranked documents. The whole process of query expansion has also been performed using each of these two test collections separately. We found that the results of Q.E. using Snippets test collections are better than other two test collections.

The work carried in this thesis clearly indicates that the performance of Hindi-English CLIR system can be improved significantly with query expansion using appropriate term(s) placed at appropriate location and the retrieved Snippets could greatly serve as the real time test collection for this purpose.