DOI 10.29042/2018-4480-4483

Helix Vol. 8(6): 4480- 4483

Cross Language Information Retrieval using Selective Documents Technique and Query Expansion

[1]Dr. Avinash J. Agrawal, [2]A. V. Zadgaonkar

[1,2] Dept. of CSE, Shri Ramdeobaba College of Engineering, & Management, Nagpur, 440013, India

Email: agrawalaj@rknec.edu, zadgaonkarav1@rknec.edu

Received: 20th September 2018, Accepted: 11th October 2018, Published: 31st October 2018

Abstract

This paper is devoted to a new method that uses combined technique of pseudo relevance feedback and co-occurring term technique for query expansion to improve cross language information retrieval along with selective document technique to reduce search space. The base is an Information Retrieval (IR) system which uses cosine similarity technique to retrieve and rank documents and a multilingual module using dictionary based translation where the tf-idf values are calculated and stored in inverted index which is term based index. The threshold value is used as a selection medium for documents which reduces the search space of index reducing the overall time for retrieval. A Query Expansion (QE) module is added in to the said system**.** The aim is to use QE to overcome the limitations of dictionary based translation, and to retrieve more relevant results. The system is validated with several measures, which shows the difference in the result before expansion and after expansion.

Keywords

Cross Language Information Retrieval, Cosine Similarity, Tf-idf, Pseudo Relevance Feedback, Query Expansion, Hindi-English Dictionary

Introduction

Cross Language Information Retrieval (CLIR) has been an important research field with the aim to provide users documents in languages different from that of query. A major challenge in CLIR is to bridge the language gap between documents and query. Query translation is now serving as a major cross-lingual mechanism in most current CLIR systems. A common approach in CLIR is to translate queries using dictionaries because of the simplicity and the availability of machine readable bilingual dictionaries. A major problem in dictionary-based CLIR systems is ambiguity. However, queries from users are often not clear or are too short, which produce more ambiguity in query translation, and reduce the accuracy of the cross language retrieval results. The language mismatch problem in CLIR is more serious than in monolingual information retrieval. Given a query in the source language, the translated query in the target language is built by selecting the "correct" translations from a list of candidate translations for each term in the initial query to build a structured query in the target language. There are two mutually exclusive techniques to address this problem. The single selection technique that tries to find one best translation for each term and the multiple selection technique, on the other hand, provides every possible translation available [8]. The detailed explanation about tf-idf, cosine similarity and inverted index is explained in previous work [5]. In this paper, pseudo relevance feedback technique is applied to improve the structured query translation by recalculating weights for query terms from top documents returned by the initial retrieval [6]. Then related terms are extracted from the top relevant documents using co-occurrence analysis, assuming that if two terms co-occur then they tend to be related. Our experiment is using a dictionary-based Hindi-English CLIR system which shows that this method helps to improve M.A.P. score. Experiments show that query expansion contributes only a minor improvement in precision; however it helps to retrieve more relevant documents. The combination of using query terms re-weighting and query expansion together is shown as a good solution.

There are many ways to solve the problem of ambiguity. Our aim is to find an efficient way to improve the performance of system, Section 2 describes some of the methods previously used for query expansion , and Section 3 describe the problem statement and proposed architecture. Section 4 explain the implementation of proposed work. Section 5 describes the analysis of the result obtained in detail.

Materials and Methods

This section presents some of the methods used for cross language information retrieval and query expansion. A paper is published by Benoit Gaillard, Jean-Leon Bouraoui, Emilie Guimier de Neef and Malek Boualem in year 2010 titled as "Query Expansion for Cross Language Information Retrieval Improvement" [2]. The idea of this paper was to use Query Expansion (QE) to overcome the problem of differences between translated data and human language. QE consists of adding new words to the initial query. It is divided into two modules CLIR module and QE module. In CLIR module before indexing the contents are translated. The original text is kept in the memory. QE module provides terms for

expansion of query Abdelghani Bellaachia and Ghita Amor-Tijani published a paper titled "Enhanced Query Expansion in English-Arabic CLIR" in Year 2008 [4]. In this paper concept of top retrieved documents is used to enhance given query. Related terms are extracted from the presumed relevant documents using co-occurrence analysis. Then the final expanded query is formed by adding those terms. The optimum effectiveness could be reached by applying Disambiguation. Using the Direct query expansion (DQE) technique, using thesaurus-based disambiguation the approach of QE is enhanced, considering that not all expanded terms are necessarily related to the query. For query expansion co-occurring words with the query terms in the set of top documents were added to the query. The terms that are directly related to the initial query are considered in the expanded query using DQE. After the query is tokenized and stemmed, query terms are translated. With this method Query Expansion proved to be effective.

Another paper published by Vivek Pemawat, Abhinav Saund and Anupam Agrawal in year 2010 tiltled "Hindi - English Based Cross Language Information Retrieval System for Allahabad Museum" [3]. The paper talks about the system developed for CLIR. The documents and images for query processing were related to Allahabad museum. The languages used were Hindi and English. The documents were stored in English language. A dictionary database was used for the conversion of Hindi words to English words, and for those not in the dictionary database, Hindi character to English character mapping was used. Vector based model was used for the retrieval of documents and images, all the documents were in English language. Clustering was done to classify documents into different classes and group to find so that we can know which document lies in which category and can find the similarity easily. Finally after retrieval of documents the output is showed in the language of user choice

A paper published by Authors Pratibha Bajpai and Parul Verma in Year 2014 titled "Cross Language Information Retrieval: In Indian Language Perspective" [1]. The paper is a review paper discussed issues in information retrieval from Indian language perspective. Also it presented the work done by various researchers in the field of Indian languages for Cross Language Information Retrieval.

Nowadays Cross Language Information Retrieval has become a crucial part. User wants to read the documents in language which they understand best. For that reason we need to develop this approach to provide accurate and better results. In particular for Hindi language not much work has been done and there are many limitations with the proposed approaches. The problem of ambiguity are still not solved completely. In the approach presented here a system is develop for CLIR that would particularly solve the problem of ambiguity using query expansion and provide better results. It uses selective document technique to reduce search space. The system is designed for the specific domain where documents related to only computer science is considered.

The documents related to computer science domain are collected in Hindi and English. The documents are Pre-processed i.e. Sentence detection, Tokenization and stop word removal is done. The Tf-Idf values for every term after tokenization is calculated and the term with respective documents ids and tf-idf value are stored in inverted index. The Hindi to English dictionary is built using the words from the processed documents. The query entered by the user can be in Hindi or English which is translated using dictionary. The query entered by the user after processing is translated. The translated terms are searched in the index and the documents above the specified threshold value are retrieved which are then used for cosine similarity calculation. If the user is not satisfied with the result the user selects query expansion option and the query is expanded using pseudo relevance feedback technique and co-occurring terms. After user selects the expansion terms, again the query is translated after processing the document ids are collected from inverted index and the cosine similarity is calculated which retrieves the relevant documents. The proposed system design is shown in figure 1.

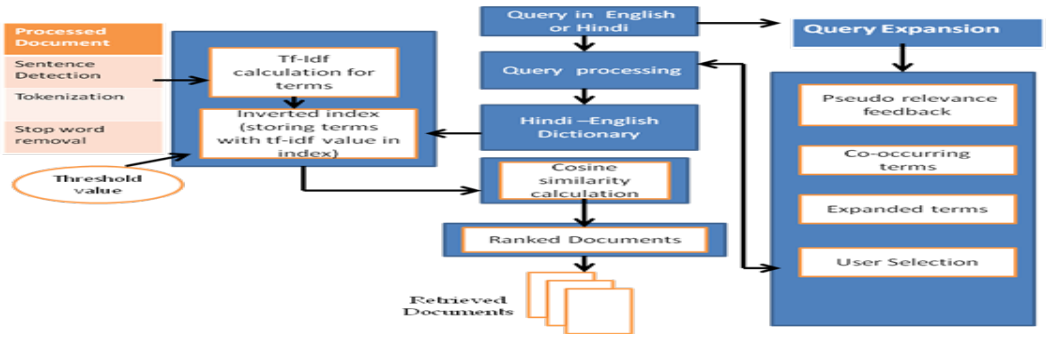

Figure 1. Proposed Framework

Pre-Processing of Documents:-The Documents are pre-processed before indexing. Sentence detection is done for every document using OpenNLP Toolkit. After sentence detection tokenization and stop word removal is done again using OpenNLP Toolkit, this is done for every term in document and its tf-idf value is calculated. The tf-idf value is calculated using tf-idf formula which is then stored in inverted index. Inverted Index is built using Lucene Library.

Vector Space Model:-Documents are represented as vectors by vector space model. It is used in relevancy rankings, and information retrieval the classic vector space model has the term-specific weights in the document vectors as the products of local and global parameters. It is known as term frequency-inverse document frequency model. Tf–idf weighting scheme are often used as a central tool in scoring and ranking by search engines. It is a numerical statistic that reflects the importance of the word for the document in a collection or corpus. The detailed explanation of tf-idf calculation is described in our previous work. [8] Using the cosine similarity between document d and query q the documents which are more similar to the query can be calculated. Cosine Similarity is used to calculate the similarity between document and query or two documents. Using the formula given below we can find out the similarity between a query and documents.

Inverted Index:-Inverted index provides access to the list of documents that contain the term in the query. An inverted index has postings lists, one associated with each term that appears in the collection. Inverted index is a data structure that we build while processing the documents that we are going to provide while answering the search queries. For a query, we use the index to return the list of documents which are relevant to the query. The inverted index contains mappings from terms to the documents that those terms appear in. Each term in the vocabulary is a key in the index whose value is its postings list. List of those documents in which the term appears is its postings list. We can also want to keep extra information in the index such as the number of different documents that the term appears in or the number of occurrences of the term in the whole collection. For a search engine indexing algorithm the inverted index data structure is a central component. The search engine implementation goal is to optimize the speed of the query. With the inverted index created, the query can be resolved by directly going to the word id in the inverted index.

English Hindi Dictionary: -For the conversion of Hindi words to English words, a dictionary database has been built. Hindi language consists of words which can have different meanings in English. Thus both the entries are stored in dictionary and while retrieving the documents both type of documents will be retrieved. To further improve the result query expansion option will be provided. We also have a option where the Hindi query can be typed in English language for those users who cannot type in Hindi. Below is the example of how the dictionary will look

English Word	Hindi Word	Hindi English Word
Computer	lax.kd	Sanganak
Study	i<kbZ djuk	Adhyayan
Structure	lajpuk	Sanrachana
Processing	çlaLdj.k	Prasanskaran
Engineering	vfHk;kaf=dh	Abhiyaantrikee

Table 1. English Hindi Dictionary

Pseudo Relevance Technique:-The following are the steps that we are performing for query expansion PRF assumes the top n documents from initial retrieval as being relevant and uses these pseudo-relevant documents to refine the query for the next retrieval

Step 1: Select top 10 documents after 1st retrieval if the documents retrieved are less than 10 consider only those documents that are retrieved

Step 2: Recalculate query terms weights based on term distribution in pseudo relevant documents. The following formula is used to recalculate query weight

$$W(t) = \sum count\,(t,d)/length(d)$$

Step 3: Change the query terms weight, use the new query weights to retrieve a new list of relevant documents for 2nd retrieval w (t) =new (tf-idf). Using new tf-idf we calculate cosine similarity after which we will get our 2nd set of retrieved documents

Step 4: After second retrieval select top 10 documents and extract the frequently co-occurring terms with the query terms. Extract Terms appearing more than once with the query term to restructure the query. The terms that are obtained are provided as option to user in form of checkbox the user can select multiple options from the provided list the selected options will be added to the query.

Step5: After getting the expanded query the initial process of collecting the documents ids and their respective tf-idf value from the inverted index which are above the specified threshold value is done. The tf-idf values of the collected documents are used to calculate the cosine similarity score. After which the third and final retrieval is done.

Results and Discussion

For experimental purpose dataset is manually created using results from google search. The Documents in Hindi and English are from computer science domain the collection of 150 documents are processed on which the dictionary is built dictionary contains 5500 terms that are used for translation of query terms. The total terms for which the tf-idf value is stored are 20561.The system is tested on set of 50 queries before expansion and after expansion**.** For the Analysis of the model, we used Mean Average Precision (MAP) and precision and recall method. MAP is the standard single-number measures for comparing search engines. Three sets of results are calculate table 3 shows the MAP score for system before expansion and after expansion. It shows improvements in MAP score after expansion

System	MAP
Original System	0.4605
System after Expansion	0.4756

Table 3. MAP Score Before Query Expansion and After Query Expansion

Conclusion

We have retrieved the documents for the user query and shown the relevant documents. The tf-idf values were used for representing the documents and query in vector form and cosine similarity was computed to retrieve and rank the documents relevant to the query. The terms were stored in a inverted index were the corresponding documents in which the terms appears and their tf-idf values in descending order were also stored A threshold value was decided and documents with tf-idf greater than or equal to that value where only considered for computing cosine similarity. This helped in reducing the time complexity. Domain specific dictionary was built, the problem of ambiguity which arises when Hindi terms are converted in English is handled by considering all the different meaning of term in English for a particular Hindi term (multiple selection technique).Further we combined the techniques of pseudo relevance feedback and co-occurring term .where the weights of term in query were recalculated and then were used to expand the query and retrieve new set of documents. The experimental results show that the MAP score show improvement in the system after expansion as compared to system before expansion. The Prcision and Recall also shows improvement.

References

[1] Pratibha Bajpai, Parul Verma “Cross Language Information Retrieval: In Indian Language Perspective” International Journal of Research in Engineering and Technology (IJRET) Jun-2014.

[2] Benoit Gaillard, Jean-Leon Bouraoui, Emilie Guimier de Neef, Malek Boualem “Query Expansion for Cross Language Information Retrieval Improvement” 2010 IEEE

[3] Vivek Pemawat, Abhinav Saund, Anupam Agrawal “Hindi - English Based Cross Language Information Retrieval System for Allahabad Museum” 2010 International Conference on Signal and Image Processing

[4] Abdelghani Bellaachia and Ghita Amor-Tijani “Enhanced Query Expansion in English-Arabic CLIR” 19th international conference of database and expert system application.

[5] Lam Tung Giang, Vo Trung Hung and Huynh Cong Phap, “Improve Cross Language Information Retrieval with Pseudo-Relevance Feedback” International Journal of Engineering Research & Technology (IJERT), June 2015

[6] Rekha Vaidyanathan ,Sujoy Das and Namita Srivastava “Query Expansion Strategy based on Pseudo Relevance Feedback and Term Weight Scheme for Monolingual Retrieval “International Journal of Computer ,November 2014

[7] Viatcheslav Yatsko, Snehal Dixit, AJ Agrawal, Sin Thi Yar Myint, Mie Mie Khin “TF* IDF Revisited” intelligence, 2013

[8] Aditi Agrawal, Avinash J Agrawal “Improving Performance of Hindi-English based Cross Language Information Retrieval using Selective Documents Technique and Query Expansion” International Journal of Science and Research (IJSR),May 2016

[9] Xuwen Wang, Qiang Zhang, Xiaojie Wang and Yueping Sun ”LDA Based Pseudo Relevance Feedback For Cross Language Information Retrieval” Proceedings of IEEE CCIS2012

[10] Avinash Agrawal, Ashwini “Information extraction using discourse analysis from newswires”, International Journal of Information Technology Convergence and Services (IJITCS), 2014

[11] Neha R Kasture, Avinash Agrawal,” A supervised Word Sense Disambiguation method using ontology and context knowledge”, Computer Engineering and Intelligent Systems,2012

[12] Avinash J Agrawal, OG Kakde “Semantic analysis of natural language queries using domain ontology for information access from database”, International Journal of Intelligent Systems and Applications, 2013

