Indian Languages IR using Latent Semantic Indexing

A.P.SivaKumar[1], Dr.P.Premchand[2], Dr.A.Govardhan[3]

[1]Assistant Professor, Department of Computer Science Engineering, JNTUACE, Anantapur
[2]Professor, Department of Computer Science Engineering, Osmania University, Hyderabad
[3]Principal & Professor, Department of Computer Science Engineering, JNTUHCE, Nachupalli

sivakumar.ap@gmail.com,p.premchand@uceou.edu,govardhan_cse@yahoo.co.in

Abstract. *Retrieving information from different languages may lead to many problems like polysemy and synonymy, which can be resolved by Latent Semantic Indexing (LSI) techniques. This paper uses the Singular Value Decomposition (SVD) of LSI technique to achieve effective indexing for English and Hindi languages. Parallel corpus consisting of both Hindi and English documents is created and is used for training and testing the system. Removing stop words from the documents is performed followed by stemming and normalization in order to reduce the feature space and to get language relations. Then, cosine similarity method is applied on query document and target document. Based on our experimental results it is proved that LSI based CLIR gets over the non-LSI based retrieval which have retrieval successes of 67% and 9% respectively.*

Keywords: Latent semantic indexing, Cross language information retrieval, Indexing, Singular value decomposition.

1 INTRODUCTION

Information Retrieval (IR) deals with representing, storing, organizing, and accessing information. This representation and organization of information is useful for user accessing. The main goal of Information Retrieval (IR) is to retrieve the information which is relevant to the users need. This Information Retrieval will be helpful in structuring of the language.

The demand for multilingual information is becoming profound as the users of the internet throughout the world are increasing. This demand creates a problem of retrieving documents in one language by specifying query in other language. This increasing necessity for retrieval of multilingual documents comes up with the new branch called Cross Lingual Information Retrieval (CLIR).

Cross Lingual Information retrieval makes use of user queries in one language (source language) and utilizes them in retrieval of documents in other language (target language). For example, if the user enters a query in Hindi language then relevant documents in English will be retrieved. These retrieved documents are semantically equal. Many information retrieval methods depend on the exact match between words in user queries and words in documents. The documents which contain the words in user query are returned to the user. So those methods will fail in retrieving the documents which do not match with the words in the user queries in a proper way. There are many standard methods like, Dictionary based method, Inverted indexing method, Probabilistic based methods are failed due to the consideration of words in user queries. The most familiar dictionary method for CLIR is also not giving efficient information retrieval, due to the limited number of indexing terms or words present in the dictionary method.

DOI : 10.5121/ijcsit.2011.3419

The contents of the paper are as follows. Section 2 outlines the previous work done by different institutions on Indexing. Section 3 gives the information regarding proposed system. Section 4 is about the experiment and results. Finally Section 5 includes future work and concludes the paper.

2 PREVIOUS WORK

Much work has already been done on CLIR systems and presently research is going on in many countries like India, Japan, China, and Portugal. Most of the proposed systems are based on indexing techniques like dictionary based indexing, inverted file system, probabilistic latent semantic indexing, ontology indexing, and language modeling which retrieve the documents based on the index terms. But, by using index terms we won't be able to get the documents which are relevant to the user query.

Using latent semantic indexing, cross language information retrieval can be performed automatically as described in [1].They tested the language independent depiction of the documents, irrespective of the user query, which means it may be short or long query. They used French and English parallel corpus for training and testing the system. They collected the corpus from Hansard collection. 982 documents were collected for training the system and 1500 documents for testing it. Totally they had used nearly 2482 documents. In English documents there are 2482 paragraphs and in French documents also there are 2482 paragraphs. The success rate in finding out the mate documents is 98%.

The reference [2] has used porter stemmer for stemming of the documents in English. Here they removed suffixes from the words. Stemming is done on the Cranfield200 collection. While stemming they calculated precision and recall. They tested porter stemmer algorithm on 10,000 vocabularies. The reduced words out of 10,000 are 1373 and the 3650 were not reduced. So by using porter stemmer the vocabulary size is nearly reduced by 1/3 rd of the original one.

The reference [3] illustrates the method of Turkish-English cross language information retrieval using LSI. In this they experimented on LSI using Singular Value Decomposition. The parallel corpus is collected from Skylife Magazine's website, which contains both Turkish as well as English articles. Those articles are converted by the interpreters. This corpus contains 1056 Turkish documents and 1056 English documents. Here each paragraph is taken as an individual document. They had matched paragraphs to their cross language mates. So finally there are 3602 document pairs and each single term is represented by document matrix. Out of 3602 documents 1801 documents are used for training the system and 1801 for testing the system. Longest Match Stemming algorithm is used for the stemming of the Turkish Documents and for English they used Porter stemmer. They had taken My SQL 5.1.11 Data base server for storing the documents. By using Latent Semantic Indexing the retrieval rate is 3 times more than the direct Matching. The success rate is 69%.

The reference [4] describes Portuguese-English Experiments using LSI. They used Los Angeles Times for English Documents only. Systran (translator) used for translating the 20 % of the English collection to Portuguese. The total documents in the collection are 22000. The success rate of the retrieval is nearly 99%.

The reference [5] describes Indexing by Latent Semantic Analysis. The method Singular Value Decomposition tested in this analysis, it gives the details about how to solve the problem of multiple terms referring to the same object. In this the relevant documents are characterized and identified properly. For example 12 term by 9 document matrix is decomposed by using SVD is given clearly.

The reference [6] describes the method of Latent Semantic Indexing Overview. It described some advantage of Liplike less dimensionality, polysemy, synonym and Term dependence .In the analysis of LSI they used 90,000 terms instead of 70,000 documents. So the term by document contains only 0.001% - 0.002% non zero, entries. To compute a [200], it had taken nearly 18 hours CPU time. In this LSI gave 16% improvement than original keyword method.

The reference [7] describes the method for retrieving of English-Greek documents using Latent Semantic Indexing for Cross Language Information Retrieval. The English and Turkish documents are clustered along the X-axis and Y-axis into a two dimensional vector. Parsing mechanism is used. Here the terms should be appearing at least more than once in the database. This paper mainly focuses on the query matching within the data base. Folding-in is another technique for the LSI generated database already exists. In this Folding- in technique each new document is represented as weighted sum of component document vector, this is appended to the existing documents.

The reference [8] describes the method of Latent semantic Indexing a fast Track Tutorial. The reference [9] describes the method of Singular Value Decomposition Singular Value Decomposition (SVD) is a mathematical technique used for reduce the dimension of a matrix. This tutorial describes how the documents are decomposed from a single matrix. This gives the relation between the correlated documents and uncorrelated documents. In this tutorial they illustrated the two dimensional data points.

The reference [10] describes the method of indexing documents by a combination of keywords neglecting the relationship between semantic words. The reference [11] describes new Chinese term measurement and MLU extractor process that none well on small corpora, and approach to the selection of MLU's in a more accurate manner. The reference [12] describes probabilistic latent semantic indexing (PLSI) models using word segmentation. Their result show that correct word segmentation improve precision of information retrievals and index based on keywords extraction obtains highest accuracy rate to PLSI model.

3 PROPOSED SYSTEM

This latent semantic indexing is the best approach for mapping of each document and query vector in to a reduced dimensional space. This is based on concept matching rather than matching of index terms. The proposed system follows many steps in retrieval of documents.

Indexing is a data structure built on the text to speed up searching. This indexing is very simple for a single language, but when coming to multilingual it is quiet difficult. So for this we are proposing Latent Semantic Indexing (LSI), by Using Singular Value Decomposition (SVD). Here input is a set of documents d1, d2, d3... and user query is q=q1, q2..., we are giving the entire document as a query.

We applied a ranking method for the documents retrieval, it gives the order of the documents (top) relevant to the user query.

In this we scale the term frequency by using following formula

$$W(t,d) = 1+\log(tf(t,d)) \quad \text{if } tf(t,d) > 0 \quad (1)$$
$$=0 \quad \text{otherwise}$$
$$Idf(t) = \log(N/df(t)).$$

Where Idf = inverted document frequency.

N = number of documents in the collection.

Here first we collect the information which is semantically equal and perform stemming on that corpus. After stemming of the documents both are placed in the same space vector. Each paragraph is considered as a single term-by -document matrix. Latent Semantic

Indexing uses a mathematical method called Singular Value Decomposition. This SVD is used for reducing dimensions of the term-by-document matrix. The formula for SVD as follows:

SVD splits a matrix (A) in to 3 matrices.

$$A=UXV^T \qquad (2)$$

Here,

U is a matrix containing the columns as the eigenvectors of AA^T. It is a concept by term matrix.

X is a matrix, the diagonal elements are singular values of A. It is a concept by concept matrix.

V is a matrix containing the columns as the Eigen vectors of the A^T a matrix. It is a concept by document matrix.

From these observations a suitable rank value (k value) is to be taken to reduce the semantic space. The selection of K value is depending on the parallel corpus that we are using in this experiment.

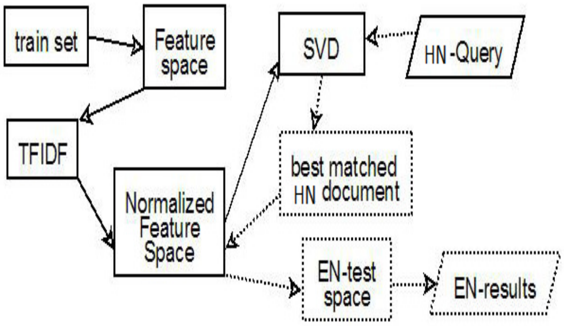

Fig.1: System overview

In this, we have created a system that can search the cross language mate of a given document. First we train the system with bilingual documents. In this stage, we have stemmed the English documents using porter stemmer and we also stem the Hindi documents manually. After stemming the documents using corresponding stemmers we remove the stop words to increase the retrieval performance.

By counting the frequency of each word in documents we created a term-by-document matrix (Feature-space). We Normalized the Feature-space using Term Frequency – Inverse Document Frequency (TF-IDF), because longer documents may affect the retrieval results. Then the normalized term-by-document matrix has been decomposed to U, S, and V matrices using singular values decomposition (SVD). For this we have used JAMA package which contains all the classes and interfaces which are used for decomposing the Feature-space.

After training the system using bilingual documents, the documents in the Hindi database have been queried to find the cross language mates. To find the similarity between the documents we use cosine similarity. For the given query document we retrieve the document which gives the value of cosine similarity almost equal to one.

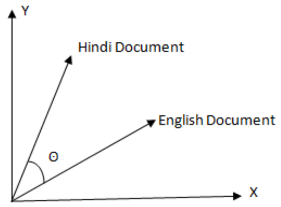

Fig.2: Cosine Similarity

Cosine similarity can be calculated by the following formula,

$$cosine\ similarity(q,d) = sim(q^T U_k \Sigma_k^{-1}, d^T U_k \Sigma_k^{-1}) \quad (3)$$

Where,

q is the query document

d is the target document

k is the rank value

4 EXPERIMENT

In this we have taken parallel corpus. We retrieved documents from India Gov[1]. This contains both Hindi and English documents which are semantically equal. Documents in both languages have been divided into paragraphs. Each paragraph is divided into a single document. So these documents are mapped to the respective translation language paragraphs. The mapping data is stored in MYSQL data base server.

The corpus consists of 180 Hindi and 180 English parallel documents. So for this purpose we used every paragraph as a single document.

Table 1. Example Document

English Document	Hindi Document
India & the World India's foreign policy seeks to safeguard the country's enlightened self-interest. The primary objective of India's foreign policy is to promote and maintain a peaceful and stable external environment in which the domestic tasks of inclusive economic development and poverty alleviation can progress rapidly and without obstacles. Given the high priority attached by the Government of India to socio-economic development, India has a vital stake in a supportive external environment both in our region and globally.	भारत और विश्व भारत की विदेश नीति में देश के विवेकपूर्ण स्व-हित की रक्षा करने पर बल दिया जाता है। भारत की विदेश नीति का प्राथमिक उद्देश्य शांतिपूर्ण स्थिर बाहरी परिवेश को बढ़ावा देना और उसे बनाए रखना है, जिसमें समग्र आर्थिक और गरीबी उन्मूलन के घरेलू लक्ष्यों को तेजी से और बाधाओं से मुक्त माहौल में आगे बढ़ाया जा सकें। सरकार द्वारा सामाजिक- आर्थिक विकास को उच्च प्राथमिकता दिए जाने को देखते हुए, क्षेत्रीय और वैश्विक दोनों ही स्तरों पर सहयोगपूर्ण बाहरी वातावरण कायम करने में भारत की महत्वपूर्ण भूमिका है।

[1] India Gov http://www.india.gov.in/

Porter stemmer has been used for stemming of English documents. For Hindi documents we performed manual stemming. So after stemming the stop word list is as follows.

Table 2. Top 20 Stop Word List

English		Hindi	
Word	Count	Word	Count
The	969	में	550
Of	577	और	445
And	483	की	378
In	389	को	241
To	337	का	215
A	202	लिए	166
For	161	से	165
With	111	ने	124
Is	108	एक	110
On	105	किया	108
As	102	पर	105
By	100	है	95
Was	73	करने	75
From	63	साथ	72
Has	63	इस	69
Also	57	भी	67
At	56	द्‌वारा	66
An	43	यह	51

As mentioned earlier each paragraph is taken as an individual document. We have mapped the paragraphs to their cross linguistic mates in the MY SQL data base server. So totally we have 360 document pairs created. In that each of them is represented as a single document in term -by-document relation.

The paragraphs which are present in the same document are semantically equal. So we used 180 documents for training the system and 180 documents for testing the system. The document set is shown in the below table.

Table 3. Corpus Overview

Set	Number of Documents	
	English	Hindi
Corpus	360	360
Training set	180	80
Hindi Test Set	00	180
English Test Set	180	00

After training the system, the documents in the Hindi testing set have been queried to the system to find their cross language mates. Cosine similarity is used to find the similarity among the documents. We also tested the system with different ranks (k values). Based on k value the results are shown in the table below.

Table 4. Cross Language Mate Retrieval Results

k	Return Rank										
	Hindi document as query										
	1	***2***	***3***	***4***	***5***	***6***	***7***	***8***	***9***	***10***	***total***
dm	01	01	01	02	02	-	06	-	01	02	16
40	40	04	01	02	-	-	-	03	-	01	51
80	80	04	02	01		04		01	01		93
120	118	03	01				01				123
160	158	03							01		162

dm: denotes direct match

The performance of the system is evaluated if we find the mate of query document in the retrieval result. After submission of query, retrieval results are ranked according to their similarity to the query document. We have given 180 test documents one by one as a query and expected to find its mate in the query results. We considered the query results as successful, if the mate of query document appears in the first 10 of ranked retrieval results. The above table shows the number of successful queries according to rank order of the mate document. For example, if we consider k=40 experiment, we obtained the mate of query document at the first rank for 40 documents. The first row in the table shows the results of CLIR, if we make direct match between documents, where no LSI and TFIDF is used. The above table also shows that, using TFIDF and LSI increases the query performance by approximately 3 times when direct matching is considered. The table shows that as k value increases the results are better but compile time is increased.

5 CONCLUSION AND FUTURE WORK

Various experiments by other researchers carried out using Latent Semantic Indexing method for other test data and other languages have produced good results. Our study on other Indian languages like Telugu, Tamil and Marathi has proved that using LSI methods increase the retrieval performance. Availability of standard test collect, remain major concern for testing LSI method. And also another important question number and size of documents to be used during training.

In this experiment we have mainly focused on improving a Hindi-English cross language information retrieval using latent semantic indexing. For that we collected parallel corpus from India.gov.in web site [12] and performed singular value decomposition to get a CLIR system. Our tests depicted that the latent semantic indexing improves the results three times to that of direct matching method. We also observed that if the value of k increases then there is no consistent performance improvement.

The CLIR system we have developed will work well for document queries but it was less informative for user generated queries. So, much work needs to be done in order to make this system work for user generated queries.

References

Susan T. Dumais, Michael L. Littman, and Thomas K.Landauer.: Automatic cross language retrieval using latent semantic indexing

Porter,M.: The Porter Stemmer is at http://www.tartarus.org/~martin/PorterStemmer/

Baturman Sen, 3Burak Gunel .: Turkish – English Cross Language Information Retrieval using LSI

Viviane Moreira Orengo, Christian Huyck .: Portuguese-English Experiments using Latent Semantic Indexing

Scott Deerwester .: Indexing by Latent Semantic Analysis Barbara Rosario .: Latent Semantic Indexing: An overview

Paul G.Young .: Cross Language Information Retrieval Using Latent Semantic Indexing.

Dr. Edel Garcia .: Latent Semantic Indexing (LSI) A Fast Track Tutorial Kirk Baker .: Singular Value Decomposition Tutorial

Dr. Edel Garcia.: Singular Value Decomposition (SVD) A Fast Track TutorialEmmett J. Ientilucci .: Using the Singular Value Decomposition

N.Tazzite, A.Yousfi.: Design and Implementation of an Information Retrieval System by Integrating Semantic Knowledge in the Indexing Phase

Chengye Lu, Yue Xu. :Web-Based Query Translation for English-Chinese CLIR

Xie Fang 1, Liu Xiaoguang 2, Hu Quan 3.:Comparison Probabilistic Latent Semantic Indexing Model In Chinese Information Retrieval

The corpus India Gov.:http://www.india.gov.in

Muhamad Taufik Abdullah1, Fatimah Ahmad1, Ramlan Mahmod1, and Tengku Mohd Tengku.: Application of Latent Semantic Indexing on Malay-English Cross Language Information Retrieval.

Md. Maruf Hasan and Yuji Matsumoto.: Japanese-Chinese Cross-Language Information Retrieval: An Interlingua Approach

Thomas Hofmann.: Probabilistic Latent Semantic Indexing

Georgios Paltoglou, Michail Salampasis, Foris Lazarinis.:Indexing and Retrieval of a Greek Corpus

Jay M. Ponte and W. Bruce *Croft*.: A Language Modeling Approach to Information Retrieval

Chung-hsin Lin and Hsinchun Chen.:An Automatic Indexing and Neural Network Approach to Concept Retrieval and Classification Multilingual (Chinese-English) Documents

J. Xu, R. Weischedel, C. Nguyen.: Evaluating a Probabilistic Model for Cross-Lingual Information Retrieval

Ulrich Schiel, lanna M. S. F. de Sousa,.:Semi-Automatic Indexing of Documents with a Multilingual Thesaurus

Hyun-Jo Lee, Hyeong-Il Kim and Jae-Woo Chang.:An Efficient High-Dimensional Indexing Scheme using a Clustering Technique for Content-based Retrieval

A. Lopez “Statistical machine translation”, In ACM Computing Surveys 40(3), Article 8, pages 1–49, August 2008.

S.E. Robertson, S.Walker,” Some simple effective approximations to the 2-Poisson model for probabilistic weighted retrieval”, Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval, Dublin, Ireland Pages: 232 – 241, 1994, ISBN:0-387-19889-X

N. Jian-Yun, M. Simard, P. Isabelle, R. Durand, “Cross-language information retrieval based on parallel texts and automatic mining of parallel texts from the web” , Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval, Berkeley, California, United States ,Pages: 74 – 81, 1999, ISBN:1-58113-096-1

J. Xu, R. Weischedel, C. Nguyen, "Evaluating a Probabilistic Model for Cross-Lingual Information Retrieval", Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval, New Orleans, Louisiana, United States , Pages: 105 - 110 , 2001, ISBN:1-58113-331-6

P.A. Chew, B.W. Bader, T.G. Kolda, A.Abdelali, "Cross-language information retrieval using parafac2", Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining, San Jose, California, USA, Pages: 143 – 152 , 2007, ISBN:978- 1-59593-609-7

C.C. Yang, C. Wei, K.W.Li, "Cross-lingual thesaurus for multilingual knowledge management", Decision Support Systems, Volume 45 , Issue 3 (June 2008), Pages 596-605 , 2008,ISSN:0167-9236

J. Gao, J. Nie, M.Zhou, "Statistical query translation model for crosslanguage information retrieval", ACM Transactions on Asian Language Information Processing (TALIP), Volume 5 , Issue 4 (December 2006), Pages: 323 - 359 , 2006, ISSN:1530-0226