Volume 5, Issue 3, March 2015 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

A Review on Information Retrieval in Indian Multilingual Languages

M. S. Madankar
Assistant Professor, Department of Computer Science and Engineering,
G. H. Raisoni College of Engineering, Nagpur, India

Abstract— In today's world of globalization, local language database and retrieval is essential for the developing nations like India. As our nation is diversified by languages and only 10% of population is aware of English language, this diversity of languages is becoming barrier to understand and acquainted in digital world. So the information retrieval (IR) has been an active field of research for decades, for much of its history it has had a very strong bias towards English as the language of choice for research and evaluation purposes. It has been found that when services are provided in local languages, it has been strongly accepted and used. The Internet is no longer monolingual, as the non- English content (Hindi, Marathi) is also growing rapidly. User want to access the information in his native language and retrieve the information in same language is a big issue. One of the crucial challenges in Cross lingual information retrieval is the retrieval of relevant information for a query expressed in as native language.

Keywords— Information Retrieval, Cross Language Information Retrieval, Multilanguage Information Retrieval, Multilingual Indian Languages.

I. INTRODUCTION

As the non- English content (Hindi, Marathi) is growing rapidly, Internet is no longer monolingual. World Wide Web is growing rapidly, and the content on Web of languages other than English is also increasing rapidly compared to English. In the past few years Hindi content has also increased rapidly on the Web. All major news papers, publication houses and Government departments have setup their web sites in Hindi Language. The globalization is reducing the significance of national borders in terms of trade and information exchange.[5] Hindi is the third most widely-spoken language in the world Marathi is also most widely spoken language in Maharashtra. Information Retrieval in Hindi, Marathi and English language is getting popularity. Currently, Google provides transliteration in Bengali, Gujarati, Hindi, Kannada, Malayalam, Marathi, Nepali, Punjabi, Tamil, Telugu and Urdu and offers searching in 13 languages, Hindi, Tamil, Kannada, Malayalam and Telugu to name a few.

Information Retrieval will be very effective research area and society get the benefit such that User can try to access the information in this native language and retrieve the information in same language without knowing in which language the information is stored in database. This system can be effectively used in application areas like e-governance, agriculture, rural health , education, national resource planning, disaster management, information kiosks etc where people from all walks of life are involved.

As far as development in IR with respect to Indian languages is concerned, a lot work is going on particularly in the field of information retrieval. Research is also going on in other related areas as well such as NLP machine translation etc. Various regional languages have been taken into consideration by researchers for IR. Even government organization like TDIL (Technology Development for Indian Languages) has made significant contributions for standardization of Indian Languages on web. In the proceeding section we present the various developments in Indian Information Retrieval, Cross language Information Retrieval, Multilanguage Information Retrieval, Query Processing and NLP system. Literature survey is classified according to different areas of NLP.

II. RELATED WORK

Cross language and Multilingual Information Retrieval

CLIR(Cross language information retrieval) deals with asking question in one language and retrieving documents in other language. MLIR(Multilingual information retrieval) deal with asking questions in one or more languages and retrieving documents in one or more different languages. Author S.M. Chaware et.al, proposed a approach of storing and processing multilingual data. Author proposed an efficient and easy way to convert local language keyword to English. The result has been tested using three local language interface namely Hindi, Marathi and Gujrathi and tested using an application like 'Shopping Mall', where there is need of local language query processing.[1]

In [2], Author N. Swapna, N. hareen kumar, B. Pdmaja Rani introduces the concept of information retrieval i.e. BLIR, CLIR and MLIR.

In [4], Author Sung Shim described a system that uses various cross-language information retrieval (CLIR) methods to provide search engines with capability of automatic bilingual search. The system accepted search queries in a local language and converted them into the corresponding queries in English using CLIR methods. The search results in English may be further translated into the local language as currently being done by some search engines.

In [5], Author Anurag Seetha, Sujoy Das, And M.Kumar evaluated the English-Hindi Cross Language Information Retrieval System Based on Dictionary Based Query Translation Method. In this strategy, author replaced the word of the query with the corresponding lexicographically, first exact match in the dictionary to produce a version of the query that can be compared with the documents in the Hindi test collection.

In [12], discussed the CLIR such as Chinese-English information retrieval. Author focused on how to obtain effective web pages and evaluate translation candidates are two challenging issues. In this paper, an approach based on maximum entropy method (MEM) was proposed to obtain effective Web pages. For obtaining a correct translation list, author established English-Chinese, Chinese-English special dictionary. The proposed method considered the context and predicts possible English meanings for searching and gives 86.8% accuracy.

In [14] Author described a method of fully automated cross-language information retrieval which does not require any query translation. Namely, monolingual queries retrieve documents from a multilingual collection which includes items from the query's source language. This is achieved by a method which combines the construction of a multilingual semantic space using Latent Semantic Indexing (LSI) and the clustering ability of Self-organizing Maps (SOM) for the generation of multilingual semantic categories.

[16] This paper explained a description on cross-lingual IR, its challenges and current methods and techniques to overcome problems for efficient and resourceful searching. This report meant for reviewing not all but some of the latest researches in the area of cross-lingual IR

In [22], Author Manoj Kumar Chinnakotla, Sagar Ranadive, Pushpak Bhattacharyya and Om P. Damani Department of CSE, IIT Bombay presented Hindi -> English and Marathi->English CLIR systems.

Author Raju Kora et,al, investigated that the query translation plays a central role in MLIR research. The language-independent indexing technology is used to process the text collections of English, Telugu and Hindi languages. Author has used multilingual dictionary based word-by-word query translation. The experimental results was evaluated to analyze and compare the performance of Average Precision (APIR) and Mean Average Precision (MAPIR) metrics of IR system with esteem to the Average Precision (APMLIR) and Mean Average Precision (MAPMLIR) metrics in MLIR system.[3]

III. METHODOLOGY

Information retrieval is concerned with making database, searching it and retrieving information. It is a separate field within computer science (closer to databases), but Information Retrieval relies on some Natural Language Processing Methods.

3.1Query Translation methods:

In [7] author described a method that uses query expansion to improve multilingual information retrieval. The backbone is an Information Retrieval (IR) system based on a search engine and a multilingual module based on statistical machine translation of documents. The aim is to use QE to overcome the limitations of machine translation, and to retrieve more relevant results. Here author suggested a novel approach for the CLIR technique that consists in indexing translated documents. This technique provided better results than query translation.

In [8] author Yilu Zhou, Jialun Qin, Hsinchun Chen, Jay F. Nunamaker explained the two fundamental approaches: query translation or document translation**.** There are three main approaches in CLIR and MLIR: using machine translation (MT), a parallel corpus, or a bilingual dictionary. This is the most popular approach because of its simplicity and the wide availability of machine-readable dictionaries.

3.2 Query Optimization techniques

In [9] Kumar Sourabh and Vibhakar Mansotra discussed the problem of Low Recall in Hindi Language. Authors have introduced the solution for Low Recall problem in Hindi Language. The queries supplied by the user are saved in query log which is a separate database used for processing the keywords for their further optimization. To accomplish this purpose author used the keyword ranking approach. Here author was worked only on the Hindi information retrieval method, where frontend and backend were both in Hindi language.

Here K. Ganesan and G. Siva proposed a way of processing multilingual information wherein the backend uses English language and the front end uses local language like Tamil. For searching multilingual information, there exist two methodologies, one is based on Phonemes and another is based on semantic matching. For semantic matching query crawler algorithm was proposed and for phonemes word crawler was proposed [10].

3.3 Multilingual Query Processing:

The main purpose of Natural Language Query Processing is to interpret an English sentence and hence a complementary action is taken. Querying to databases in natural language is convenient method for data access, especially for newbie's who have less knowledge about complicated database query languages such as SQL.

In [11], author proposed a novel approach for multilingual query processing where author proposed phonetic distance based measure for searching proper name data in Indian language script. System allows query in language of user's

choice. A cross lingual search is conducted where query is one language and document is in another language. Phonetic distance based process involved three steps :

1. Conversion of query into a language independent Common Ground (CG) representation.
2. Query matching at CG level in the phonetic space.
3. Conversion of search result to the query language.

Here author had generated a distance matrix that can be used by the DTW algorithms for matching the query in database record. Future scope of this implementation is to improve the efficiency of DTW algorithms and implement on more Indian languages.

A new concept of query processor based on finite automata and natural language processing is implemented in [6].

A system that is capable of handling simple queries with standard join conditions are introduced here but not all forms of SQL queries are supported.

3.4 Existing Algorithms for Query translation:

There are various query translation algorithms like Iterative Page Rank-Style Algorithm, Query translation based approach using bi-lingual dictionary. Transliterated using a simple rule based approach corpus to return the 'k' closet English transliteration.[22],[25]. Meaning Matching Approach based on translation probability.[23]

1. Main approaches in CLIR & MLIR system for query translation discuss in[8]
   a. Machine translation
   b. Parallel Corpus
   c. Bilingual dictionary
   d. Thesaurus based method
2. Query translation based approach using Multi-lingual dictionary-word by word query translation.[3]

IV. PROPOSED PLAN

Above are the various techniques for information retrieval in multilingual and cross lingual information retrieval.

One of the crucial challenges in Cross lingual and multilingual information retrieval is the retrieval of relevant information for a query expressed in as native language. The amount of non-English information on the Web has proliferated so rapidly in recent years that it often difficult for a user to retrieve documents in an unfamiliar language. For retrieving the information from the database in native language, translation of the whole document is very difficult and time consuming. Hence a new approach agent based information retrieval in multilingual Indian languages is proposed to analyse the access time for query processing, accuracy and precision

V. CONCLUSIONS

The India is a multilingual country. But on web mostly information are available in the form on English content. The information retrieval in their native language is very big issue and research area for the new researchers. Here I have studied various query translation methods, query optimization techniques, query processing methods, cross language and multi language information retrieval system, their advantages and drawbacks. By various surveys it is conclude that user can not retrieve information easily by submitting the query in his native language and information retrieve in the same language without knowing that data stored on the web in any other language.

REFERENCES

[1] S.M.Chaware, Srikantha Rao. (2009). "Information Retrieval in Multilingual Environment," Second International Conference on Emerging Trends in Engineering and Technology, ICETET-09, IEEE Computer Society, pp 648-652

[2] N. Swapna1 , N.Hareen kumar2, B. Padmaja Rani3, (September 2012) "Information Retrieval In Indian Languages: A Case Study On Cross-Lingual And Multi-Lingual", International Journal of Research in Computer and Communication technology, IJRCCT, ISSN 2278-5841, Vol 1, Issue 4,

[3] Raju Korra#1, Pothula Sujatha*2, Sidige Chetana*3, Madarapu Naresh Kumar, (June 2011) "Performance Evaluation of Multilingual Information Retrieval (MLIR) System over Information Retrieval (IR) System" IEEE-International Conference on Recent Trends in Information Technology, ICRTIT 2011 MIT, Anna University, Chennai. Pp-722-727.

[4] Sung J. Shim (2005) "Using Cross-Language Information Retrieval Methods for Bilingual Search of the Web", IEEE Computer Society, International Conference on Computational Intelligence for Modelling, Control and Automation, and International Conference on Intelligent Agents, Web Technologies and Internet Commerce.

[5] Anurag Seetha, Sujoy Das, M. Kumar (2007) "Evaluation of the English-Hindi Cross Language Information Retrieval System Based on Dictionary Based Query Translation Method", 10th International Conference on Information Technology, IEEE Computer Society, pp- 56-61.

[6] Jasmeen Kaur , Bhawna chauhan , Jatinder Kaur Korepal, "Implementation of Query Processor Using Automata", International Journal of Scientific and Research Publications, Volume 3, Issue 5, May 2013 1 ISSN 2250-3153

[7] Benoit Gaillard, Jean-Leon Bouraoui, Emilie Guimier de Neef, Malek Boualem, "Query Expansion for Cross Language Information Retrieval Improvement" 978-1-4244-4840-1/10/ IEEE-2010.

[8] Yilu Zhou, Jialun Qin, Hsinchun Chen, Jay F. Nunamaker , "Multilingual Web Retrieval: An Experiment on a Multilingual Business Intelligence Portal" Proceedings of the 38th Hawaii International Conference on System Sciences – 2005 pp-1-10.

[9] Kumar Sourabh, Vibhakar Mansotra, "Query Optimization A Solution for Low Recall Problem in Hindi Language Information Retrieval", International Journal of Computer Applications (0975 – 8887) Volume 55– No.17, October 2012-pp-6-17.

[10] k. Ganesan and G. Siva, (2007) "Multilingual Querying and information processing", Information Technology Journal ISSN-1812-5638, PP 751-755

[11] Sriram S. Partha talukdar, Sameer Badskar, "Phonetic distance based cross lingual search",
Qin Chen1,Lei Liu2,Lin Ma "Application of Maximum Entropy Method in Chinese-English Cross Language Information Retrieval", IEEE 2008, pp-1192-1195

[12] Jolanta Mizera-Pietraszko, "Interactive Document Retrieval from Multilingual Digital Repositories", IEE 2009, pp- 423-428.

[13] Nikolaos Ampazis, Helen Iakovaki , "Cross-Language Information Retrieval using Latent Semantic Indexing and Self-organizing Maps", IEEE 2004, pp-751-755

[14] Mohammad Shamsul Arefin*, Yasu, "Multilingual Content Management in Web Environment", Chittagong University of Engineering & Technology, Bangladesh, IEEE 2011.

[15] B.Ashwin Kumar, "Profound Survey on Cross Language Information Retrieval Methods (CLIR)", 2012 Second International Conference on Advanced Computing & Communication Technologies, IEEE Computer Society, pp-64-68.

[16] Bao-Quoc Ho, Van B. Dang, Minh V. Luong and Thuy T.B. Dong, "English-Vietnamese Cross-Language Information", IEEE 2008, pp-107-113.

[17] Zhao Rongying, "Visual analysis on the research of cross-language information retrieval", 2008 IEEE, pp-107-113.

[18] Sumam Mary Idicula, David Peter, S "A Multilingual Query Processing System using Software Agents", Journal of Digital Information Management _ Volume 5 Number 6 _ December 2007, pp-385-390

[19] Bin Xue, "Research on Multi-agents Information Retrieval System Based on Intelligent Evolution", 2nd International Conference on Computer Science and Network Technology, pp-1042-1045

[20] Ashish Almeida, Pushpak Bhattacharyya, "Using Morphology to Improve Marathi Monolingual Information Retrieval" *FIRE-2008.*

[21] Manoj Kumar Chinnakotla, Sagar Ranadive, Pushpak Bhattacharyya and Om P. Damani, "Hindi and Marathi to English Cross Language Information Retrieval"at CLEF 2007.

[22] Tan Xu1 and Douglas W. Oard, "Maryland: English-Hindi CLIR" FIRE-2008

[23] TechnologyDr.M.Hanumathappa1, Mallamma.V. Reddy2 "Natural Language Identification and Translation Tool for Natural Language Processing", International Journal of Science and Applied Information , Volume 1, No.4, September – October 2012, ISSN No. 2278-3083- pp-107-112.

[24] Mallamma V Reddy, Dr. M. Hanumanthappa, "Kannada and Telugu Native Languages to English Cross Language Information Retrieval" Mallamma V Reddy et al, / (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 2 (5) , 2011, pp-1876-1880.

[25] Dr.S.Saraswathi, Asma Siddhiqaa.M, Kalaimagal.K, Kalaiyarasi.M, "BiLingual Information Retrieval System for English and Tamil" JOURNAL OF COMPUTING, VOLUME 2, ISSUE 4, APRIL 2010, ISSN 2151-9617 pp-85-90.

[26] Pinaki Bhaskar, Amitava Das, Partha Pakray and Sivaji Bandyopadhyay "Theme Based English and Bengali Ad-hoc Monolingual Information Retrieval in FIRE 2010", Forum for Information Retrieval Evaluation (FIRE) 2010.

[27] Kumar Sourabh Vibhakar Mansotra, "Factors Affecting the Performance of Hindi Language searching on web: An Experimental Study", International Journal Of Scientific & Engineering Research, Volume 3, Issue 4, April-2012 1 ISSN 2229-5518, pp-1-15.

[28] Kumar Sourabh, Vibhakar Mansotra, "An Experimental Analysis on the Influence of English on Hindi Language Information Retrieval", *International Journal of Computer Applications (0975 – 8887) Volume 41– No.11, March 2012*

[29] Roger Bradford, John Pozniak, "Combining Modern Machine Translation Software
with LSI for Cross-lingual Information Processing", 2014 11th International Conference on Information Technology: New Generations, pp-65-72.

[30] Maria Pia di Buono, Mario Monteleone, Federica, Johanna Monti, "Knowledge Management and Cultural Heritage Repositories Cross-Lingual Information Retrieval Strategies", 2013 IEEE, pp-295-302.

[31] Sandeep Chaware, 2Srikantha Rao,"Ontology Supported Inference System for Hindi and
Marathi", 2012 IEEE,

[32] Fuminori Kimura, Akira Maeda, Kenji Hatano, Jun Miyazaki, "Cross-Language Information Retrieval by Domain Restriction using Web Directory Structure", Proceedings of the 41st Hawaii International Conference on System Sciences – 2008, pp-1-8.

[33] Hassan Alam, Aman Kumar, "Multi-Lingual Author Identification and Linguistic Feature Extraction – a Machine Learning Approach", 2013 IEEE, pp-386- 389

BIOGRAPHY

Mangala S. Madankar received the B.E degree in Computer Engineering from the Nagpur University, India, in 2004 and the M.E. degree (Distinction with CGPA 9.14) in Wireless Communication and Computing, in 2012. She worked with industry for 3 years 2005 to 2008 focused on php/mysql. She joined G.H.Raisoni College of Engineering, Nagpur, India, as Lecturer in 2008 and became an Assistant Professor in 2012 and currently working**.** Her area of specialization are Wireless Communication, Android, Mobile Computing, Network Security, Theory of Computation, Natural Language Processing.

