



A SURVEY ON INFORMATION RETRIEVAL METHODS IN REGIONAL LANGUAGES

H C Vijayalakshmi*

Department of Computer Science
JSS Science and Technology University, Mysuru, India
vijilakshmihc@sice.ac.in

Bhavana S Dixit

Department of Computer Science
JSS Science and Technology University, Mysuru, India
dixit.bhavana8@gmail.com

Manuscript History

Number: IRJCS/RS/Vol.06/Issue07/JLCS10082

Received: 02, July 2019

Final Correction: 11, July 2019

Final Accepted: 20, July 2019

Published: July 2019

Citation: Vijayalakshmi & Dixit (2019). A Survey on Information Retrieval Methods in Regional Languages. IRJCS:: International Research Journal of Computer Science, Volume VI, 654-658. doi://10.26562/IRJCS.2019.JNCS10082

Editor: Dr.A.Arul L.S, Chief Editor, IRJCS, AM Publications, India

Copyright: ©2019 This is an open access article distributed under the terms of the Creative Commons Attribution License, Which Permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited

Abstract— Data available on the web is growing at an exponential rate, creating Knowledge or extracting information is of paramount importance. Information Retrieval (IR) plays a crucial role in Knowledge management as it helps us to find the relevant information from the existing data. This paper compares the performance of keyword-based retrieval and other architectural styles in information retrieval system with ontology-based retrieval on documents in regional language

Keywords— Ontology; Information Retrieval; Knowledge Base; natural language queries; SQL; natural language processing;

I. INTRODUCTION

Information Retrieval can be defined as finding the required information from the relevant data source which can be structured or unstructured arrangement of data. Ontology is one of the well-explored models in the Information Retrieval fields. Ontology is “a formal, explicit, specification of the shared conceptualization” [17]. The main advantage of using Ontology is that it helps in re-use of stored information and it helps to induce conceptual knowledge in the input search query. Ontologies form significant part in an Information Retrieval system. It is commonly accepted as explicit specification of a conceptualization (Gruber,1995). Ontology based Information system overcomes the disadvantages of keyword-based information retrieval methods such as high computation expense and low system efficiency. This methodology is constructed to re-use the stored information. It constructs the user’s query in a well-built way and eliminates ambiguities. It computes the semantic understanding among the input documents, and thus increases the precision of the information retrieval system. It provides the diversified knowledge integration and generic knowledge representation. Real world example of ontology-based architecture is Texpresso. Texpresso is Ontology based system implemented by the California Institute of Technology to extract the information on biological literature [14].

This paper is organized as follows. A brief introduction is given in section 1, section 2 discusses the various methodology and architecture, section 3 addresses the Comparative Analysis followed by conclusion and future work in section 4.

II. DETAILED DESCRIPTION OF ARCHITECTURAL STYLE

R Mahesh et al [3], have discussed in detail the challenges faced while using natural language processing approach, in regional language. Researchers have used Shallow syntactic approach to implement the conceptual information of the Knowledge base. The system is designed to perform analysis on shallow syntactic and semantics, strategies for generating inter-query contexts and for generating back query to the user. R Mahesh et al [3], have discussed in detail the challenges faced while using natural language processing approach, in regional language. Researchers have used Shallow syntactic approach to implement the conceptual information of the Knowledge base. The system is designed to perform analysis on shallow syntactic and semantics, strategies for generating inter-query contexts and for generating back query to the user.

In [4] Debasis Mandal et al, have presented two experiments on two cross lingual and one monolingual English text retrieval at Cross Language Evaluation Forum (CLEF) in ad-hoc track. Data source used in this experiment is 'Shabdanjali', a Hindi English bilingual lexicon consisting approximately 26000 words which is built by IIIT Hyderabad. Corpus processing, Query Generation, Document Retrieval are the implementation stages adopted in this research. Automatic Query Generation and Machine Translation are the key methodologies adopted for this experiment. The best Mean Average Precision (MAP) values for Bengali, Hindi and English queries for our experiment were 7.26, 4.77 and 36.49 respectively. Observations from this experiment pointed out by the authors are, The system is built with limited resource scenario with basic Machine Translation Approach and also it is pointed out that good computational approaches such as Pseudo Relevance Feedback for query expansion, Multiword expression detection, Word sense disambiguation and Structured query translation will increase the performance of the system.

Jagadeesh Jagarlamudi et al [5], have proposed the word alignment table learned by a Statistical Machine Translation (SMT) and trained on parallelly aligned sentences, in cross language information retrieval system. The model implemented by authors take the input query in Hindi, Telugu, Bengali, Marathi and Tamil and relevant English documents are retrieved. 73% result was achieved by this methodology for mono lingual system and 43% accuracy was achieved for cross lingual systems. In the training phase of the machine translation system, statistically aligned Hindi to English word alignments were used as the training data. Word by word translation is carried out for the given Hindi Input query. Translation probability is calculated for each word and the word with the highest probability is picked. The following are the experiments conducted on the proposed approach. (1) CLEF Data set for Adhoc track. (2) Word alignment table as Bilingual Dictionary.

Dr. S Saraswathi et al [6], have developed a protocol, to build bilingual information retrieval system for English and Tamil. The system is built by following modules. (a) User Interface (b) Keyword extraction (c) Information Retrieval (d) Information Extraction. Ontological tree is adopted to build a tree where every node has two entries from source language as well as target language. Page Rank algorithm which assigns numerical weightage to documents to measure the relative importance is implemented by researchers to rank the documents in the domain. It is shown by the authors that the system accuracy ranges up to 40 % (for English) to 60% (for Tamil). This model can handle any two language for cross language Information Retrieval.

Jagdish S Kallimani et al [7]. have designed a model for text summarization from large documents. Auto Sum is the Text summarization method used to carry out various computations such as percentage summarization, keyword extraction to generate accurate results. Auto sum is used to parse the input document and rank the sentences in the document based on their sentence score and the content from these sentences are used to summarize the input document. It is observed that percentage of summary size is directly proportional to percentage of common sentences.

In [9] Raja Sukumara A et al, the authors have proposed a new model to process the Natural language query in Malayalam. They have further discussed about the problems involved in natural language processing. They have built a strong domain specific linguistic model which is used in semantic computational method in the system. The proposed model analyzes the input query both syntactically and semantically. Once the input query is understood by the system, it is made to initiate a reasoning process to identify the type of query and the result slots which are required. The query investigator is then used to identify the latent information of the input query submitted by the user. Researchers have demoed that model has 87.50% Precision and Recall is 80% for 70 input queries. Authors have concluded that with more robust methods of semantic computational model the scope of the system can be extended to retrieve documents.

Murugan et al In [10], have discussed various computational styles in Ontology based Information Retrieval methods such as Vector Space Model, Probabilistic Models, Context aware model, Semantic Based model. Semantic similarity is applied to derive the concepts from the knowledge base. Exploring equivalent words using Word net and mathematical model to evaluate weightage of concepts are the stages in semantic similarity in an Information Retrieval system. Semantic association method is adopted to derive the direct and indirect association of the concepts in Knowledge base.

Semantically indexed elements are computed by Query document relevance and query concept recognition is executed to identify the candidate relevance in Ontology and used to evaluate the accuracy of relevant documents retrieved. Authors have derived that semantic association and semantic similarity approach has more Precision and hence lead to retrieve more relevant documents. It can be observed that Information Retrieval (IR) methods are discussed in general and no key focus is given to challenges faced for IR in regional languages.

S.P. Bansu et al, in [11] have discussed the importance of ontology-based retrieval system and have established that Ontology based IR models are more efficient than keyword-based approach. Authors have proved that conceptual framework proposed will overcome the limitations of keyword-based retrieval system and will increase the hit-miss ratio by rendering the relevant information to the users. The model implemented by authors take Marathi natural language query and retrieve documents in Marathi language. Each token in the input query is made to run through the ontological tree built from the knowledge base of the language. The system developed by researchers is made to handle language translation, semantic matching and Information Retrieval (IR) The authors have concluded by extending the idea of applying the framework to other languages.

Mangala Madankar et al. in [12], have presented a detailed survey paper for Cross Lingual Information Retrieval (CLIR) methods, which are Machine Translation, Bilingual dictionary, Parallel Corpora, Morphological Analyzer, Transliteration, word sense disambiguation. The authors have focused the discussion on Machine translation (MT) methodology. Corpus based Machine translation, Dictionary based Machine Translation, Example based machine translation are the key methods presented by the researchers. Key focus of this paper is Machine Translation (MT) approach. Authors have deep dived into MT approach where architecture of Rule based, corpus based, Dictionary based, and Example based Machine Translation are discussed in detail. Researchers have discussed not all, but some of the latest methodologies in Information Retrieval field.

Vijay Kumar Sharma et al. in [13], have proposed a Wikipedia API based query translation approach. N-gram is a concept defined in Natural Language Processing (NLP) to process sequence of N number of words. The method used in NLP to predict the occurrence of the words based on the occurrence of N-1 previous words. This technique is used to tokenize the multi word query. Each N-gram is given as a search query for Wikipedia API to retrieve article titles in the source language, and the target language inter-wiki link is used to retrieve the title of the article if it matches with the N-gram. Else N-gram is merged by removing the white spaces and newly formed N-gram is searches in the Wikipedia knowledge base and the source language article titles are retrieved. If no title is selected with 80% likelihood for the translation, then title with the maximum likelihood is selected. Terrier search engine which supports vector space and BM25 are used to retrieve documents in the target language. This approach provides 0.2685 Mean Average Precision (MAP).

Manasamithra P et al. in [15] have discussed hybrid system to query and retrieve the data from the database using NLP in English. M-way B Tree is used for the keywords storage and semantic methods are used to retrieve the data. The system has following modules (1) Pre-processing. This method is used to eliminate the noise in the data. (2) Natural Language Processor. This approach used the combination of semantic and keyword-based approach and Stanford Dependency Parser is used to implement this technique. (3) Knowledge base. In Order to increase the efficiency of the application well known Data structure M-way B-Tree is used. For Table names, attributes, constants, escape character / stop words separate B-Tree are used. (4) Query Translator. Based on the comparisons tokens are classified and based on this SQL queries are formulated. The system implemented by authors successfully handle the Natural Language Query given by user in English language, but they have not considered the languages in other Indian Regional language.

Pratibha Bajpai et. Al [16] have identified the problem of retrieving the relevant documents when the document language and the language in which input query is placed are different. In their work, optimization in Document Information Retrieval system is achieved by Two level disambiguation algorithm. The methodology followed makes the source language words unambiguous while translating them to target language at two levels. To aim for the likelihood of the translation in terms of search query First level deals with the translation candidates in pairs, and Second level aims to find the translation which achieved better probability. The Mean Average Precision (MAP) of two-level disambiguation algorithm which is more than 75% of monolingual search with both search engines has proved the effectiveness of the algorithm and no favouritism of search engine. Addition of Component Analyzer to the two-level disambiguation method in the basic cross language information retrieval method, increased effectiveness of the system. The researchers have achieved good accuracy and quality target language translations.

Language Independent Information Retrieval from Web (LIIRW) is implemented by R. Seethalakshmi et al in [17]. The system modules are categorized as follows. (a) Translator (b) Crawler and Indexer (c) Search Engine. As the authors call "software_morph_parser" is translator element in the system. The module is encompassed of natural language processing principles. Authors have implemented "Dyn_crawler" as the crawler and indexer, to crawl through the documents and retrieve the meta tag associated with it.

Input for “Dyn_crawler” is the output of “software_morph_parser”. There are two kinds of database used in this, Translator Database and Indexed Database. Authors have developed the prototype for language independent Information Retrieval with Linux Apache MySQL PHP (LAMP) Architecture.

III. COMPARATIVE ANALYSIS

A comparative analysis of various proposed methods by different researchers is tabulated which includes the paper name, Dataset used in the implementation methodology, brief explanation of the method, Accuracy and is there any scope for improvement sections.

TABLE I- COMPARITIVE ANALYSIS OF INFORMATION RETRIEVAL TECHNIQUES

S.No	Paper Name (Authors)	Dataset used	Approach	Result Accuracy	Scope for Improve ment
1.	Prathibha Bajpai et al [16]	Hindi Word Net	Two level disambiguation Model	81.1%	Yes
2.	Vijay Kumar Sharma et al [13]	Forum for Information Retrieval Evaluation (FIRE) 2010 and 2011 Data set	N-gram Technique	0.2685 Mean Average Precision (MAP)	Yes
3.	Jagdeesh S Kallimani et al [7]	Proprietary Data set	Extractive Summarization Method – Auto Sum	Not Quantified	Yes
4.	Dr. R.M.K Sinha et al [3]	Proprietary Data set	Human Computer NLP Schema	Applicable only in restricted domain.	Yes
5.	Debasis Mandal et al [4]	Linguistic Data consortium	Parallel corpora-based approach to build statistical lexicon	50% – 80%	Yes
6.	Raji Sukumara A et al [9]	Proprietary Data set	Natural Language Query Processing & Ontological information retrieval system	Precision – 87.50% Recall – 80%	Yes
7.	Dr. S Saraswathi et al [6]	Proprietary Data set	Ontological Tree	40% - 60%	Yes
8.	Shilpali Pankaj Bansu et al [11]	Proprietary Data set	Conceptual Framework of Ontology	Not Quantified	Yes
9.	Jagadeesh Jagarlamudi et al [5]	Cross language Evaluation Forum (CLEF) 2007 Data set	(a) CLEF Data set for ADHOC Track (b) Word alignment table as bilingual dictionary	54.4% - Cross Lingual IR System , 73% - Monolingual IR system	Yes
10.	Kalyani Lokhande et al [14]	FIRE 2010 set of Marathi News Corpus	Cross Language Information Retrieval System using Query Expansion	Precision and Recall lies between 0.5 to 1	Yes
11.	Manasa Mithra P et al	Multiple proprietary Data sets	M-way B-Tree is used in keyword based approach and Stanford Dependency Parser was used for semantic analysis	Not Quantified	Yes
12.	Comfort T. Akinribido et al [8]	Proprietary Data set	Fuzzy Ontological Information Retrieval System (FOIRS)	Not Quantified	Yes
13.	Vallet D et al [2]	Proprietary Data set	Ontology based Schema and semi-automatic annotation of Documents	Not Quantified	Yes

IV. CONCLUSIONS

With the advent of the Information Retrieval technology, concept of Ontology is playing a major role in Information Retrieval when the source and target languages are different. Comparisons in this paper proves that Information Retrieval in a cross-language domain is improved by use of Ontology based approach. This paper puts the limelight on the example where Ontology based systems are used in real world scenarios such as academic systems, Military systems too. With the approaches studied in this paper, a new model will be developed on Ontology based Information Retrieval for Kannada Language in our future work. As there is very few literatures available to retrieve the documents when query is placed in Kannada language. Hence, we conclude that there is lot of scope for exploring document retrieval in the Indian Regional Language Kannada

ACKNOWLEDGMENT

We express our sincere gratitude to the pioneer researchers who developed the methodologies that has enabled us to compare and understand various Information Retrieval algorithms.

REFERENCES

1. Studer R, Benjamins V R, Fensel D. "Knowledge Engineering: Principals and methods." Data and Knowledge Engineering, pp 161-197, March 1998
2. Vallet D., Fernández M., Castells P. An Ontology-Based Information Retrieval Model. In: Gómez-Pérez A., Euzenat J. (eds) The Semantic Web: Research and Applications. ESWC 2005. Lecture Notes in Computer Science, vol 3532. Springer, Berlin, Heidelberg, pp 455 – 470, 2005
3. Raveendranatha P. Mahesh and Koustuv Sinha. "On Design of a Question-Answering Interface for Hindi in a Restricted Domain.". Proceedings of the 2006 International Conference on Artificial Intelligence, ICAI, Las Vegas, Nevada, USA pp 319-324, 2006
4. Debasis Mandal , Sandipan Dandapat , Mayank Guptha , Pratyush Banarjee , Sudeshna Sarkar, " Bengali and Hindi to english Cross Language text retrieval under limited Resources" Cross Language Evaluation Forum , Budapest,Hungary, pp 19-21, 2007.
5. Jagadeesh Jagarlamudi, A Kumaran "Cross lingual Information Retrieval system for Indian Languages", The 2nd International workshop on "Cross Lingual Information Access",pp 80-87 , 2008
6. Dr. S Saraswathi, Asma Siddhiqaa M , Kalaimagal K, Kalaiyarasi M "Bilingual Information Retrieval system for English and Tamil" , Journal of Computing , Vol 2 , 2010
7. Jagdeesh S Kallimani, Srinivas KG, Eswara Reddy B "Information Retrieval by text summarization for an Indian Regional Language", International Conference on Natural, Language Processing and Knowledge Engineering, Beijing, pp. 1-4, 2010
8. Comfort T. Akinribido, Babajide S. Afolabi, Bernard I. Akhigbe, Ifio J. Udo. "A Fuzzy-Ontology Based Information Retrieval System for Relevant Feedback." International Journal of Computer Science, World Academy of Science, Engineering and Technology,8(1),pp.382-389,2011
9. Raji Sukumara, Babu Anto P, "Intelligent Query Processing in Malayalam", International Journal on Computer Science and Applications, Vol 3 , pp 51-59 ,April 2013
10. Sakti Murugan R, P Shanti Bala, Dr.G Aghila "Ontology based Information Retrieval -An analysis" , International Journal of Advanced Research in Computer Science and Software Engineering , Volume 3, Issue 10,pp 486-493, 2013
11. S. P. Bansu, S. S. Govilkar and J. W. Bakal, "Conceptual framework for Ontology Based Information Retrieval System for Indian regional language," International Conference on Advances in Communication and Computing Technologies (ICACACT 2014), Mumbai, pp. 1-4, 2014
12. Mangala Madankar, Dr.M B Chandak, Nekita Chavhan , "Information System and Machine Translation :A Review", Procedia Computer Science , pp 845-850, 2016
13. Information Retrieval system for Indian Languages", The 2nd International workshop on "Cross Lingual Information Access",pp 80-87 , 2008
14. Lokhande, Kalyani & Tayade, Dhanashree. "English- Marathi Cross Language Information Retrieval System". International Journal of Advanced Research in Computer Science and Software Engineering.
15. Manasamithra P., H.C Vijayalakshmi, "NLP for Information Retrieval using B Trees", International Journal of Computer Applications(0975-8887) Volume 182-No5, July 2018
16. Pratibha Bajpai, Parul Verma, Syed Q Abbas "English - Hindi Cross Language Information Retrieval System", Journal of computer science, 14(5) , pp 705 – 703, 2018
17. R. Seethalakshmi, Ankur Agrawal, Ranjit Ranjan, "Language Independent Information Retrieval from Web", 2019