A Survey of Information Retrieval Models for Malayalam Language Processing

Arjun Babu
Department of Computer Science
College of Engineering Poonjar.

Sindhu L.
Department of Computer Science
College of Engineering Poonjar.

ABSTRACT

Information retrieval is an important research area in computer science. Information retrieval is concerned with the storage of text documents and their subsequent retrieval in response to user's requests for information. There are several Information Retrieval models developed for efficient document retrieval. In this survey paper we describe major IR models which are used for various document retrieval purposes. A suitable method for an effective Malayalam monolingual information retrieval system is also proposed here.

General Terms

Information Retrieval (IR), Malayalam, IR Models

Keywords

Boolean Model, vector space Model, Probabilistic Model, Bayesian network Model, Inference Network Model, Latent Semantic Indexing, Neural Network Model, Ontology

1. INTRODUCTION

Information Retrieval is the process of finding relevant information or a document which satisfies the user's information need. A good monolingual information retrieval finds a document that satisfies the information need from within large collection (Usually stored on computer). In the retrieval process the user submits the query in natural language and the system responds with a set of relevant documents in natural language. Given the user query, the system has to retrieve the documents related to the query and if the size of the document collection is large, indexing technique must be used for better performance [5].There are several IR systems developed such as in English, Arabic, Chinese, Spanish etc. It is very limited in the case of Indian languages.

Malayalam is a Dravidian language which is a free word order and highly agglutinative in nature. The government of India declared Malayalam as a classical language in 2013.It is the mother tongue of three Crores of peoples in Kerala. Lacks of people living in urban area are very poor in English. These people need information about government, nation, rules, agriculture, etc. A Malayalam IR system can help those people find precise answer. Malayalam is free order language and morphologically rich, so it is difficult to develop an IR model. The ultimate aim of this survey is to study different IR models and find an effective IR model for Malayalam information retrieval.

The paper is organized as follows: section 2 brief introduction of types of IR, Section 3, followed by IR models. Section 4 brief description of related works and section 5 covers conclusion

2. TYPES OF IR

In classical IR search engine, both the query and retrieved documents are in the same language. The variation of IR is Bilingual Information Retrieval (BLIR), Cross Lingual Information Retrieval (CLIR), and Multi Lingual Information Retrieval (MLIR). But to perform these variants of IR a variety of translation methods are needed. CLIR deals with asking questions in one language and retrieving document in another language. There are various works done in CLIR Nikesh P.L.et.al [6] describes about the methodology of CLIR system in which the system responds Malayalam documents when the user queried in English. Another work in CLIR JinxiXu.et.al [1] describes cross lingual IR in which queries is in English and documents are in Chinese and Spanish. In CLIR the query and the document to be retrieved are different, so queries need to be translated, the translation process causes reduction in retrieval performance.

Multi Lingual Information Retrieval (MLIR) system deals with asking questions in one or more language and retrieving documents in one or more different languages. The paper [5] deals with Chinese, English and Japanese MLIR using the complex knowledge representation formalism for information retrieval.

3. IR MODELS

The fundamental classical IR models, includes Boolean model, Vector space model and probabilistic model, Alternative probabilistic model includes Bayesian network model and inference network model and Alternative Algebraic model describes latent semantic indexing and Neural network model. The rest of this section briefly describes these models.

3.1 Boolean Model

The Boolean model is a simple retrieval model based on set theory and Boolean algebra. In this model the queries are specified as Boolean expression such as AND, OR and NOT [3].For example, the query" all the hotels in Las Vegas or San Diego, but not Chicago" is typed by the user as:

[[Las&Vegas]|[San&Diego]]&hotel&!Chicago]

Each term is represented as 0|1 vector form. To answer query: take the vector for Las Vegas, San Diego and Chicago(complemented) and perform a bitwise AND .The output of the system is a list of relevant documents. The Boolean is very rigid AND means "all" and OR means "any" .All matched document will be returned. It predicts that each document is either relevant or non-relevant. There is no partial match to the query condition. Partial match allows retrieval of documents that approximate the query conditions.

3.2 Probabilistic Model

The probabilistic retrieval model is based on the probability relevance, which states that an information retrieval system is rank the documents according to their probability of relevance to the query, given all evidence available. Documents and queries are represented by binary vectors and each vector element indicating whether a document attribute or term occurs in the document or query, or not Instead of probabilities. For probabilistic model the index term weight variables are binary. In this model the documents are ranked according to their decreasing order of probability.

3.3 Vector space model

In this model, documents and queries are represented as vectors in multidimensional space. Then the retrieval is based on the similarity between the query vector and document vectors. The retrieved documents are ranked according to their similarity. The similarity is based on the occurrence frequencies of the keywords in the query and in the documents. in this models the documents and queries assigned a non-binary weight. Document and query is represented as

$$d_j = (w_{1j}, w_{2j}, \ldots, w_{n,j})$$

$$q_k = (w_{1k}, w_{2k}, \ldots, w_{n,k})$$

Vector Space Model have been introduce term weight scheme known as t_f-id_f weighting [7, 14]. These weights have a term frequency (t_f) factor measuring the frequency of occurrence of the terms in the document or query texts and an inverse document frequency (id_f) factor measuring the inverse of the number of documents that contain a query or document term [4].

The t_f factor is given by,

$$t_f=freq_{i,j}/max_i freq_{i,j}$$

The id_f factor is given by,

$$id_f=log(N/n_i)$$

The t_f-id_f weighting scheme uses weights which is given by, $w_{i,j}= t_f\times id_{f..}$The cosine similarity measure work better. The formula is the same as the inner product, but it is normalized by the length of the documents and the length of the query. The cosine measures the angles between the two vectors (the higher the cosine value closer to, the smaller the angle between Document vector and the query vector, therefore a more relevant document).

$$sim(d_j, d_q) = \frac{\overset{\rho}{d}_j \cdot \overset{\rho}{d}_q}{\left|\overset{\rho}{d}_j\right|\left|\overset{\rho}{d}_q\right|} = \frac{\sum_{i=1}^{n} w_{i,j} w_{i,q}}{\sqrt{\sum_{i=1}^{n} w_{i,j}^2}\sqrt{\sum_{i=1}^{n} w_{i,q}^2}}$$

3.4 Alternative Probabilistic Model

In alternative probabilistic model here we discuss Bayesian Belief network and Inference networks.

3.4.1 Bayesian network Model

Bayesian network is a graphical model that represents a set of random variables and their conditional dependencies via directed acyclic graph (DAG). Bayesian networks are also known as belief networks, probabilistic independence networks, influence diagrams and causal nets.

In Bayesian network model the set of all keywords as the universe of discourse U, this is the sample space. Let t be the total number of keywords. Then U={k_1,k_2,…..k_3}.For each keyword k_i. This variable is one indicates that keyword is observed. A document d_j indicates that a set of selected keywords. If all the variables are in on state indicates that the document has been observed.

3.4.2 Inference Network Model

In inference network model random variables are associated with index term, documents and query. All of these variables are represented as nodes in an inference network.

3.5 Alternative Algebraic model

In this model we describe latent semantic indexing and neural network model.

3.5.1 Latent Semantic Indexing

LSI is a concept based retrieval method that exploits the idea of vector space method and singular value decomposition (SVD).LSI can retrieve documents even when they do not share any words with the query. It is similar to the concept of vector space model. Neelam Phadnis and Jayant Gadge[10] explains latent semantic indexing by using singular value decomposition(SVD).SVD means matrix A can be decomposed into a product of three other matrices

$A = U\Sigma V^T$, Where

U is an orthogonal matrix,

Σ is a diagonal matrix and

V is the transpose of an orthogonal matrix. .

Neelam Phadnis and Jayant Gadge[10] this paper proposes a Latent Semantic Indexing Algorithm for efficient retrieval of documents. The LSI approach is useful for finding texts in large collection of data. It solves the problem of vocabulary problems such as synonymy and polysemy. In future this approach can be combined with WorldNet to improve web information retrieval.

3.5.2 Neural Network Model

Neural network models are a simplified graph representation of interconnected neurons in the human brain. The nodes in the graph are processing units and edges play the role of synaptic connections and a weight is assigned to each edge of neural network. The query term nodes are the ones which initiate the inference process by sending signals to the document term nodes. Following that, the document term nodes might themselves generate signals to the document nodes.Igor MOKRIS and Lenka SKOVAJSOVA [12] describes neural net work model of information retrieval system for Slovak language.

3.5.3 Ontology-based Information Retrieval

Conceptual specification of a word or a term is called ontology. An Ontological tree is defined as entities and grouping of entities, classification of entities in a hierarchical order and grouping of entities based on similarities and differences. Each keyword identified is matched with entry in every node in the Ontological tree. The exact location of the keyword in the tree is identified.

4. RELATED WORKS

Jinxi Xu.et.al[1] proposes a probabilistic cross lingual retrieval system. The system uses a generative model to find the probability that a document in one language is relevant given query in another language. It produces slightly better results than using a machine translation system for CLIR. The probabilistic CLIR system achieves roughly 90% of monolingual performance in retrieving Chinese documents and 85% in retrieving Spanish documents.

Pinaki Bhaskar.et.al[2] explain monolingual information retrieval task for English and Bengali languages. Documents are clustered using a set of theme keywords. Information retrieval is based on these set of theme keywords. Each language data consists of four consecutive years of news from the two reputed newspapers.TF-IDF weighting scheme is used for information retrieval. The limitation of the system is only certain query topics were quite good.

Jeongwoo ko.et.al [5] proposes two probabilistic models for answering ranking in the multilingual QAS, which find exact answers to a natural language questions written in different languages. The independent prediction model directly estimates a probability that an answer is correct given a multiple answer relevance features and answer similarity features. The joint prediction model uses an undirected graphical model to estimate the joint probability of all answer candidates from which the probability of an individual candidate is inferred. One advantage of the joint prediction model is that which is useful for list questions and JP model is less efficient than the independent model.

NikeshP.L.et.al[6] describes about an English-Malayalam information retrieval system. Here the system retrieves Malayalam documents in response to query in English. If the input is in English a translation process is needed for that a bilingual English-Malayalam dictionary is used. The user query is passed through processes like tokenizing, stop word removal and stemming. The documents are indexed and stored in a hash table called the inverted index file. The index is usually noun and proper nouns. Vector Space Model and tf-idf weighting scheme is used for document ranking and retrieval. The limitation of the system is for CLIR need translation it reduces the performance of MLIR .

Reshma O.K.et.al [8] describe an efficient information retrieval for Malayalam produces a good results using vector space model. The document and query preprocessor consists of Tokenizer, POS tagging, Stemming and Symantic mapping. For IR index terms are obtained after stemming the noun terms of each document. The system uses a vector space model and tf-idf weighting scheme to assign weights to the index terms of the document and query.

Marco Antonio Pinheiro de Cristo.et.al[9] explain how Bayesian networks can be used to represent the classic vector space model. Bayesian network models were first introduced in IR by Turtle and Croft. In their model, index terms, documents and user queries are seen as events and are represented as nodes in a Bayesian network. This model performs better than traditional probabilistic models for document ranking. A second model was proposed by Ribeiro-Neto and Muntz[4] .A combination of vector space model and Bayesian network. It yields better results than the use of a vector space ranking alone. A third model was proposed by Acid can be used to efficiently compute the relevance probabilities of the documents. Marco Antonio Pinheiro de Cristo.et.al [9] demonstrates that the combination of evidence from past queries with the vector space ranking yields better results than the use of a vector space ranking alone.

A.P.SivaKumar.et.al [11] uses LSI technique with Singular Value Decomposition (SVD) to achieve effective indexing for English and Hindi languages. This paper uses the Singular Value Decomposition (SVD) of LSI technique and it solves the problem of polysemy and synonymy. Cosine similarity method is applied on query document and target document. Applied a ranking method for the documents retrieval, it gives the order of the documents relevant to the user query. This paper also shows that, using TFIDF and LSI increases the query performance by approximately 3 times when direct matching is considered. The limitation of the system is It is expensive so in practice it works only in small text collection.

Igor MOKRIS and Lenka SKOVAJSOVA [12] describes neural net work model of information retrieval system for Slovak language. Slovak language is more complicated therefore; it is difficult to recognize keywords in Slovak text. Using neural net work model the Slovak text analysis can be simplified.The system is divided into three subsystems, administrator, indexation and user subsystem. The first sub layer of the system is query sub layer, the second one is keyword sub layer and third one is document sub layer. For Slovak language it produces better results.

S.Saraswathi.et.al [13] describes bilingual information retrieval for English and Tamil based on the keywords with the use of the Ontological tree. Usage of the Ontological tree is the striking feature of this system. A Part-Of-Speech (POS) Tagger is used to determine the keywords from the given query. Each keyword identified is matched with entry in every node in the Ontological tree finally, the solution for the query is translated back to the query language (if necessary) and produced to the user. The efficiency of the system is 40% for English and 60% for Tamil.

Table 1. Comparison of different IR Models.

Model	Related work	Method	Advantage	Limitations
Boolean Model	No works done.	i) Based set theory and Boolean algebra. ii) Queries are represented by Boolean expressions iii) Terms are combined with AND,OR ,&NOT.	i) Easy to implement. ii) Computationally efficient.	i) No partial matching ii) No ranking iii) No weighting of terms. iv) Difficult to control output.
Probabilistic Model	i)Jinxi Xu.et.al[1] ii)Jeongwoo ko.et.al [5]	i) Based on probability ranking principle. ii)Based on relevance and non-relevance of data.	i)Ranking of document ii)Does not consider index inside a document (ie,all weights are binary)	i) Only partial ranking of documents. ii) Estimation of needed probabilities. iii) Prior knowledge needed.

Vector space Model	i)NikeshP.L.et.al[6] ii)Reshma O.K.et.al[8]	i) Documents and queries are represented by vectors. ii) A weighting Scheme is used. iii) Rank documents by similarity. iv. Using cosine similarity.	i)Term weighting improve retrieval performance ii) Partial matching allows documents that approximate query conditions.	i)Assumes independence of index terms
Latent Semantic Indexing	i)A.P.SivaKumar.et.al [11] ii) Neelam Phadnis and Jayant Gadge[9]s	i) Retrieving text based on concept. ii) Use Singular Value Decomposition.	i) It solves the problem of polysemy and synonymy. ii) Retrieve documents even they do not share any keyword in the query.	i) It does not allow fast retrieval. ii)LSI technique expensive so it works only in small text collection
Neural Network Model	Igor MOKRIS and Lenka SKOVAJSOVA [12]	i) Based on neurons. ii) A weight assigned to each edge of Neurons.	i) it allows the retrieval of documents which are not initially related to the query terms.	i) Difficulty with understanding the internal processes of the network.
Inference network model.	No works done.	i) Random variables are associated with index term, documents and query. ii) Combine the evidence.	i) it provides a framework in which different ranking strategies can be used.	i) Boolean query formulation.
Bayesian Network model	i)Marco Antonio Pinheiro de Cristo.et.al[9]	i) A probabilistic graphical model. i) Represents a set of random variables and their conditional dependencies via directed acyclic graph (DAG).	i) Providing a separation between the document space and the query space.	i) Specify the queries as Boolean expressions.
Model	**Related work**	**Method**	**Advantage**	**Limitations**
Ontology-based Information Retrieval	i)S.Saraswathi.et.al [13]	i) Classification of entities in a hierarchical order. ii) Based on keyword matching.	i)Reusability and Sharing of the Ontology with other applications.	i)it is a time-consuming ii)Difficult to create ontological tree . ii)It is laborious to manually add a new concept into an existing ontology.

5. CONCLUSION

This paper discussed about the different Information Retrieval techniques. Each of the methods has different criteria in extracting document for user’s query. Malayalam preprocessing requires large works compared to other languages so we have to need a better and easy IR model. At last we conclude that, for Malayalam information retrieval vector space model using tf –idf weighting scheme may produce better result. Vector space model find the similarity measure and rank them according to their similarity, That is why vector space using Tf-Idf model gives better result for long documents as compared to term-count model. Vector space model was used to rank text documents in response to user query. In future this approach can be combined with POS tagging to improve web information retrieval.

6. REFERENCES

[1] JinxiXu, Ralph Weischedel and Chanh Nguyen"Evaluating a Probabilistic Model for Cross-lingual Information Retrieval September" 9-12, 2001,ACM .

[2] Pinaki Bhaskar, Amitava Das, Partha Pakray and Sivaji Bandyopadhyay, **"**Theme Based English and Bengali Ad-hoc Monolingual Information Retrieval in FIRE 2010"

[3] Ricardo Baeza-Yates and BerthierRibeiro-Neto, "Modern Information Retrieval,"ACM Press, New York, 2009.

[4] B. Ribeiro-Neto, R. Muntz, "A belief network model for IR" in: Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM Press, 1996, pp. 253–260.

[5] Jeongwoo ko and Eric Nyberg and Teruko Mitmura, "Probabilistic Models for Answer-Ranking in Multilingual Question-AnsweringACM" Transactions on Information Systems, Vol. 28, No. 3, Article 16, Publication date: June 2010.

[6] Nikesh P.L., Sumam Mary Idicula, and David Peter S, "English Malayalam Cross-Lingual Information Retrieval - an experience,"IEEEISSN: 978-1-4244-2030-8, Vol. 3, No.2, 2008

[7] Nath Singh, Sanjay Kumar Dwivedi Jitendra "Analysis of vector space model in Information Retrieval"National Conference on Information Technology & its impact on Next Generation Computing CTNGC 2012 Proceedings published by International Journal Of Computer Application(IJCA).

[8] Reshma O.K., Sreejith C, P.C. Reghu Raj "An Effective Malayalam Information Retrieval System Using Query Expansion",2013 International Conference on Control Communication and Computing (ICCC) 13-15 December 2013.

[9] Marco Antonio Pinheiro de Cristo ,Pavel Pereira Calado, Maria de Lourdes da Silveira,Ilmerio Silva, Richard Muntz,Berthier Ribeiro-Neto **"**Bayesian belief networks for IR "2003 Elsevier.

[10] Neelam Phadnis ,Jayant Gadge "Framework for Document Retrieval using Latent Semantic Indexing"International Journal of Computer Applications (0975 – 8887) Volume 94 – No.14, May 2014.

[11] A.P.SivaKumar,Dr.P.Premchand,Dr.A.Govardhan "Indian Languages IR using Latent Semantic Indexing" International Journal of Computer Science & Information Technology (IJCSIT) Vol 3, No 4, August 2011.

[12] Igor Mokris, Lenka Skovajsova "Neural Network Model Of System For Information Retrieval From Text Documents In Slovak Language" ActaElectrotechnica et Informatica No. 3, Vol. 5, 2005

[13] 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.

[14] Viatcheslav Yatsko "TF*IDF Revisited" International Journal of Computational Linguistics and Natural Language Processing Vol 2 Issue 6 June 2013.