Volume 8, No. 5, May-June 2017
ISSN No. 0976-5697
International Journal of Advanced Research in Computer Science
RESEARCH PAPER
Available Online at www.ijarcs.info

A Trend Analysis of Information Retrieval Models

Dr. M. Balamurugan
Associate Professor,
School of Computer Science, Engineering and Applications,
Bharathidasan University,
Tiruchirapalli, India

E.Iyswarya
Research scholar,
School of Computer Science, Engineering and Applications,
Bharathidasan University,
Tiruchirapalli, India

Abstract: Information retrieval technologies behind web search engines in the field of computer science were bought up in the year of 1950s. It is a process of retrieving the relevant documents based on the queries raised by the user. It deals with the representation, storage and access of information items. In this system, the generated outputs are ranked according to their relevance. The information retrieval (IR) uses data models that make a retrieval process easier when compared to the traditional IR database model. In this work, we analyse the most popular information retrieval models such as boolean, vector space, probabilistic and latent semantic analysis and evaluate the performance of the models by using the underlying parameters like concept, representation, word occurrence, information type, output, pros and cons of the models. This study aims to determine the appropriate model for different situations and additionally describes the indexing methods for decrementing search space and different probing (searching) techniques to retrieve the information.

Keywords*:* information retrieval, boolean, vector space, probabilistic model, latent semantic analysis

I INTRODUCTION

Information retrieval is generally designed to help the users to quickly get the relevant information on the web. It is considered as a subfield of computer science that deals with the representation, storage and access of information [8]. The web search engine is a kind of website which is used to retrieve the information from the World Wide Web. The search engine consists of multimedia information and information retrieval system is the way to retrieve this information. In this study, the notion on information retrieval system is been explored, as information retrieval (IR) system is a software program that stores, manages information from the documents and assists users in finding the information they needed. It doesn't explicitly return the information rather the system informs the existence and location of documents that might contain the desired information. This suggested information will satisfy the user's information needed. This desired document is called as relevant document. A perfect retrieval system will retrieve only the relevant document but a perfect retrieval system will not exists because retrieving the relevant information will based on different users and their relevance also based on their subjective opinion[5].

The main goal of information retrieval is to "find the relevant information or document that satisfies the user's information need". To satisfy the user's query the IR system implement the processes as follows: Indexing process (represent the documents in summarized content form), filtering process (filters the stop words) and searching process (there are various techniques to retrieve matching information of user's need) Likewise, the quality of information retrieval is measured by:

Precision (positive predictive value) is the number of relevant documents retrieved is divided by the total number of documents retrieved.

$$\text{Precision} = \frac{[\{\text{relevantdocuments}\ \}\cap\{\text{retrieveddocuments}\ \}]}{[\{\text{retrieved}\ \ \text{documents}\ \}]}$$

Recall (sensitivity) is the number of relevant documents retrieved is divided by the total number of documents [2].

$$\text{Recall} = \frac{[\{\text{relevantdocuments}\ \}\cap\{\text{retrieveddocuments}\ \}]}{[\{\text{relevant}\ \ \text{documents}\ \}]}$$

The two key problems exist in the process of IR systems: first, it fetches some irrelevant information together with the relevant one. Second, search engines are not capable to perform the retrieval of all relevant documents [7].

In this paper, different retrieval models are explained to help the users to retrieve the information as per the user's opinion. The structure of this paper is as follows. Section I gives the perception on information retrieval along with their quality measures. Section II describes the information retrieval models. Section III shows the comparative analysis of these models and section IV explains the indexing techniques. Section v discusses the searching techniques. In section VI the related works were discussed and ended with the conclusion in last section.

II IR MODELS

To make information retrieval efficient, the documents are transformed into a suitable representation. Now such type of information is retrieved efficiently with the help of IR models. The models are categorized according the properties of model as: set-theory model and statistical model. The set-theory model represents documents as sets of words or phrase. Similarities are usually derived from set theoretic operations on those sets [15]. Boolean model is said as set theory model. Likewise the vector space and probabilistic model is said as statistical model which use the statistical information in the form of term frequencies. And all these three models are said as traditional models. There are three basic processes an information retrieval system supports: the representation of the content of the documents (indexing process), the representation of the user's information need (query formulation process) and the comparison of two representations (matching process). [1]

Figure 1, shows the process of information retrieval and the squared boxes represent the data and circled boxes represent the process. In the case of IR, a retrieval model

specifies the representations used for documents and information needs, and how they are compared. The descriptions of these models are given below.

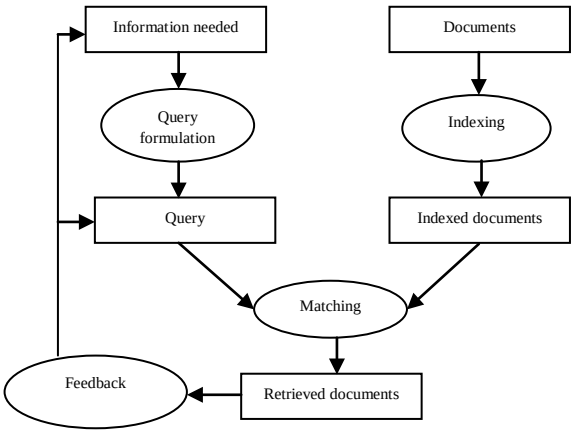

Figure1. Information retrieval process

A. Boolean model

The Boolean model of information retrieval is a classical first model and said as an exact-matching model (query specifies precise retrieval criteria every document either matches or fails to match query). It is based on set theory (studying on collection of sets) and Boolean algebra (true or false). In this the documents to be searched and user's queries are represented in the form of set of terms, each is viewed as Boolean variable and valued as true it is present in the document. In Boolean model, documents are associated with a set of keywords and for query formulation operators like AND, OR, NOT are used [6]. The search engine returns all documents that satisfy the Boolean expression. And it cannot rank documents in decreasing order of relevance. The retrieval function of this model treats a document as either relevant or irrelevant.

B. Vector space model

It is a simple model based on linear algebra. The vector space model represents documents and queries as vectors in multidimensional space, whose dimensions are the terms used to build an index to represent the documents. The procedure of this model is divided into three stages: Document indexing (content bearing terms are extracted from the document text), weighting the indexed terms (enhancing the retrieval of documents is relevant to user) and ranking the documents (by similarity measure).A common similarity measure is known as cosine measure determines angle between the document vector and the query vector [7]. When the user requests for some information, the output are generated based on the similarity between the query vector and the document vector. The user query is treated as vector [17].

$$\text{sim}(d_j, q) = \frac{dj.q}{\|dj\|\|q\|} = \frac{\sum_{i=1}^{N} w_{i,j} w_{i,q}}{\sqrt{\sum_{i=1}^{N} w_{i,j}^2} \sqrt{\sum_{i=1}^{N} w_{i,q}^2}}$$

Documents and vectors are represented in vectors

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

$$q = (w_{1,q}, w_{2,q}, \ldots, w_{t,q})$$

Vector space model introduces the term weight scheme known as if-idf weighting. These weights have a term frequency(tf) factor measuring the frequency of occurrence of the terms in the document or query texts and an inverse document frequency(idf) factor measuring the inverse of the number of documents that contain a query or document term[9].

C. Probabilistic model

This model is introduced in 1976 by Robertson and Sparck Jones, which later became known as the binary independent retrieval (BIR) model. The probabilistic retrieval model is based on the Probability Ranking Principle, which states that an information retrieval system is supposed to rank the documents based on their probability of relevance to the query. Documents and queries are represented by binary vectors ~d and ~q, each vector element indicating whether a document attribute or term occurs in the document or query or not. Instead of probabilities, the probabilistic model uses odds O(R), where O(R) = P(R)/¬P(R), R means "document is relevant" and ¯R means "document is not relevant". [9]

D. Latent Semantic Indexing model

Latent Semantic Indexing (LSI) is an extension of vector space model that indexes and uses mathematical technique called singular value decomposition (SVD), which identifies the pattern in an unstructured collection of text and finds relationship between them. A method to improve the quality of similarity search in text is called LSI in which the data is transformed into a new concept space. This depends upon the document collection in question, since different collections would have different sets of concepts. LSI is a technique which tries to capture this hidden structure using techniques from linear algebra [16]. The contents of a webpage are crawled by a search engine and the most common words and phrases are collated and identified as the keywords for the page. This model is used to overcome the problem of finding the relevant documents from the search words by mapping both words and documents into a "concept" space and doing the comparison in this space. It uses a mathematical technique called singular value decomposition (SVD) to identify patterns in the relationships between the terms and concepts contained in an unstructured collection of text [17].

To perform the latent semantic indexing, the following steps are performed: first, convert each document in index into a vector of word occurrences. Second, scale each vector so that every term reflects the frequency of its occurrence. Next, combine these column vectors into a large term-document matrix. Rows represent terms, columns represent documents. Perform SVD on term-document matrix. According to (Deerwester et al) three major advantages of using the LSI representation are: synonymy (same underlying concept can be described using different terms), polysemy (represents words that have more than one meaning. Large numbers of polysemous words in the query can reduce the precision of a search significantly) and term dependence (represents first-order approximation) [10]. The disadvantages are: storage, lack of efficiency [11] and output are in slow rate while using collaboration with large documents [14].

III. COMPARATIVE ANALYSIS OF IR MODELS

The Table 1 shows the comparative analysis of information retrieval models.

Table I. Comparison of Information Model

IR models(IR mod)/ attributes(A)	Boolean (IR mod)	Vector space (IR mod)	Probabilistic(IR mod)	Latent semantic indexing(IR mod)
Concept(A)	Based on set theory and Boolean algebra	Based on the concept of vectors	Based on probability ranking principle	It is an extension of vector space model
Representation(A)	Documents are represented by the index terms extracted from documents, and queries are Boolean expressions on terms.	Represented in the form of weighted-term vectors. Cosine measure is used to find the similarities	Documents and queries are represented in binary vectors	Documents are represented in the form of term-document matrix.
Information type(A)	Does not consider semantic information	It consider semantic information	Considers the semantic information	Considers the semantic information
Word occurrence(A)	Number of occurrence are not mentioned	Tells about the number of occurrence	Occurrence based on the probability relevance	Based on term-document matrix
Output(A)	Exact match of the output to the query	Best match of the query	It gives best match of output	Best match of the query
Advantages(A)	Easy to implement	Simple model, weights are not in binary	Theoretical adequacy: ranks by probabilities	Synonymy and polysemy
Disadvantages(A)	Does not rank documents, retrieves too many or too few	Suffers from synonymy and polysemy. It theoretically assumes that terms are statistically independent	Binary weights, ignore frequencies and independence assumption	Not clear about similarity between words

IV. INDEXED TECHNIQUES

There are common indexing techniques in information retrieval process like signature files and inverted indices.

A. Signature file

In signature file indexing technique each document return a bit of string, (that is, signature) using hashing method on its text and superimposed coding. The final output of document signatures is stored in a separate file and this file is called as signature file. The signature file is much smaller than the original file, and it can provide high search rate.

B. Inverted index

Each document can be represented by a list of some reference words called keywords which explains the contents of the document for retrieval purpose. Fast retrieval can be obtained if we invert on those keywords. All the reference words are stored alphabetically in a file called index file. For each keyword there is a list of pointers to the characterize documents in the postings file. This method is mostly used by all the commercial systems.

V. SEARCHING TECHNIQUES

There are different searching techniques, including linear search, brute force search and binary search and these searching techniques are describe as follows:

A. Linear search technique

This search technique is a simplest search algorithm. It is a basic technique of finding a particular word or keyword from a list of words or array that checks presence of every element in list, one at a time and in a sequence. One major disadvantage of linear search is its searching speed is very poor or slow especially in case of ordered list. This type of search is also called as sequential search.

B. Brute force search technique

It is a very common problem-solving technique that consists of consistently itemizes all possible participants for the solution and determines whether each participant satisfies the problem's statement. This searching technique is simple to apply and it will always return a solution if it exist.

C. Binary search technique

It finds the position of a particular input value (that is, the search key) within an array sorted by some key value. For this technique, the given array should be arranged in some order that is, ascending or descending. In each step, this algorithm examines the search key value of middle element key value of the given sorted array. If the value of both keys matched, then a matching item has been found and its index or position is returned. Differently, if the search key value is less/greater than the middle element's key value, then the method repeats its steps on the sub-array to the left/right of the middle element. If the leftover array to be searched and it is found empty, then the search key cannot be found in this empty array and a particular bit of string is indicated as "Not Found" is returned. [12]

VI. RELATED WORKS

R. John and b. Killoran states that the information retrieval process represents the information in the form of query. The query is stated as formal statements. Several researches have been done on the information retrieval models [3]. In earlier the researches shown that Boolean model is the simplest of the entire model for retrieving the information which represents in the form of Boolean operators it is easy to implement.

Suresh Kumar, Manjeet Singh and Asok De classify the information retrieval models into exact-match retrieval and best – match retrieval. In Exact-match retrieval model, exact keyword matching is carried out. This is suffering from the problem of synonymy and polysemy. Best-Match retrieval model is designed to overcome these problems [13].

Vaibhav Kant Singh, Vinay Kumar Singh explained that vector space model is easy to understand, cheaper to implement considering to the fact that the system should be cost effective i.e. should follow the Space/Time constraint and shows the best result to the query [4].

Ashwini Deshmukh and Gayatri Hedge discussed the latent semantic indexing which uses the indexing technique reveals that it is very effective method of retrieving information from even a large amount of documents [16][15].

VII. CONCLUSION

On the focus towards to retrieve desired result for information the IR models are used efficiently. On the other hand, all retrieval models are based on different concepts and assumptions. The evaluation of the models based on parameters encloses that in case of implementing the simple queries; boolean model is the simplest model. The vector space also used to generate the queries for small documents only. The latent semantic analysis represents the documents that are somewhat economical and it tries to overcome problem of lexical matching by conceptual matching and it is a very effective method of retrieving information from even for large documents. There is a variation of latent semantic indexing model is called as latent dirichlet allocation model which is a probabilistic model for finding latent semantic topics in large collection of documents.

VIII. REFERENCES

[1] Akram Roshdi and Akram Roohparvar, "Review: Information Retrieval Techniques and Applications", International Journal of Computer Networks and Communications Security, vol. 3, no. 9, September 2015, 373–377.

[2] Yi Shang Longzhuang Li: Precision Evaluation of Search Engines. World Wide Web (2002).

[3] R. John b. Killoran, "How to Use Search Engine Optimization Techniques to Increase Website Visibility**,"** vol. 56, NO. 1, March 2013.

[4] Vaibhav Kant Singh, Vinay Kumar Singh, "vector space model: an information retrieval System", International Journal of Advanced Engineering Research and Studies, Proceedings of BITCON-2015 Innovations for National Development National Conference on: Information Technology Empowering Digital India.

[5] Vijaykumar, S., S.G. Saravanakumar, Dr. M. Balamurugan: "Unique Sense: Smart Computing Prototype", Procedia Computer Science, Issue. 50, pp. 223 – 228, 2015. doi: 10.1016/j.procs.2015.04.056

[6] Manish Sharma1, Rahul Patel, "A Survey on Information Retrieval Models, Techniques And Applications", International Journal of Emerging Technology and Advanced Engineering (ijetae)Volume 3, Issue 11, November 2013.

[7] Jitendra Nath Singh, Sanjay Kumar Dwivedi, "Analysis of Vector Space Model in Information Retrieval", CTNGC 2012 Proceedings published by International Journal of Computer Applications (IJCA).

[8] S.Vijaykumar, S. G. Saravanakumar, "Future Robotic Memory management", Advances in Digital Image Processing and Information Technology Communications in Computer and Information Science .Volume 205, 2011, pp 315-325. ISSN: 1865-0929. DOI:10.1007/978-3-642-24055-3_32

[9] D.Hiemstra, P. de Vries, "Relating the new language models of information retrieval to the traditional retrieval models", published as CTIT technical report TR-CTIT-00-09, May 2000

[10] Deerwester, Dumais, Furnas, Lanouauer, and Harshman, "Indexing by latent semantic analysis", Journal of the American Society for Information Science, 41 (1990), pp. 391-407.

[11] Vijaykumar S, Saravanakumar SG, Balamurugan M. "Unique sense: Smart computing prototype for industry 4.0 revolution with IOT and bigdata implementation model", Indian Journal of Science and Technology, 2015 Dec; 8(35). DOI: 10.17485/ijst/2015/v8i35/86698.

[12] Balwinder Saini, Vikram Singh, Satish Kumar, "Information Retrieval Models and Searching Methodologies: Survey", International Journal of Advance Foundation and Research in Science & Engineering (IJAFRSE), Volume 1, Issue 2, July 2014.

[13] Suresh Kumar, Manjeet Singh, Asok De, "Information Retrieval Modeling Techniques for Web Documents", Proceedings of the 5th National Conference; INDIACom-2011 Computing For Nation Development, ISSN 0973-7529, March 2011.

[14] Barbara Rosario, "Latent Semantic Indexing: An overview", INFOSYS 240 Spring 2000, Final Paper.

[15] Balamurugan M, Vijaykumar S, Saravanakumar SG. Analysis of High Performance Parallel Computing Instruction Sets. Indian Journal of Science and Technology. 2016 Dec; 9(48). DOI: 10.17485/ijst/2016/v9i48/97098.

[16] Ashwini Deshmukh, Gayatri Hegde, "A Literature Survey on Latent Semantic Indexing", International Journal of Engineering Inventions, Volume 1, Issue 4 (September 2012) PP: 01-05, ISSN: 2278-7461.

[17] Bhavna Arora, Abhinav Bhardwaj, "Analysis of Information Retrieval models", International Journal Of Engineering And Computer Science, Volume 3 Issue 10, October 2014 Page No. 8551-8554, ISSN:2319-7242 .

