Review Article **Bioinformatics**

International Journal of Pharma and Bio Sciences

ISSN 0975-6299

STUDY OF APPLICATIONS OF CONTENT BASED INFORMATION RETRIEVAL MODELS IN HEALTHCARE

SENTHIL MURUGAN BALAKRISHNAN [*1]

[1*]*School of Information Technology and Engineering, VIT University, India.*

ABSTRACT

Information Retrieval plays an important role in a distributed collection of documents. In this paper, we discuss the various information retrieval models for an efficient and effective matching. We analyze each and every model for its benefits and issues in its usage in a distributed environment. The models are compared on the basis of the characteristics such as input query, implementation, ranking and precision rate. As the Healthcare industry turns out to be progressively reliant on electronic data the requirement for modern data retrieval frameworks and for educated individuals to plan, buy, and utilize them additionally increments. This paper also analyzes the cases of these information retrieval schemes in the Healthcare sector.

KEYWORDS: **Boolean Retrieval, Probabilistic Model, Indexing, Ranking**.

SENTHIL MURUGAN BALAKRISHNAN*
School of Information Technology and Engineering, VIT University, India.

Received on : 28-09-2016
Revised and Accepted on : 21-11-2016
DOI: http://dx.doi.org/10.22376/ijpbs.2017.8.1.p125-128

INTRODUCTION

In recent years, the need for information storage and retrieval has increased rapidly, the main objective of the system is to store the contents, locate the contents whenever require by the users and maintain the contents for future use. Many information retrieval methods are developed to identify the document in a huge collection that exactly matches with the user's query[1]. An information retrieval model is said to be effective if it specifies the details of document information, query representation and the matching function. In earlier days, traditional information retrieval models are used for high structured documents and the complex current generation models are used to handle document with complex internal structure[2]. Retrieval of accurate contents from the huge collections requires efficient searching methods with high scalability. Distributed platform express a way to exhibit efficient and fast retrieval over a large scattered data. When a searching is performed for large collections they are analyzed and then queried, which will enable fast processing and efficient data management. Searching methods efficiency is evaluated based on the response time and the main concept involved in searching method is indexing. Indexing[3] use the divide and conquer method, where the large collections are divided into small sets, each and every sets are processed using parallel algorithms to enable high computing speed and the final result are displayed. From the healthcare perspective information retrieval deserves special attention. The Electronic Health Records (EHR) serves as the major source of information. The information retrieval relevant to the medical record as the base for research is called as "passage retrieval"[4].

EXISTING APPROACHES OF INFORMATION RETRIEVAL

There are many retrieval models for textual documents. They are broadly classified into exact-match, vector space and probabilistic model[5]. Information retrieval means identifying the documents in the given collection that exactly match with the user input query. Information retrieval starts with Boolean Systems, but with the increase in data many new model were developed for efficient content retrieval.

Boolean Retrieval Method

Boolean Retrieval method[6] is the first retrieval method which follows the Boolean rules, called as traditional Boolean Retrieval System. In the proposed retrieval system includes a binary valued variable called as Feature variable[7] and Boolean operators (AND, OR and NOT). The feature variable contents the text extracted from the documents. When the search keyword is given, based on the search function the results are displayed in two sets – relevant documents and irrelevant documents. Each and every document is assigned with true value if the document contains the keyword. Firstly, the relevant documents with true value will be displayed and then the irrelevant documents. No ranking algorithms so they are ranked based on the document properties like order by date, order by size of file. Some of the advantage is- it is easy to implement, standard model for large scale retrieval system, it is very effective. The disadvantages are - it is difficult to write a Boolean expression for queries, this method does not differentiate natural language terms AND, OR, NOT with the Boolean operators and no ranking algorithms are used[8]. To overcome the disadvantages of traditional Boolean system, new methods are developed namely Smart Boolean and Extended Boolean Model. Smart Boolean methods[9] main goals are structure search, more users friendly and effective. In this method we don't use much Boolean operators and conceptual query representation is absent. Extended Boolean model[10] approaches are P-norm and Fuzzy set theory that are used to assign weight to the documents and rank them. In the attempt for improving search over EHR, Davit et al[16] suggests that, from the data analysis point of view, many research works extracted fields from the query and generated different indices depending on the use of separate fields or not (FIELDS or COMBINED). In this the use of FIELDS is implemented as a Boolean search over the fields in which OR operator is used to join the results on each field index and ranking of documents.

Vector Space Method

Vector Space Model[12] is the model which assigns numeric value to document and rank them. In this model the document content and search keywords are represented as vector of terms in a multi dimension. They are independent dimension in a dimensional vector space. If term belongs to a particular document then value will be a non-zero and if does not belong then it will be zero, along the dimension space corresponding to the term. Most system is operated in a positive quadrant. There are many methods[7] used to compute the vector value. Most common one is using similarity function between the document vector and the query vector, that is angle between the vector are the divergence measure and cosine of the angle is the numeric value (cosine method) or the dot product between the vectors are also used. Once the numeric value is calculated, then the document can be ranked according to the relevance .This model uses the linear algebra and partial matching of content. Some limitation of this model is that order of representation in document differ with vector representation, terms are statistically independent. Vector space model is extend to generalized vector space model, Topic based vector space model, latent semantic indexing. These methods use the same concept in different ways to obtain an efficient result. The earlier work[13] by Zhang et al (2013) shows that vector space model and its semantic enhancement are complementary in its application over medical disorders information retrieval. Further they develop a baseline information retrieval method using vector space model and semantic vector space model with dissimilar random indexing and evaluates the influence of query expansion using query datasets containing shared task on information retrieval of Medical Disorders in 2013.

Probabilistic Model

Probabilistic Model[10] is also a model which assigns some value to the document and ranks them. This

model uses the probabilistic ranking principle (PRP) like 0/1 loss, which states that the document should be ranked based on the probability of relevance to the query. There are many probabilistic methods have be proposed, the most traditional one which is used is binary implementation method. In the probabilistic method[7] the document are represented as D by P(R|D) where R is the relevant document and P(~R|D) is represented for non-relevant document using the Bayes transform. When the query is Q (split into terms), the probability of term present in relevant document for all terms in query and document and the probability of term absent in relevant document for all terms in query and not in document is obtained. Once the probability is calculated we can rank the document by adding some constant to it, then the documents are displayed in descending order of relevance. In the context of healthcare, the concept of probability is used for understanding test results and for making medical decisions that involve uncertainty which is inherent. An understanding of what probability is and of how to adjust probabilities after the acquisition of new information is paramount in clinical consultation systems.

Indexing Model

Indexing plays an important role in the distributed environment. For fast processing and efficient search, divide and conquer method can be introduced. In this method the large collection of document are divided into small nodes and finally all the contents are merged to get the final output. There are many frameworks which implements this method with high scalability, simple programming model and economical. The most commonly used framework in the MapReduce, which takes key/value as input and produces output key/value pairs. Mostly used indexing method is inverted index method which is the fundamental for all the information models. For the given query, each and every term will have a posting list. This list contains the occurrence of the term along with document id. So, the document with more posting list will be ranked highest. But there are many difficulties when we use inverted index in a distributed environment. So by using Hadoop MapReduce various indexing methods like Per-term and Per-token methods are implemented. In Per-Token[14] indexing the map function will generate the <token, document id> for each token in the document and the reduce function will count the number of term frequency for each document and add the contents to the Posting List. But in this method if a token emits “n” times then $< term, document\ id >$ will be repeated “n” times, so the intermediate data between the map and reduce will be high which will affect the overall job execution time. In Per-Term[14] indexing the map function will generate <term, (document id, term frequency)> for the tokens in the documents and the reduce function will pass only the terms with term frequency is not equal to zero. So it will generate one per document and overall intermediates will be reduced, but this method use the inverted index method which is less efficient in distributed environment. The major area where indexing based retrieval is used in medical field is indexing and retrieval of medical images[15]. As indexing of medical images using text or numbers is difficult and it requires more time to memorize, the novel methodologies adopting retrieval through query by text and query by image called Content-Based Image Retrieval (CBIR) deserves special attention.

Table 1
Comparison of Models

Factors	Boolean Retrieval Method	Statistical Models	Indexing Models
Input Query	Structural and conceptual representation	Query formulation representation	Index representation
Implementation	Easy to Implement	Best Match method, complicated	Efficient and fast retrieval
Ranking Method	No Ranking method	Based on ranking function	Based on term weight
Precision Rate	Relevant and irrelevant documents	Relevant documents based on query	Exact matches document

DISCUSSIONS

In this survey I have summarized the observations on the various information retrieval models with respect to the factors such as input query, implementation complexity, ranking method adopted and precision rate in Table I.

CONCLUSION

In this paper, we have discussed many information retrieval methods and analyzed. In a distributed environment we can implement these models but each and every model have its own issue .With the analysis result indexing is the best retrieval method which can be implemented for an effective and good search result based on the query when compared to other models. This survey could be helpful in addressing medical image indexing and retrieval problem by proposing appropriate experimental design and compare their efficiency and performance. As a part of future work, the healthcare information retrieval is implemented using data analytics frameworks such as Hadoop, Spark, R and so on. The results obtained can be is analysed from the perspective of factors stated in Table 1.

CONFLICT OF INTEREST

Conflict of interest declared none.

REFERENCES

1. Salton G, McGill MJ. Introduction to modern information retrieval
2. Belkin NJ, Croft WB. Information filtering and information retrieval: Two sides of the same coin?. Communications of the ACM. 1992 Dec 1;35(12):29-38.
3. McCreadie R, Macdonald C, Ounis I. Comparing distributed indexing: To mapreduce or not?. Proc. LSDS-IR. 2009:41-8.
4. Melucci M. Passage retrieval: A probabilistic technique. Information Processing & Management. 1998 Jan 31;34(1):43-68.
5. Dong H, Hussain FK, Chang E. A survey in traditional information retrieval models. InIEEE International Conference on Digital Ecosystems and Technologies 2008 Feb 26 (pp. 397-402).
6. Lashkari AH, Mahdavi F, Ghomi V. A boolean model in information retrieval for search engines. InInformation Management and Engineering, 2009. ICIME'09. International Conference on 2009 Apr 3 (pp. 385-389). IEEE.
7. Turtle HR, Croft WB. A comparison of text retrieval models. The computer journal. 1992 Jun 1;35(3):279-90.
8. Khankasikam K. A comparison of information retrieval models applied to Thai digital library. InComputer and Automation Engineering (ICCAE), 2010 The 2nd International Conference on 2010 Feb 26 (Vol. 1, pp. 335-338). IEEE.
9. Salton G, Fox EA, Wu H. Extended Boolean information retrieval. Communications of the ACM. 1983 Nov 1;26(11):1022-36.
10. Croft WB, Harper DJ. Using probabilistic models of document retrieval without relevance information. Journal of documentation. 1979 Apr 1;35(4):285-95.
11. Lee JH, Kin WY, Kim MH, Lee YJ. On the evaluation of Boolean operators in the extended Boolean retrieval framework. InProceedings of the 16th annual international ACM SIGIR conference on Research and development in information retrieval 1993 Jul 1 (pp. 291-297). ACM.
12. Wong SM, Raghavan VV. Vector space model of information retrieval: a reevaluation. InProceedings of the 7th annual international ACM SIGIR conference on Research and development in information retrieval 1984 Jul 2 (pp. 167-185). British Computer Society.
13. Zhang Y, Cohen T, Jiang M, Tang B, Xu H. Evaluation of Vector Space Models for Medical Disorders Information Retrieval. InCLEF (Working Notes) 2013.
14. McCreadie R, Macdonald C, Ounis I. MapReduce indexing strategies: Studying scalability and efficiency. Information Processing & Management. 2012 Sep 30;48(5):873-88.
15. Chandrakar A, Thoke AS, Singh BK. Indexing and Retrieval of Medical Images Using CBIR Approach. InAdvances in Parallel Distributed Computing 2011 (pp. 393-403). Springer Berlin Heidelberg.
16. Martinez D, Otegi A, Soroa A, Agirre E. Improving search over Electronic Health Records using UMLS-based query expansion through random walks. Journal of biomedical informatics. 2014 Oct 31;51:100-6