www.ijecs.in
International Journal Of Engineering And Computer Science ISSN:2319-7242
Volume 3 Issue 10, October 2014 Page No. 8551-8554

Analysis of Information Retrieval models

Bhavna Arora[1], Abhinav Bhardwaj[2]

[1]Department of IT, Amity University
Uttar Pradesh
arora.bhavna244@gmail.com

[2] Department of ECE
ITM University
abhibhard99530@gmail.com

Abstract: **The information on the web search engine is escalating alarmingly and the efficiency of the information generated depends on the way of retrieving the information. There are several models which help to retrieve the pertinent information.**
In Information Retrieval system, the generated outputs are ranked according to their relevance. Thus information retrieval process commences as soon as the user inputs a query to hunt sufficient information. The aim of information retrieval system is to retrieve the documents attaining the user query. This paper expounds the architecture of the search engine and several different information retrieval models such as Boolean model, Vector space model, Latent semantic Indexing (LSI) model and Latent Dirichlet allocation (LDA) model. It also covers the comparative analysis of these models.

Keywords: search engine, Information retrieval, Boolean Model, Vector Space model, Latent Semantic Indexing (LSI).

1. Introduction

The web search engine is a kind of website which is used to retrieve the information from the World Wide Web. There is a plenty of information including audio, pdfs, video and other multimedia information obtainable from the internet. The search engine is the engine of multimedia information and the information retrieval system is the way of retrieving this multimedia information. The user has to find the relevant information according to his need. The information can be searched either via using search engines or by using directories [1]. If web search engine is used, it makes the use of the queries. The input is taken as a query and relevant output is generated as a linear list of documents. Despite of being simple and efficient, it takes too much time to generate the output. The web directories, if used, makes the use of catalogs being dependant on several factors like quality of classification signifying the way through which techniques is applies, the accuracy of generated output, etc.
In this paper, different techniques are explained which helps to generate the documents as per user needs. The outline of paper is as follows. First section explains the literature survey of the search engine and different information retrieval techniques. Section 2 describes the architecture of the search engine and its various components including crawling, indexing, and ranking of the documents. Section 3 defines the. information retrieval system and its different techniques. In another section, brief comparative analysis of these information retrieval models is done which is then ended with the conclusion in the last section.

The information retrieval is the process of representing the information and the way of storing the information so that efficient information could be retrieved. Moreover, the information is stored in free form and not structured. The web search engines consist of several multimedia information including audio, video and other collection of text. Thus, the information retrieval process is a way of representing the information in the form of query. The queries are represented as formal statements. Based on the query given as an input, efficient information is generated. Several researches have been done on the information retrieval models [2]. Earlier researches shown that Boolean model is the simplest model for retrieving the information. It represents the information in the form of Boolean operators and queries and shown the exact match to the query. It is easy to implement.[1]. Arash Habibi Lashkari and Fereshteh Mahdevi explained that the vector space model represents the documents in the form of vectors. Researches shows that the vector space model is more efficient and shows the best result to the query. The latent semantic model introduces a variance called Latent Dirichlet Allocation model which considers both the synonym and polysemy words.

2. Web Search Engine Architecture

The web search engine, in today scenario, is the primary tool for searching the information; the large

search engine is increasing with the large amount of the information. This information is managed based on some algorithm. The efficiency of the documents generated depends greatly on the input given [11][12]. The search engine generates a long list ranked by the degree of relevance. The architecture of the search engine is shown in figure1

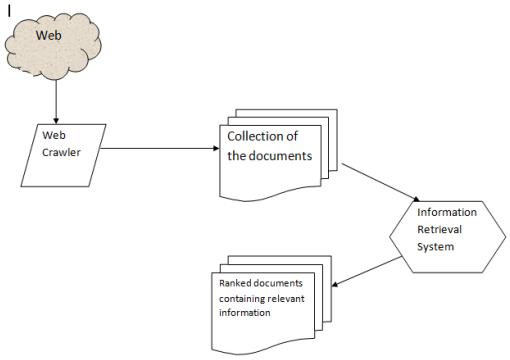

Figure 1: Search Engine Architecture

The search engine generates the output by performing the crawling of the information followed by indexing and ranking of the generated output.

- *web crawler*

Web crawler is the component with which the user can search for any word in the web page. They are used for selecting web pages for generating large corpus of text. It creates a cache copy of the processing pages which indexes the downloaded pages for searching efficiently. It follows a crawling policy which is a combination of several policies. These policies may be the selection policy, revisit policy, parallelization policy etc. The web crawlers are used to collect information from the web. There is a list of URLs maintained in the queue. Crawlers consider all sort of web pages like HTML page, XML page, PDF and other document types. For Google, the crawler used is Googlebot.

- *Indexing*

Indexing is used to store and represent the information. The aim of indexing is the optimization of the performance in retrieving relevant information. It is used to retrieve the most relevant document followed by next degree of relevance.

- *The retrieval of information*

The information is retrieved based on some information retrieval models. These models may include the Boolean model, the Vector Space Model, Latent Semantic Indexing model or Latent Dirichlet Allocation models. The models generate the different outputs based on different inputs given.

- *Ranking of the documents*

The list of documents generated is ranked according to the degree of relevance..

3. Information Retrieval systems

The information retrieval system generates the output of documents based on the query entered by the user. The documents that are generated as an output are ranked according to their relevance. The user request is treated as a query and the output satisfying the query is the list of documents containing relevant information [8]. The main aim of such systems is to generate the output that satisfies the used request and for generating the outputs, different techniques or models are used. These models may include the Boolean model, Vector Space Model, Latent Semantic Indexing (LSI) model and Latent Dirichlet Allocation (LDA) model.

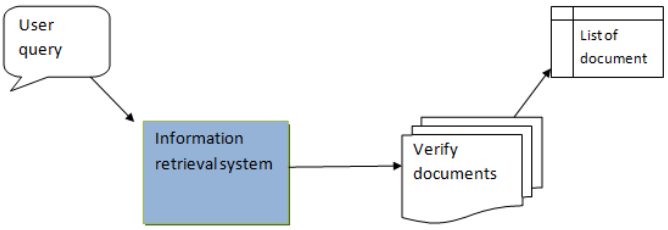

Figure 2: Information Retrieval Process

3.1 Boolean model

The Boolean model is based on the set theory and the user request is treated as a query. This query is represented as the Boolean expressions of the keywords. It makes the use of Boolean operators like AND, OR, NOT and also the parenthesis to define and signifies their scopes. The output generated from this model is the list of relevant documents where no ranking to the document is being given [1].

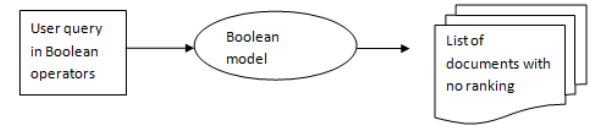

Figure 3: Boolean model

Boolean model gained the popularity because of its simple structure. Some of the other advantages of the Boolean model are: Simple and easy to implement, efficient. Users find it easy to search for the phrases and synonyms that are useful in the generation of the query [5].

Despite of being simple, it suffers from some drawbacks such as difficulty in ranking or prioritizing of the output and not suitable for complex queries as writing query in the form of Boolean operations is quite difficult and challenging task.

3.2 Vector Space model

The vector space model was used because of inclusion of weighted term vectors for each document. It is basically an algebraic model which represents the documents containing the text as the vector of identifiers. Thus, it represents the documents and the queries as the vectors in multi dimensional space.

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 [6][7].

This model has so many advantages like simple model based on the linear algebra. It ranks the documents based on their relevance of the position. But it suffers from some limitations also. Since long documents are having poor similarity value, they are poorly documented. Moreover, the order of the appearance of the term is lost.

3.3 Latent Semantic Indexing model

The Latent Semantic Model (LSI) is an extension of vector space model which is based on the technique of the Singular Vector Decomposition (SVD) to identify the patterns and relationships between the terms and the concepts [9].

In this model, the keywords are being replaced by the 'concepts' which considers the synonym of the keywords as the relevant document.

The advantages of this model are: It is not restricted to work with words only, replaces the keywords by the term concept which helps in identifying the relationships and patterns between the words more efficiently, helps in capturing the associative structure between the words. The limitations of this model includes: quite expensive if used in collaboration with large documents, the output generated at a slow rate.

3.4Latent Semantic Analysis

Latent semantic indexing is related to the analysis of hidden meanings of the words and how they often occur in the document. The meaning of the word can be inferred from the relationship between them. All these words are put together in a document called latent semantic space[13].
Based on the information in the document and a principle of Singular Vector Decomposition (SVD), term document matrix is generated.
It suffers from some limitations as the input domain does not represent any structural information. It is also unclear whether the similarity between the words actually holds or not

4. Comparative analysis of Information Retrieval models

If comparison is done among these models, several distinct features come up. The comparison among the Boolean model and Latent semantic models reveals that the Boolean model is based on the Boolean operators whereas Latent Semantic Indexing model is based on Singular Vector Decomposition principle. The Boolean model represents the query in the form of Boolean expressions consisting of AND, OR operators but the Latent Semantic Indexing model represents the query in the form of weighted vectors. If Latent Semantic Indexing is compared with the vector space model, analysis shows that in the latent semantic model, 'keywords' are replaced by the term 'concept' but the vector space was based on the concept of 'keyword'.

s.no.	Parameter	Boolean model	Vector space model	Latent Semantic Indexing model
1.	concept	Based on Boolean operators where query is given as an input including the Boolean operators.	Based on the concept of keywords and vectors	Extension of vector space based on the concept of Singular Vector Decomposition.
2.	Representation	Query in the form of Boolean operators. No specific matrix is generated.	Represents the documents in the form of vectors.	Represents the document in the form of matrix called term-document matrix.
3.	Type of the information	It does not consider any semantic information.	Considers semantic information.	Considers semantic information.
4.	Advantage	Easy to implement.	Simple	Not restricted to words only since it replaces 'keywords' by 'concepts'
5.	Disadvantage	Does not rank documents.	Poor documentation.	Unclear about similarity between words.
6.	Word occurrence	Does not tell about number of occurrence.	Tells about number of occurrence.	Tells about number of occurrence of words by term-document matrix.
7.	Output	If gives the exact match of the output to a query.	It shows the best match.	It shows the best match.

Table 1: comparison among different models

5. Conclusion

The information retrieval models are efficiently used to generate the output of the desired information. But all the models are based on different concepts and assumptions. Among all the models, the Boolean model is easiest model to be implemented in case of simple queries. If queries become complex, this model is quite difficult to be implemented. The Latent Semantic model used Singular Vector Decomposition (SVD) of word-by-document matrix but does not tell anything about structural information. There is a variation of latent semantic indexing model called Latent Dirichlet Allocation model which is a probabilistic generative model for finding latent semantic topics in large collection of data.

References

[1] Arash Habibi Lashkari and Fereshteh Mahdavi, "A Boolean Model in Information Retrieval for Search Engines," presented at the International Conference on Information Management and Engineering, 2009© IEEE. DOI: 10.1109.

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

[3] Stefan Pohl, Alistair Moffat, and Justin Zobe," Efficient extended Boolean Retreival,"vol.26, no.6 June2012.

[4] Shuo Wang, Kaiying Xu, Yong Zhang, Fei Li," Search Engine Optimization Based on Algorithm of BP Neural Networks," presented at the Seventh

International Conference on Computational Intelligence and Security, 2011©IEEE. DOI: 10.1109
[5] Jasmina Armenska, Katerina Zdravkova," Comparison of Information Retrieval Models for Question Answering,"2012©ACM.
[6] Cataldo Musto," Enhanced Vector Space Models for Content based Recommender System,"2010©ACM.
[7] Claire Fautsch, Jacques Savoy," Adapting the tf-idf Vector- Space Model to Domain Specific Information Retrieval," 2010©ACM.
[8] Sonia Gaied Fantar, Habib Youssef,"Applying Information Retrieval to Distributed Hash table (DHT) System,"2011©IEEE.
[9] Ravina Rodrigues, Kavita Asnani," Concept Based Search Using LSI and Automatic Keyphrase Extraction,"2010©IEEE.DOI:10.1109.
[10] Bill Kules, Jack Kustanowitz and Ben Shneiderman," Categorizing Web Search Results into Meaningful and Stable Categories using Fast-Feature Techniques,"2006©ACM.
[11] Diana Inkpen," Information Retrieval on the Internet,"
[12] Alexandros Ntoulas, Gerald Chao and Junghoo Cho," The Infocious Web Search Engine: Improving Web Searching Through Linguistic Analysis,"2005©ACM.
[13] April Kontostathis," Essential Dimensions of Latent Semantic Indexing (LSI)", presented at the Proceedings of the 40th Hawaii International Conference on System Sciences, 2007.
[14] Juan Cao, Jintao Li, Yongdong Zhang and Sheng Tang," LDA-Based Retrieval Framework for Semantic News Video Retrieval, presented at International Conference on Semantic Computing, 2007©IEEE.DOI:10.1109.