2012 International Conference on Communication, Information & Computing Technology (ICCICT), Oct. 19-20, Mumbai, India

Performance Comparison of Various Information Retrieval Models Used in Search Engines

Shivangi Raman, Vijay Kumar Chaurasiya, Venkatesan S.
Division of MS(CLIS) and MBA(IT)
Indian Institute of Information Technology
Allahabad, India
rs72@iiita.ac.in, vijayk@iiita.ac.in, venkat@iiita.ac.in

***Abstract* – The process by which information is retrieved from the repositories is called as information retrieval. This process had been a manual process since many centuries but since the advent of computers since the past few decades, it has become automated. Certain models form the basis of information retrieval in automated systems and are categorized as conventional and unconventional. This is because the motive of information retrieval systems is not only to render information to the regular users of internet but also, help the novice users of internet to skim out the documents that they need.**

In our paper, we have suggested that the information retrieval models must contain some questions regarding the user intent as well so that the search engine is able to search the websites that may prove to be informative for the users. These questions may enable the system to find out the user intent because entering one word may give varied results based upon the various uses to which it is put so that refining of requirements may be helpful in yielding appropriate results.

Keywords - information retrieval, models, comparison, web search

I. INTRODUCTION

Information retrieval models have been used not only in the retrieval of text documents from various automation systems, but have also been used in other fields of natural language processing. The concept of information retrieval rests on the usage of three basic models of information retrieval: Boolean model, probabilistic model and the vector space model.

Looking into the three basic models of information retrieval, we find out that the information retrieval models are basically the different ways by which one would be enabled to find out the right document that would fulfill one's requirements. These ways are either conventional ways that look up for exact matches or the unconventional ways that have the approach of partial matching [1][2]. An instance of the conventional ways is the Boolean approach in which the exact match of the query is found and the document giving this match thereby satisfying the query is retrieved [1][2].

Apart from understanding the models of the various web searches, it is also significant for us to find out the places where each model can be more useful as compared to the other models. This implies that before incorporating the models of information retrieval in our automation systems, we should find out which model suits our requirement in the best possible manner. This can be known only after knowing the characteristics of the end-user [3] and the parameters of measuring the performance of the various search engines [2][4]. The study of the end user varies based upon the search environments provided by the various search engines [3]. Only when a programmer is able to understand the demands of his user, he becomes able to design a system that would be best suitable to serve his needs.

How one analyzes which model is better for which type of search, or which particular type of users depends upon the users as well as the results that one is able to get by using a particular type of model for information retrieval from search engines. The performance of retrieval models depends upon factors like:

a) *Precision*
b) *Recall*

The vector space model, unlike the Boolean model has the approach of partial matching. This model takes the weights of the terms of various documents in a document corpus. Then it decides upon the documents that are to be retrieved based upon the query entered by the user who intends to find the entered information.

The probabilistic model is another basic model that follows the approach of partial matching. Unlike the vector space model and the Boolean model, the probabilistic model has the approach of calculating the probability of the number of times a word may occur in a document and henceforth the number of documents that may be better fitting for the query that needs to be answered through the retrieval of documents. The probabilistic model works through the iterative approach. This implies that the probabilistic model tends to improve the search and get better results with every repetition of the query. Although this increases the effectiveness of the results, time consumed to get the desired result becomes more, which makes the result less valuable for the user seeking immediate information.

In our paper, we shall study these three models and their different ways of information retrieval and then compare them based upon the various parameters. Also, we shall conclude as to which model would be more effective when being used in a web search engine and the reason supporting our deduction.

An important aspect to foster the quality of results is to investigate the users' actions. The links on which the users

978-1-4577-2078-9/12/$26.00©2011 IEEE

click often lead us to know about their site preferences. This has been found by the analysis of the search engine transaction logs as performed by [5].

The paper proceeds as follows: Section I introduces the motive of the paper. Section II underlines the concepts of the various information retrieval models that have existed and are being used in the web search engines. In Section III, we shall compare the different models and try to analyze which model would be the best if used in the web search engines. In Section IV, we will conclude this paper.

II. IR MODELS USED IN WEB SEARCH ENGINES

The web search engines deploy the basic information retrieval models for performing web searches. These are the three basic models: the Boolean model, the vector space model and the probabilistic model[8][9][10]. The only difference is that these models are used in this case for the retrieval of information from an ever changing repository of information rather than being a static database that has got limited information stored into it.

One of the significant features that are needed by the search engines is the identification of the intent of the user behind entering a particular query. This when identified can solve the problem of decidability of the type of model to be applied while particular information is being found out from the repositories [6]. Jansen et al. [6] have made three categories of the users' intent: that which is informative, which can be navigated and which is transactional. Also, they have categorized these as the three categories of user intent. The users' intent was found out based upon his queries after analyzing millions of queries from the transactional logs. The queries can also be transactional, navigational and informative based upon the intent of the users. We further infer that the user's intent decides whether the queries should be satisfied with the use of the Boolean model, the probabilistic model or the vector space model. This can be decided in information retrieval systems where the search places an option before the user if he wishes to get results based upon the exact match or partial match.

Lee et al. [7] have also tried to identify the intentions of the user in web search and has laid emphasis upon the amount of effect the intentions play upon the resultant retrieved.

We have tried to throw some light upon the various information retrieval models in this paper as recapitulation and analyses. The precise description of the three models of information retrieval is as given further.

a) The Boolean model:

Basically the Boolean model is a conventional model that works with the underlying principle of finding the document matching the exact terms of the query [1][10]. The Boolean model does not perform partial matches. We can understand this by the concept of 1s and 0s as followed in the Boolean algebra – either yes or no, or, in other words, either a document of exact match is found or it is not found. This may work out to a great extent for those who are adept at the usage of information retrieval systems, but it may possibly not be good for those who are new to the use of computers or information retrieval systems to use the Boolean information retrieval techniques.

In case of web search, we have to consider not just the researchers and the scientists or computer professionals, but also those people who are not used to of using computers as a medium of information gathering. Boolean model may not be so useful in case if the web search is being carried out by novices. Also, in Boolean search, unless otherwise programmed, the system will not be able to search out for synonyms because of which the user may sometimes lose the significant documents. It is seen that the significant documents also use synonyms in some cases because of which the documents may not seem similar for Boolean search but they are actually much related to the topic being searched. For such cases, it is advisable that the searches are made by using other techniques like vector space model or the probabilistic model which may although be challenging to implement yet yield better results.

b) Vector space model:

The vector space model is another model for information retrieval in which the information is retrieved through the unconventional methods. The retrieval of information is done through partial matches. The method of term weighting is applied in this model. Frequency of each word in each document is found followed by the weight of words in the documents with respect to the other documents and the query as entered by the user is found.

The vector space model gives much better results as compared with the Boolean model of information retrieval. Also, as compared to the probabilistic model, the vector space model is not very complex either. It gives results of 'nearly exact matches', in which sometimes such documents are also found which may not be completely useful, but may be utilized to some extent. It deploys the use of TF-IDF weighting and documentation of the documents from which the information is to be retrieved. Vector space model is considered to be one of the earliest historical models that may be used for information retrieval[12].

c) Probabilistic model:

The probabilistic model of information retrieval uses the unconventional methods of matching of documents in which nearly exact match is found and not the exact match. As the name of the model suggests, this particular model of information retrieval when used in search engines or in other systems for the purpose of document retrieval, gives better results[11].

Although the results of retrieval are better when they are given with the use of the probabilistic model, they are retrieved with comparatively lesser speed because the retrieval is improved gradually.

The term frequency inverse document frequency algorithm decides upon whether a certain word should be made a keyword or not, but ultimate finding of the document rests upon the type of match carried out by the different search engines that have been designed for fulfillment of this purpose.

In [13], the authors have proposed a gravitation based model for information retrieval which they have derived based upon Newton's laws of gravitation. This model is used in structured document retrieval and associates the attributes of a document to three different terms: type, mass and diameter. The model is peculiar in itself. On one hand, where it uses rectangular shapes to represent documents that have continuous values or hidden meanings of the terms in a document, it uses spheres to represent discrete values or terms that are explicitly used in it.

$$\text{Recall} = \frac{\text{Number of relevant documents}}{\text{Number of relevant retrieved documents}}$$

These are tabulated in Table 1.

Other parameters that can be taken into consideration before deciding upon the models that should be used for information retrieval are f-measure, fallout, average precision, r-precision, mean average precision[8].

TABLE 1: A COMPARISON OF THE VARIOUS RETRIEVAL MODELS BASED UPON CERTAIN SIGNIFICANT PARAMETERS

S. No.	Parameters	Boolean Model	Vector Space Model	Probabilistic Model
1.	Orientation	Query-oriented, exact match	Partial match	Partial match
2.	Basic Approach	Conventional	Non-conventional	Non-conventional
3.	Precision	Either all or none of the documents are retrieved	Document retrieval based upon the weights of the terms with respect to the query and the other documents.	Document retrieval based upon the probability of occurrence of a term occurring in one document as compared to the same term occurring in the entire document corpus.
4.	Recall	Either all documents are retrieved or none of them is retrieved because there is an exact match.	Better recall rates than the Boolean model because of the 'partial matches' and term weighting. The effect of larger number of documents may be adverse on the rate of recall, since it is not always possible to find out the weights of words in an ever extending and changing repository.	Better recall rate than the vector space model and the Boolean model because it does not make much difference to the calculations overloading in case of probabilistic model.

III. COMPARATIVE ANALYSIS OF VARIOUS MODELS

The three models can be compared based upon various parameters which are as follows:

a) *Orientation*: Whether there would be exact query matching or partial matching.

b) *Basic Approach*: Conventional approach or non-conventional approach.

c) *Precision*: Ratio of the number of relevant documents relevant to the number of documents that have been retrieved [4].

$$\text{Precision} = \frac{\text{Number of relevant documents}}{\text{Number of retrieved documents}}$$

d) *Recall*: Ratio of the number of relevant documents actually available in the repository to the number of relevant documents that have been retrieved[4].

Therefore, we see that the three models of information retrieval have their own benefits and their own areas losses. Where on one hand, the Boolean model is the one in which the exact match is obtained, this model does not give an answer in places where the user is not sure of what he wishes to retrieve. In the same way, the vector space model and the probabilistic model have their own unconventional way of performing matches in information retrieval by carrying out partial matches.

It is an essential feature of search engines to show a variety of results to the users as it is not possible to find out the exact intent of the user at once. Researchers underline the fact repeatedly that the web was not designed in an orderly fashion and is still not well-ordered[14]. The fact that the web still remains to be tamed is an implicit and well-realized requirement of most semantic web researchers.

Web has enhanced word repository added to it quite frequently in it. This causes the web to become a plethora of words that may be lexically analyzed and may be related to

each other as being synonyms or opposites of each other based on their usage in different contexts[15]. This, however, does not solve the problem of the identification of intent of the user who aims to extract information about a certain word or phrase from the web. It does take time to know about the user's intent in many cases [9].

Web, unlike human beings cannot understand the abstract and implicit meanings behind the instructions given by the users. Mining the indexes and analyzing the content of the web, thereby, trying to guess the intent of the user has become the need of the hour in the case of search engines to enhance the searches made by the user. Page-count-based-metrics [15] consider the co-occurrence of words in documents and calculate the number of times they occur in a document. So, in order to guess the exact intent of the user, it becomes important that the search engine first retrieves all the information that is in relation to a particular query and then let the user check out. For example, if we take the word "Maya" and try to find it on Google search engine, we get the following result:

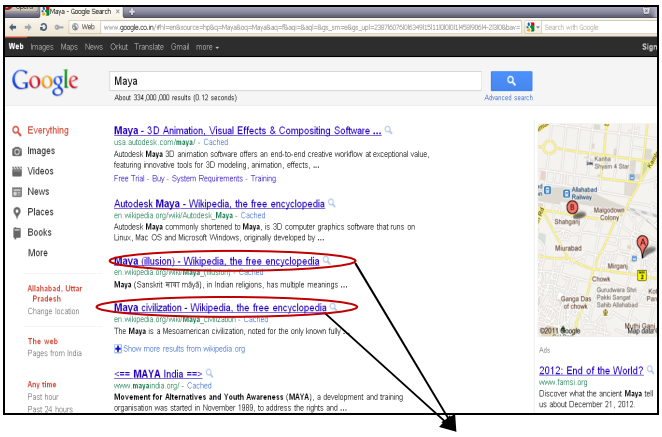

Figure 1: Entering the word "Maya" on Google Search Engine yields varied results

In the example illustrated in figure 1, we can observe that the two results of the word "Maya" are entirely different although they are meaningful. In such a case it entirely depends upon the users to determine whether the data provided would serve as a piece of information for them or not.

IV. CONCLUSION

We outlined the conditions that govern the performance of different information retrieval models. Also, we analyzed these models based upon their properties. We conclude that search engines should use some such hybrid model in which they can incorporate the properties of all the three models so that the search engine is able to identify the user intent and is able to decide upon the matches to proceed for, because it is indeed essential for the search engines to identify the intent of information retrieval of the user, or, in other words, the reason for or the use of the information that is being searched for.

To understand this better, we take another example of the asbestos sheets that are brought to use in chemistry laboratories as an apparatus, as a building material to safeguard from fire (fire resilience) and in electric heaters for cooking or keeping the room warm during winters. We need to make sure of the intent of the user if he wishes to search for information that would list the use of asbestos sheets or that a catalog that would help a user to find the rates at which the asbestos sheets are available in varied quantities.

REFERENCES

[1] Arash Habibi Lashkari, Fereshteh Mahdavi, Vahid Ghomi, "A Boolean Model in Information Retrieval for Search Engines" in 2009 International Conference on Information Management and Engineering, 2009, pp. 385-389.

[2] Jiang Hua, "Study on the Performance of Information Retrieval Models" in 2009 International Symposium on Intelligent Ubiquitous Computing and Education, 2009, pp. 436-439.

[3] Deitmar Wolfram, "Search characteristics in different types of Web-based IR environments: Are they the same?" in Elsevier's Journal of Information Processing and Management, Vol. 44, 2008, pp. 1279-1292.

[4] Mei Kobayashi, Koichi Takeda, "Information Retrieval on the Web" in ACM Computing Surveys, Vol.32, No. 2, June 2000, pp- 144-173.

[5] Ying Zhang, Bernard J. Jansen, Amanda Spink, "Time series analysis of web search engine transaction logs" in Elsevier's Journal of Information Processing and Management, Vol. 45, 2009, pp. 230-245.

[6] Bernard J. Jansen, Danielle L. Booth, Amanda Spink, "Determining the user intent of web search engine queries" in WWW 2007, ACM, 2007, pp. 1149-1150.

[7] Uichin Lee, Zhenyu Liu, Junghoo Cho, "Automatic identification of user goals in web search" in WWW 2005, ACM, 2005, pp. 391-400.

[8] Modern Information Retrieval Addison-Wesley Longman Publishing Co., Inc. Boston, MA, USA ©1999 ISBN: 020139829X.

[9] Olivier Chapelle et al., "Intent based diversification of web search results: metrics and algorithms" in Journal of Information Retrieval, LLC 2011, May 2011.

[10] A Singhal, "Modern Information Retrieval: A Brief Overview" in IEEE Data Engineering Bulletin, Special Issue on Text and Databases, Vol.4, No. 4, December 2001.

[11] M. Steyvers, T.L. Griffiths, "Rational Analysis as a link between Human Memory and Information Retrieval" in The Probabilistic Mind: Prospects from Rational Models of Cognition, Oxford University Press, 2008, pp. 327-347.

[12] C. Zhai, "Statistical Language Models for information Retrieval A Critical Overview" in Foundations and Trends in Information Retrieval, Volume 2, No. 3, 2008, pp. 137-213.

[13] S. Shi et al., "Gravitation-Based Model for Information Retrieval" in SIGIR'05, 2005.

[14] C.C. Marshall, F.M. Shipman, "Which Semantic Web?" in ACM HT'03, 2003, pp. 57-66.

[15] E. Iosif, A. Potamianos, "Unsupervised Semantic Similarity Computation between Terms Using Web Documents" in IEEE Transactions on Knowledge and Data Engineering, Vol. 22, No. 11, November 2010, pp. 1637-1647.