IEEE International Conference on Advances in Engineering & Technology Research (ICAETR - 2014), August 01-02, 2014, Dr. Virendra Swarup Group of Institutions, Unnao, India

Comparative Study of Information Retrieval Models Used in Search Engine

Javed Ahmad Khan
Dept. of Computer Science & Engineering
Ambedkar Institute of Advanced Communication Technologies & Research, GGSIP University
Delhi, India
javedahmad1989@gmail.com

***Abstract*—Information retrieval is the methods of retrieving all the relevant information from a large collection of information. We will focus on the performance of the different information retrieval model based upon these models properties. Search engines used the property of these models for identifying the users given query for retrieving the information from the World Wide Web and form some specific application domain. If the user query is matched the search engine is able to precede the user query and retrieve the relevant information that is being searched by the user. Search engines which usually allow the information retrieval accuracy by using the Boolean Model, Vector Space Model and Probabilistic Model. In our paper we analyze precision and recall of these three models and compared by using different parameters.**

Keywords-Information retrieval, web search, Comparisions, Precision, Recall.

I. INTRODUCTION

Information retrieval is the procedure of retrieving relevant data from collections unstructured data that meet the user demands. Keyword based search method employed to retrieve the relevant text file from the indexed text files that are relevant for the user. The semantic web is an elongation of the current web in which information provides well-defined meaning that enables system and people for more honest understanding and can enable to operate effectively by seeing information from different informants [1]. 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. Information retrieval models retrieve the information basically different ways by which one would be make possible to find out the correct query that would fulfill the user requirements. These models are based upon the conventional or unconventional approach of partial matching [2] [3]. Boolean model use the conventional approaches in which the exact match of the query is found and the retrieved query satisfying the user needs. 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. It is difficult for us for choosing that which model is better in our automation system, and is suitable for our requirements in the best possible manner. This can be only possible when we know the characteristics' of the end user and the parameter of measuring the performance of the various search engines [4] [5]. When programmers know the needs of the users he becomes able to design a system that would be suitable to serve user needs. The performance of retrieval models depend upon these two factors like:

1) Precision
2) Recall

This paper is organized in the following six sections: 1) introduction, 2) Related Work 3) Precision and Recall for various information retrieval model 4) Comparisons of information retrieval models 5) Drawback of Information Retrieval models 6) Conclusion

II. RELATED WORK

The web search engines deploy the basic information retrieval models for performing web searches. These are the three basic models: Boolean model, the vector space model and the probabilistic model [6] [7] [8]. 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. A main important feature of the search engine is that to identify the intent of the user behind entering a particular query. If they identified the user need then they can solve the problem of decidability of the types of model to be applied while the specified information is being found out from the repositories [9]. After analyzing the millions of queries from the transactional log the user query was found that are satisfy the user needs. 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 interrogations should be satisfied with the use of the Boolean model, the probabilistic model or the vector space model. Lee et al [10] 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.

978-1-4799-6393-5/14/$31.00 ©2014 IEEE

III. Precision And Recall of Information Retrieval Models

We describe the results of three information retrieval models by taking a collection of 10 documents and giving the query as an input. On the basis of these analyses we compare these three models precision, recall and F-measures.

A. Vector Space Model

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 collection. 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. We show the process of vector space model in figure 2.

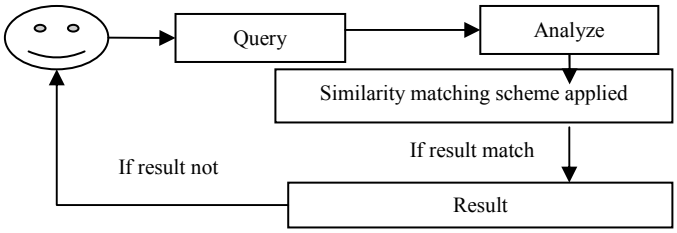

Fig.2. Show the process of vector space model

Vector space model used the unconventional methods for retrieving the information in which the retrieval of information is done through the partial matches. Query weighting concepts are used in this model. We calculate weight for each document or query that is given by the user. Vector space model gives much better results as compared with the Boolean model of information retrieval. Vector space model 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 [11]. Then a vector of index term weights is computed as the internal document representation. These weights are calculated by most often used TF-IDF scheme [15] [16].

$$W_{ij} = TF_{ij} \times IDF_i \quad \ldots\ldots (1)$$

$$TF_{ij} = FREQ_{ij} \div MAX\ FREQ_{ij} \quad \ldots\ldots (2)$$

$$IDF_i = Log\ (N \div n_i) \quad \ldots\ldots (3)$$

TF = Term Frequency, IDF = Inverse Document Frequency. $FREQ_{ij}$ is the number of occurrences of term T_i in document d_j, N is number of documents in collection, and n_i is the document frequency for term Ti in the whole document collection. In order to find some relevant document to a specific query Q it is necessary to represent the query Q in the same way as a document D_i (i.e. a vector of index term weights). Similarity between a query Q and a documents Di is computed as cosine of those two normalized vectors (document and query vectors).

$$SIM_{tf\text{-}idf}\ (Q, D_i) = ((Di \times Q) \div |D|\ |Q|) \quad \ldots\ldots (4)$$

We will take the set of data collection that consist the following documents.
D1: New York Times, D2: New York postal,
D3: Los Angeles Times, D4: Times of India,
D5: New Bharat Times, D6: Indian Times Coaching,
D7: Indian Postal Office, D8: Los Angeles Airport,
D9: Indian Postal Code, D10: Indian postal Tracking
Query: Indian, Postal, New, York, Los
TF of data collection are shown in table I. IDF, Similarity of documents and IDF for query are shown in table II.

TABLE I. Show the Term Frequency

Collection Set	D1	D2	D3	D4	D5	D6	D7	D8	D9	D10
Airport	0	0	0	0	0	0	0	1	0	0
Angles	0	0	1	0	0	0	0	1	0	0
Bharat	0	0	0	0	1	0	0	0	0	0
Coaching	0	0	0	0	0	1	0	0	0	0
Code	0	0	0	0	0	0	0	0	1	0
Indian	0	0	0	1	0	1	0	0	1	1
Los	0	0	1	0	0	0	0	1	0	0
New	1	0	0	0	0	0	0	0	0	0
Of	0	0	0	1	0	0	0	0	0	0
Office	0	0	0	0	0	0	1	0	0	0
Postal	0	0	0	0	0	0	1	0	0	1
Time	1	0	1	1	1	1	0	0	0	0
Tracking	0	0	0	0	0	0	0	0	0	1
York	1	1	0	0	0	0	0	0	0	0

TABLE II. Show the IDF, Similarity of documents and IDF for query

Collection Set		Similarity of documents	Documents	Query	IDF
Airport	1	0.01562	D1	Indian	.05
Angles	0.05	0.01562	D2		
Bharat	1	0.125	D3	Postal	.11
Coaching	1	0.125	D4		
Code	1	0	D5	New	.25
Indian	0.2	0.01	D6		
Los	0.5	0.0463	D7	York	.25
New	0.5	0.125	D8		
Of	1	0.0463	D9	Los	.25
Office	1	0.0363	D10		
Postal	0.33				
Time	0.2				
Tracking	1				
York	0.5				

By using above TF-IDF scheme we calculate the similarity of the document and query and draw the graph for the precision and recall for the user query that are shown in Figure 3.

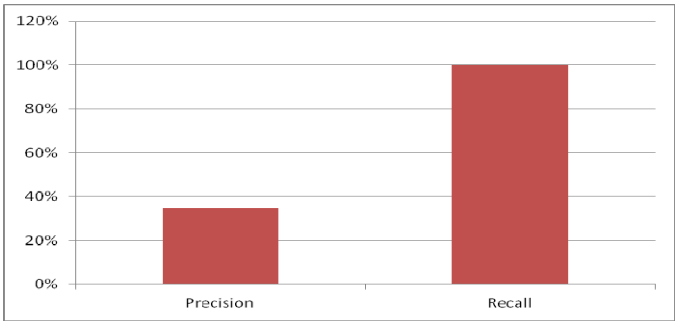

Fig.3. Graph show Precision and Recall for vector space model

B. Probablistic Model

The probabilistic model works through the iterative approach. Probabilistic model gives better result and improve the search result when user repeat the query multiple times. This model also show the improved result and time consumed when user retrieve the information. Probabilistic latent semantic index (PLSI) [12] is a technique in which the question is expanded and then Term Frequency Indexing (TFI) is applied to compute the document scores for information recovery. The probabilistic model uses the concepts of unconventional methods of matching, in which the user given query exactly match or not equal. This information retrieval model takes less time and yields better results [13]. Passed on a query Q and a aggregation of documents DI, a subset R of DI is taken to exist which takes precisely the relevant documents to Q. The probabilistic retrieval models, then ranks documents in decreasing order of probability of belonging to this set (i.e. of Being relevant to the information need). Which is noted as P (R | Q, D_i), where Di is a document in D.

Following the previous notation:

• D: documents are represented as a vector of words or index terms occurring in a document. Each term in the document, that is, each pair (T_i, D_i), has a binary associated weight 1 or 0, denoting the presence or absence of the term in the document.

• Q: queries are represented by a vector of words or index terms that occur in the query. Each term in the query, that is, each pair (T_i, Q) has a binary weight 1 or 0, denoting the presence or absence of the term in the query.

• SIM measures the degree of similarity of a document Di to a query Q as the probability of D_i to be part of the subset R of relevant documents for Q. This is measured in the probabilistic model as the odds of relevance, as given by.

$$SIM = P(R \div D_i) \div (\neg R \div D_i) \quad (1)$$

Once an initial subset of documents V is retrieved and ranked by the probabilistic model, the probabilities can be refined to:

$$P(T_i \div R) = |V_i| \div |V| \quad (2)$$

Where V_i is the set of retrieved document containing T_i.

$$P(T_i \div \neg R) = ((n_i - |V_i|) \div (N - |V|)) \quad (3)$$

That the non retrieved document is not relevant following process recursively we get

$$P(T_i \div R) = (|V_i| + n_i \div N)) \div (|V| + 1) \quad (4)$$

$$P(T_i \div R) = ((n_i - |V_i| + (n_i \div N)) \div (N - |V|) + 1 \quad (5)$$

If we have V= (D1, D2) and we want to compute the relevance Of D1.

Query**:** (New, Times)

We calculate the relevance and non relevance for these queries so assume

1) P (new / R) = 0.733 and P (time / R) = 0.75 both are represented by P1.

2) P (new / ¬R) = 0.022 and P4 = P (time / ¬R) = 0.50 both are represented by P2.

TABLE III. Probabilistic model data collection

Documents	Set of data collections												
	Airport	Angles	Bharat	Coaching	Indian	Los	New	Of	Office	Postal	Time	Tracking	York
D1	0	0	0	0	0	0	1	0	0	0	1	0	1
D2	0	0	0	0	0	0	1	0	0	0	0	0	1

On the basis of user query we draw the graph for relevance and non relevance the precision and recall that are shown in Figure 4. P1 show the relevance for New and Time and P2 show irrelevance for New and Time query.

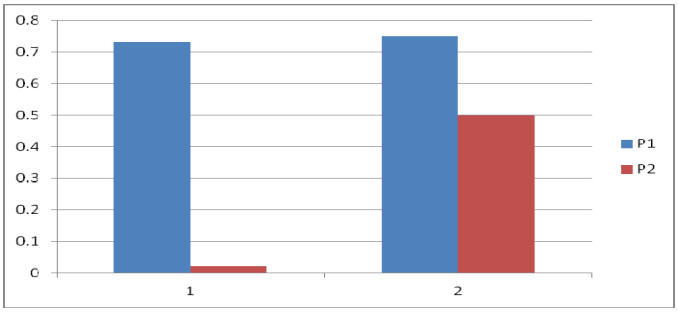

Fig.5. Graph show the relevance and non relevance queries for probabilistic model

C. Boolean Model

Basically the Boolean model is a conventional model that works with the underlying principle of finding the document match the exact terms of the query [2] [8]. The Boolean model performs exact 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. Boolean model may not be so useful in case if the web search is being carried out by inexperienced person. Boolean model not be able to search out for synonyms because the user loss the significant information sometimes. Vector space model and probabilistic model are playing important role to retrieve the synonyms information's.

The Boolean Model is a simple retrieval model based on set theory and Boolean algebra.
Documents are represented by the index terms extracted from documents, and queries are Boolean expressions on terms.
• D: The elements of D are represented as sets of index terms occurring in each document. The factors of D are represented as sets of index terms occurring in each text file. Terms are treated as logic propositions, denoting whether the condition is either present (1) or absent (0) in the papers. Text files can thus be seen as the Conjunction of their conditions.

• Q: Queries are represented as a Boolean expression composed by index terms and logic operators (AND = ∧, OR = ∨, NOT = ¬) which can be normalized to a disjunction of conjunctive vectors (i.e. in DNF2, disjunctive normal form).

• SIM is defined by seeing that a document is predicted to be relevant to a question if its index terms satisfy the query formulation. Boolean model results on the basis of the query are show in figure 7.
Query: **(**new, york, time**)**

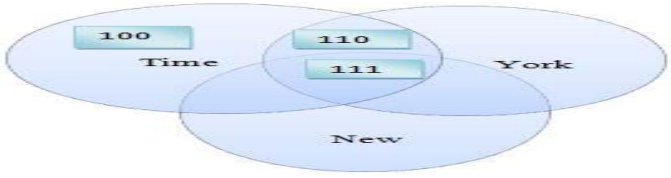

Fig.7. Represent the relevance and non relevance for given query with the help of Boolean models

This query is composed of three different terms: Time, New and York, and it can be written in a disjunctive normal form as qdnf = [(1,1,1)∨ (1,1,0)∨ (1,0,0)]. where each of the elements is a binary weighted vector associated with the tuple (Time, New, York). These binary weighted vectors are sent for the conjunctive components of qdnf. Caved in the query q, the subsets of documents that satisfy the query are:

- Those containing the three terms: (1, 1, 1).
- Those containing the word retrieval, but neither York nor New: (1, 0, 0).
- Those containing the word Time and York, but not New: (1, 1, 0).

IV. COMPARISION OF INFORMATION RETRIEVAL MODELS

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

a) Basic method:

The three models used the Conventional approach or nonconventional approach.

b) Precision:

Ratio of the number of relevant records retrieved to the total number of irrelevant and relevant records retrieved. It is usually expressed as a percentage.

Precision = (Number of relevant document) ÷ (Number of relevant document) ×100

c) Recall:

Ratio of the number of relevant records retrieved to the total number of relevant records in the database. It is usually expressed as a percentage.

Recall = (Number of relevant documents) + (Number of not relevant document + Number of relevant document retrieved)) ×100

d) Adaptation:

Whether there would be exact query matching are found or partial matching found.

e) Real time application:

A Real time application (RTA) is a program that functions within a given time frame.

f) Preprocessing time application:

Where something is maintained and not after or before the event has happened. This processing method is used when it is essential that the input request is dealt with quickly enough so as to be able to control an output.

g) Retrieval time application:

Presents information based on a user local context in an accessible yet non-intrusive manner.

TABLE IV. A comparison of the various retrieval models based upon certain significant parameters

Parameters	**Information retrieval models**		
	Boolean Model	**Vector Space Model**	**Probabilistic Model**
Basic Method	Conventional	Non conventional	Non conventional
Precision	Either all or none match retrieve	Retrieve according to documents weight	Retrieve document according to their occurrences probability
Recall	Either all or none documents retrieve because they used the concept of exact match	Better recall rate compare to Boolean model because partial match and document weighting concepts are used	Better recall rate compare to Boolean and vector space model
User Query	Query Oriented Match	Partial Match	Partial Match

Real Time Application	Boolean models gives average performances in real time application	Vector space model Gives better performance in real time application	Probabilistic model gives good performance in real time application
Preprocessing Time Application	Average preprocessing time	Good preprocessing time	Better preprocessing time
Retrieval Time Application	Good retrieval time	Average retrieval time	Better retrieval time

V. DRAWBACK OF INFORMATION RETRIEVAL MODELS

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. Boolean model suffers from two major drawbacks.

a) Retrieval strategy is based on a binary criterion (i.e. a document is predicted to be either relevant or non relevant) and therefore it does not provide a proper basis for ranking the retrieved results, which may likely result in low precision levels when the retrieval space is too big.

b) It is not always easy for most users to translate an information need into a Boolean expression with logic operators, which significantly decreases the usability of the latter.

Vector space model and the probabilistic model have their own unconventional way of performing matches in information retrieval by carrying out partial matches. Web was not designed an orderly fashion and is still not well-ordered [14]. So does not solve the problem of the identification of the user given queries, whose aims to retrieve the information about a particular query from the web. Its take time to know about the user intention in many cases. Search engine used concept of page-count-based-metric [15]. Consider the co-occurrence of query in document and calculates the number of times they occur in a document. For retrieving the exact user gives query it's become important that the search engine first retrieve all the information that is relation to particular query and then let the user check out.

VI. CONCLUSION

Information retrieval models follow two basic concepts that are conventional and unconventional approach. These three models retrieve the user query from the search engine after using the similarity matching concepts they provides the relevant information to users. . In our paper we have provide the concepts that the search engine is able to search the websites that are provide the reliable information to the user according to their need. These three models are helping the user in different conditions. We also show the comparisons of these three models in various parameters as shown in Table IV. When a user enters a word here are many information related to these words are retrieved but these models helpful in yielding appropriate results.

REFERENCES

[1] Berners-Lee, J. Hendler, and O. Lassila, The Semantic Web, Scientific american, vol. 284-5, 2001, p. 34-43.

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

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

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

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

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

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

[8] 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.

[9] 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.

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

[11] 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.

[12] S. Dumai s. Improving the retrieval of information from external sources. Behavior Research Methods, Instruments, and Computers, 23(2):229–236, 1991.

[13] Laurence A. F. Park and Kotagiri Ramamohanarao, "Efficient storage and retrieval of probabilistic latent semantic information for information retrieval," The VLDB Journal - The International Journal on Very Large Data Bases archive, vol. 18-1, 2009 p. 141-155.

[14] 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

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

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