Published by :
http://www.ijert.org

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 10 Issue 09, September-2021

Comparison of basic Information Retrieval Models

Manal Sheikh Oghli
Web Science program, Syrian Virtual University
Damascus, Syria

Muhammad Mazen Almustafa
Web Science program, Syrian Virtual University
Damascus, Syria

***Abstract*— it was necessary to provide an information retrieval model capable of meeting user requirements in an effective manner, in light of the increasing growth and the huge amount of digital information in recent decades.**

The Information Retrieval process depends on the matching process largely between representations of the user's desire, which is expressed through the query, and stores of information to return results related to the user's desire.

As the challenge today is no longer, providing information stores, therefore, the biggest challenge that facing researchers is the ability to retrieve an appropriate information related to the needs of the user.

In this paper, we review basic information retrieval models, that are classified according to the mathematical dimension to arrive to a description of the most effective model in retrieval operations, and to demonstrate to the most widespread among other models, which is, Vector Space Model and its strength to excel in the event that the weaknesses points, that it suffers from are addressed, which is represented by not setting fixed standards, for terms' weighting ,in addition to this model that assumes independence of terms from each other.

Keywords*— *Information Retrieval Models IRM, Boolean Model, Vector Space Model VSM, Probabilistic Models, Term Weighting)

I. INTRODUCTION

The term Information Retrieval was first used during Calvin Mooers' presentation of a research paper at a 1950 conference, as he wrote, "The problem under discussion here is machine searching and retrieval of information from storage according to a specification by subject… ". [1] [2]

Mooers used this term to describe the process by which a user could convert their need for information into an actual list containing a set of useful references, and explained that information retrieval is another, more general name for producing a demand bibliography. [2]

Information retrieval models considered a blueprint for implementing an actual retrieval system as the retrieval system predicts and explains what the user wants by analysing the user-defined query. [3]

Models provide different techniques and methods for matching stored documents to a query. The main goal of information retrieval models is to find documents relevant to the information needs of a large group of documents. [4]

II. THE AIM OF THE RESEARCH

A brief historical overview of the emergence of the concept and development of Information Retrieval, and its multiple models, with a brief description of the working method and processing algorithms in each of them, with a focus on the aspects that distinguish each model from the other, and then compare these models with the aim of describing the best model can meet with the requirements of User in terms of performance and related results.

Despite the emergence of many information retrieval models and the development that occurred in this area of knowledge, most models still suffer from their limitations in meeting the user's desire to obtain the required information. In addition, the many models that appeared in information retrieval systems depend in their way of working on the basic models represented by the Boolean model, the probabilistic model and the vector space model.

The aim of this paper is to compare these models to guide workers in this field to the simpler and more efficient model by reviewing the points of strengths and deficiencies in it to overcome them in future researches in order to reach an effective and efficient retrieval model.

III. CLASSIFICATION OF INFORMATION SYSTEMS RETRIEVAL MODELS:

Information retrieval systems models differ among themselves in general in the way of representing documents and queries, and the methods of matching and arranging. These models could be classified according to two dimensions: the mathematical base and the characteristics of the model.

Retrieval models were classified according to the mathematical base dimension into:

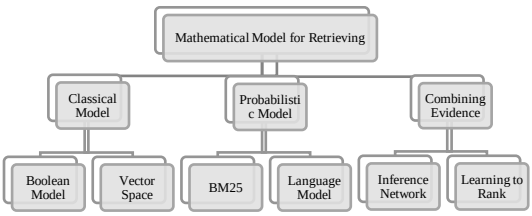

Fig. 1: Classification of Mathematical Information Retrieval Models [5]

The following is a review of the most important features that distinguish each of these models, explaining the Positives and limitations of each.

A. Classical Models:

1) Boolean Model:

The Boolean model is the first form of information retrieval [3]. One of the oldest and simplest models in this field, as it based on logical algebra [4], and the principle of Exact Match [3]. There is no room for partial matching in this form.

Where documents are represented by a set of terms (also known as index terms) [4] [6] .Then it is classification into a class in which the terms of the query mentioned, and a class in which the terms not mentioned. This classification means that there is no sort of arrangement in evaluating the relevance of the documents to the query.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 10 Issue 09, September-2021

The user's information needs are identified by a combination of basic logical transactions defined by George Bool, which are three (AND, OR, NOT) meaning intersection, addition and difference, and are used during formulation of the query [3] [4] [5].

Despite their limitations, these models still used in many retrieval systems. It also gives expert users the feeling that they are able to control the system largely more than others [7].

The development of these models appeared, for example, the Extended Boolean Model, which was described in an article published in ACM in 1983 by (Gerard Salton, Edward A. Fox, and Harry Wu).

Through this model, it became possible to perform partial matching and term weighting. It combines the characteristics of a Vector Space Model (it will be explained in the next paragraph) with the characteristics of a Boolean model [6]. However, the judgment in this model on the importance of the documents to the query, depending on whether or not the term mentioned in them, without taking into account the repetition or its mentioning in the one document that is taken into account by the other models.

2) Vector Space Model:

Gerard Salton and his colleagues suggested this model in 1983 [8]. It was based on the similarity criterion proposed by Hans Peter Luhn in 1957, who was the first one suggested the statistical model for searching for information based on the similarity criterion between inquiries and documents. HP Luhn formulated the similarity criterion as follows:

"The more two representations agreed in given elements and their distribution, the higher would be the probability of their representing similar information…" [9]

Based on this criterion, Salton and his colleagues considered that both documents and queries could be represented as vectors in Euclidean space, so that each term is assigned an independent dimension, and then they calculated the similarity between vectors using the cosine between the vectors representing both the document and the query. [8]

This model was considered one of the algebraic models and the most widespread. The text is represented in it by a vector of terms independent of each other. The terms are words and phrases that represent indexing terms. [7]

According to the statistical model, the content of the document is viewed as a Bag-of-Words [6].

This means that the document content includes unordered and irregular frequency terms within the document content.

Index terms set weights for both documents and queries. Then measuring the similarity or distance between the document and the query through several methods, we mention as example (Dot Product, Euclidean, Manhattan, Cosine…). [5]

The biggest challenge facing this model is to set the appropriate value for the vector components, and this problem is known as Term Weighting in addition to terms independence as this model does not take into account the terms link between them. [3]

B. Probabilistic Models:

Maron, Kuhns suggested the initial idea of probabilistic models in information retrieval systems in a paper titled Probabilistic Indexing and Information Retrieval published in 1960 [7] [10].

It was considered the first scientific work to deal with the use of the probabilistic approach in retrieving information, and on this basis, what is known as Probabilistic Indexing appeared, and since then many models have been developed that rely on different techniques to estimate probabilities.

Probabilistic models are based on the Probabilistic Ranking Principle (PRP) [6] [7]:

"If a reference retrieval system's response to each request is a ranking of the documents in the collection in order of decreasing probability of relevance to the user who submitted the request, where the probabilities are estimated as accurately as possible on the basis of whatever data have been made available to the system for this purpose, the overall effectiveness of the system to its user will be the best that is obtainable on the basis of those data." [11].

Feedback plays an important role in such probability models, as it uses the historical information of the document to calculate its probabilities of relevance to the query.

Probabilistic models differ depending on the assumptions on which they are based [6].

The classic probabilistic model introduced by S.E. Robertson and K. Parck Jones in 1976 assumes the independence of the term as this model was defined as the binary independence retrieval model (BIR). This model was depended on the concept that the set of documents stored in an information retrieval system was divided into two independent binary groups from each other; the first group is known as the join group whose content is related to the query, and the second group is known as not join group whose content is irrelevant to the query. [12]

As the set of all possible outcomes is called the sample space and thus the probability of P(R) in the sample space can carry one of two values {0,1} where Irrelevant = 0, Relevant = 1. [3]

We must mention that the value of P (...) changes with the change of the random variable converter R, so when we assign different values to the random variable converter R or different values of the sample space. Therefore, we are talking about different values of probability P. [3]

This model depends on the method of using probability theory as the basis for the treatment process. Where, according to the followed probability model, the probabilities in which the document is relevant to the user's query are calculated or estimated.

The basic idea of this model is the hypothesis, that the information retrieval system includes documents related to the user's query, that are completely relevant and there is another group that is far from this relevance, according to this model, the related group of documents is called the ideal answer set, and by providing a complete description of this set of documents (ideal answer set). The problems of retrieving the content of documents diminish, and yet another obstacle appears in the difficulty of definitively knowing what these characteristics and features are.

The primary effort of a probabilistic model is the initial guessing to determine the characteristics of the keywords contained in the documents, which have linguistic and idiomatic connotations, which contribute to the marking of these characteristics, allowing the creation of a probabilistic

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 10 Issue 09, September-2021

initial description of the ideal response set to the documents query.

In spite of the results achieved by this model, some saw that it isn't better than the results achieved by the classical models, which led, from their viewpoint, to the system developers not convinced of switching to this model largely.

Among the probabilistic models, we mention:

1) Best Match BM25:

This model was developed in the 1970s and 1980s by Stephen E. Robertson, Karen Spärck Jones, and others. It is a currently widely used probability model, in which documents are classified based on the estimated probability of the documents fit in the query. [6]

This model was developed from the Binary Independence Model (BIM), which is a classic probabilistic model that represents both documents and queries as binary vectors by combining range frequency and Document Length Normalization. [6]

This model is usually referred to as the Okabi BM25 because the Okabi retrieval system implemented in London in the 1980s was the first to implement this model.

The BM25 model has a lot in common with the TF-IDF terms weighting algorithm. Both algorithms use the term frequency and the inverse document frequency, but the definition of factors differs slightly between the two models. [13]

Both models define the weight that is given to each term as a result of combining the inverse document frequency and the term's frequency and then calculating the term's weight for the entire document for the specified query. The most prominent difference between the two models is that the BM25 takes into account the length of the document while it has no effect in the traditional TF- IDF used in the classic form. [13]

2) Language Models:

Language models applied to information retrieval by a number of researchers in the late 1990s, among them we mention: (Ponte and Croft 1998, Hiemstra and Kraaij 1998, Miller etal. 1999). [3]

Language models developed in the early 1980s for automatic speech recognition systems. It studies the probability distribution over all words sequence in a language. [3] [6]

The language model estimates the probability of words sequences. There is a language model associated with each document and this model may contain the queries most relevant to it. Language model-based methods are a widely used model. [6]

Once we know the probabilities of individual words, i.e. assign a probability value to each term, we can calculate the probabilities for any phrase or sentence in the language. The higher the probability of the sentence, the more likely that the sentence will be of interest. For example: Assume that the probabilities associated with words (information, retrieval, forms) are (0.1, 0.15, 0.05) respectively, then the probability of the phrase: Information Retrieval Models is 0.3. [6]

Some other probability models may specify a very high probability of the word "retrieval", indicating that the probability of this message that we are writing, for example, will be a strong candidate for retrieval if the query contains this word. [5] [3]

The language models take the same starting point, as the probability indexing model, that was set by (Bill Maron and Larry Kuhns), is a probability value that is assigned to the different indexing terms, which the document contains, so that each document has a set of indexing terms and each term has a probability value, that determines its importance for the query Including the terms it contains. Then, the language model for each document will be designed to follow this approach. [5] [3]

One of the advanced models that emerged using this technology is the Natural Language Processing Model.

As it does not rely on terms of query and document only, but it processes sentences and formulas, and this model works on matching them. Therefore, it requires building systems that work on natural language texts on three levels of processing: syntactic analysis, semantic analysis, Pragmatic analysis.

C. Combining Evidence:

In these models the technology for understanding the content is used for the documents and queries, and then it used also for concluding the probable relations between documents and queries. Therefore, the information retrieval process is an inference process or logic thinking in which we can evaluate the probability of the extent of the documents' fitness with inquiries which determines the user' need.

One of these models is:

1) Inference Network:

In this model, the documents retrieval is modeled as a process of logical inference, and its probability was evaluated to meet with the user' need for information in which we can express it by one or more of queries by analyzing the document as inference network will be the mechanism of concluding these relations kinds [14].

Most of the techniques used by information retrieval systems could be applied within this model. [7]

In the simplest application of this model, the document gives the term a certain power, and then the values for the terms contained within the query are combined to calculate the numerical result of the query in relation to the documents.

In other words, the power given to a term may be considered as the weight of the term in the document. Thus, the classification of documents in this model becomes similar to the arrangement in the vector space model or the probabilistic models. The strength of the term is indefinite and therefore any algorithm or form of term strength can be used within the document or query. [7]

This model relies on three basic things:

- Support the use of multiple document representation schemes.
- Allow the merging of results from different types of queries, which retrieve different documents for the same specific need for information.
- Flexible matching between terms or concepts mentioned in the queries and those specified for documents. This is done by improving recall by using cognitive matching of query concepts and documents and their representations without significantly degrading accuracy. [14]

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 10 Issue 09, September-2021

	Relevant	Non Relevant
Retrieved	***True Positive TP***	***False Positive FP***
Not Retrieved	***False Negative FN***	***True Negative TN***

2) Learning To Rank:

The learning to rank algorithm is part of the information retrieval for large documents. Data consist of queries and documents which are represented as vectors. [6]

It is divided into three models (Pointwise, Pairwise, and List wise). In the first model, pointwise, the arrangement is done as a traditional classification process, so the result is Class, so the goal is to reduce misclassification of queries and documents. In the second model, Pairwise, which is the process of converting the arrangement to a pointwise classification process, the goal of this process is to increase the number of pairs that were classified out of order, and the third, list wise, is very similar to the pair's wise model except that it deals with lists of rows and classes. [6] [15]

This form is applied to the Training Set test, and the documents are sorted according to their relevance and importance. [6] [15]

IV. Evaluating in IR Systems:

Retrieval effectiveness is a measure of how well the documents retrieved by a system meet users' needs. The process of determining the retrieval effectiveness for a given query is referred to as effectiveness evaluation [6].

One approach to measuring effectiveness of an IR system which is widely used is precision and recall. [6] [16].

However, precision and recall are inversely related. That mean, obtaining higher levels of precision can be obtained through lower recall values [6] [16].

Effective information retrieval systems must retrieve the largest possible number of relevant documents (i.e. have a high recall), and must retrieve a very small number of irrelevant (i.e. high-precision) documents. Unfortunately, experience has shown that these two aims are completely opposed. [7] Therefore, in some evaluating systems, the F1 factor is used as a measure of combining precision and recall. [16]

Below we explain the calculation methods for: Recall, Precision, and F1.

$$Recall = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad , \quad Precision = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} = \frac{2 * \text{TP}}{2 * \text{TP} + \text{FN} + \text{FP}}$$

V. Comparison of Main Models:

By reviewing the previous models, we find that the basic models, which the rest of the other models depend on, are three: Boolean model, the vector space model, the probabilistic model and all other models, which are attempts to develop these three basic models or combine them.

We have found that the Boolean retrieval model allows users to formulate complicated logical expressions and queries that may be difficult for an ordinary user, and this model does not provide an arrangement for the retrieved documents. This set of documents will either be nearly empty (low recall) or contain too many documents (low precision) due to the use of an exact match criterion. Consequently, this model is more useful for data recovery than information retrieval because all terms in it are of equal weight. [17]

While we found in the VSM the possibility of applying a set of values to each term, either in the representation of documents or in the user's query. Common terms are not important in this model due to the application of inverse document frequency as importance is given to rare terms.

In the VSM model, a long document that can contain the same query terms only in the title and the abstract may be very relevant to the query, but in this model it will be of low importance compared to a short document that contains the same terms in the appendix, which is considered one of the drawbacks of this model. Another disadvantage of representing VSM documents is that the terms arrangement is missing and documents containing query terms close to each other cannot be preferred over documents that contain separate terms in different parts of the document. [17]

The probabilistic retrieval model relies on assumptions that have explicitly made, such as assuming that 50% of the document containing the term is closely related to that term. However, not all assumptions correspond to reality. Therefore, the total number of relevant documents must be estimated, and the calculation of the probability value P (…) which is a constant is not always correct. Therefore, the probabilistic retrieval model to achieve accurate results requires that the terms will be independent. It ignores the calculation of weight to repeat the term and position within documents, and thus it is more suitable for long documents than for short documents. [17]

The following table summarizes the most important differences between the three main retrieval models:

Tab.1:

	Positives	Negatives
Boolean Model	– ***Simple and uncomplicated form and thus easy to apply and investigate.*** – ***Predictable and easy to explain.*** – ***Experts feel more in control of the system.***	– ***The vocabularies of indexes are the same, as the vocabularies of inquiries, it uses the complete matching,*** – ***There is not possible to apply the partial matching.*** – ***The retrieved documents are not arranged or classified.*** – ***There is no weighing of the inquiries and index ' terms*** – ***Difficulty in constructing logic inquiries if they are long.***
Vector Space Model	– ***The simplest model based on linear algebra*** – ***It depends on the weighting of terms.*** – ***It is based on the calculation of the degree of similarity between inquiries and documents.*** – ***Partial matching is possible.***	– ***The terms were assumed statistically independent in theory.*** – ***Long documents are poorly represented and thus have limited expressive ability.*** – ***The keywords must be completely identical to the document terms.*** – ***It lacks the linguistic structure to represent important linguistic features.***

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 10 Issue 09, September-2021

Probabilistic Model	– ***Effective model.*** – ***Mathematical & theoretical model.*** – ***Suitable for long documents.***	– ***Probabilities are difficult to estimate.*** – ***Unrealistic assumptions due to independence of the term.*** – ***Logical relationships are neglected.*** – ***There are many models, and thus it is difficult to determine the best one, because it requires prior knowledge.***

VI. CONCLUSION:

After reviewing the different models of information retrieval systems, we found that the vector space model considered a flexible and clear at the same time, as it represents one of the most widespread models to date, and whose results depend largely on the process of term weighing, but it has the following two main problems: independence of terms and weighting of terms. [6]Consequently, working to overcome these points enables us to find a sophisticated information retrieval mechanism capable of obtaining better results than those achieved by Boolean models without entering into the complexities of the calculations of the probabilistic model.

We suggest working to increase the effectiveness of terms weighting process by defining descriptors of terms in documents in order to overcome the weaknesses of the vector space model. So that those descriptors give quantitative or qualitative indicators that determine the value of the information in them, and its importance to the document in an objective manner through the analysis of the linguistic structure of the terms and then, a value representing the degree of membership of the term in the document based on these descriptors, which leads to more accurate results and thus obtaining an effective information retrieval system. That is not restricted by exact match and simple

as the case of Boolean retrieval systems and depends in its operations on a thoughtful representation of the terms of texts, and not assumptions that may not correspond to reality as the case of the Probabilistic model, which ignores some important descriptors of the terms.

REFERENCES

[1] M. Sanderson and B. Croft, "The History of Information Retrieval Research," *IEEE*, vol. 100, no. Special Centennial Issue, pp. 1444 - 1451, 2012.

[2] C. Mooers, "Zatocoding applied to mechanical organization of knowledge," *Journal of the Association for Information Science and Technology*, vol. 2, no. 1, pp. 20-32, 1951.

[3] A. Göker and J. Davies, "Information Retrieval Models," *Wiley*, pp. 1-19, 2009.

[4] T. Gondaliya and H. Joshi, "Journey of Information Retrieval to Information Retrieval Tools-IR&IRT A Review," in *CALIBER-2017*, CHENNAI, 2017.

[5] D. Mabrouk, S. Rady , N. Badr and . M. , "Modeling using Term Dependencies and Term Weighting in Information Retrieval Systems," *Eighth International Conference on Intelligent Computing and Information Systems (ICICIS)*, vol. 42, pp. 321-328, 2018.

[6] V. Gudivada, D. L.Rao and A. R.Gudivada, "Information Retrieval: Concepts, Models, and Systems," in *Handbook of Statistics*, vol. 38, United States, 2018, pp. 331-401.

[7] A. Singhal, "Modern Information Retrieval: ABrief Overview," *IEEE*, 2001.

[8] G. Salton and M. J. McGill, Introduction to Modern Information Retrieval, 1983: McGraw-Hill Book Company, New York.

[9] H. Luhn, "A statistical approach to mechanized encoding and searching of literary information," *IBM Journal of Research and Development*, vol. 1, pp. 309-3017, 1957.

[10] M. and K. , "Probabilistic Indexing and Information Retrieval," *Journal of the ACM (JACM)*, vol. 7, pp. 216-244, 1960.

[11] E. Gaussier and F. Yvon, Textual Information Access: Statistical Models, WILEY, 2012.

[12] K. Jones and S. Robertson, "Relevance weighting of search terms," *Journal of the American Society for Information Sciences*, vol. 27, no. 3, pp. 129-146, 1976.

[13] Mastering Elasticsearch 5.x, Third ed., UK: Packt Publishing Ltd, 2017.

[14] H. Turtle and W. B. Croft, "Inference networks for document retrieval," *ACM SIGIR Forum*, vol. 51, no. 2, 2017.

[15] T.-Y. Liu, "Learning to Rank for Information Retrieval," *Foundations and Trends in Information Retrieval*, vol. 3, no. 3, p. 225–331, 2009.

[16] T. Sabbah, A. Selamat, M. H. Selamat, F. S. Al-Anzi, E. H. Viedma and O. Krejcar, "Modified Frequency-Based Term Weighting Schemes for Text Classification," *Applied Soft Computin*, vol. 58, pp. 193-206, 2017.

[17] M. Pannu, A. James and R. Bird , "Comparison of Information Retrieval Models," in *The 19th Western Canadian Conference on Computing Education - In-Cooperation with ACM SIGCSE*, 2014.

ABOUT AUTHORS

Manal SHEIKH OGHLI is an Information Technology IT graduated, she is a student in Web Science master's program at Syrian Virtual University (SVU). Her main research topics are the web science.

Muhammad Mazen ALMUSTAFA is an assistance professor at the international university for science & Technology (IUST). He works also as a part-time at Syrian Virtual University (SVU). He has two masters and two PhD degrees. The first PhD degree in Computer Information Systems (CIS). The second PhD Degree in Computers and their networks. His main research topics are the web science. He has published a number of papers on this topic in conferences and journals.