ISSN 2395-1621

A Survey on Information retrieval techniques and their performance measures

[1]Swati Bhat, [2]Neha kale, [3]Mioule Thereza Fernandes, [4]Jason Charles Lawrence, [5]Manjusha Sanke

[1]swatibhat710@gmail.com
[2]nehakale21198@gmail.com
[3]mioule.ferns@gmail.com
[4]charleslawrence2204@gmail.com
[5]manj176@gmail.com

Information Technology, Shree Rayeshwar Institute of engineering & Information Technology Shiroda, India.

ABSTRACT

The availability of tremendous amount of information has necessitated the need for retrieving useful information from a huge collection. Information retrieval systems assist the user to retrieve relevant information from huge corpus depending on user need specified in the form of query.
In this paper, we discuss various information retrieval models and methods for evaluating a model. Also we discuss an application where an IR model can provide a solution.

Keywords: Information retrieval, Tokenization, Cosine Similarity, Vector Space, Boolean, Semantic, Indexing, Scoring, Ranking, Search Engine

ARTICLE INFO

Article History

Received: 8th March 2020

Received in revised form :

8th March 2020

Accepted: 10th March 2020

Published online :

11th March 2020

I. INTRODUCTION

Information Retrieval (IR) is the process by which a collection of data is represented, stored, and searched for the purpose of knowledge discovery as a response to a user request [2]. Almost all applications that handle information on the internet would fail without the support of information retrieval technology. IR model is a base of a search engine which is a vital component of World Wide Web. IR systems perform ranking of documents based on their estimation of the usefulness of a document for a user query. Many IR systems assign a numeric score to every document and perform ranking of documents based on its computed score. The traditional IR models [3] are basically classified as Set-theoretic, Algebraic and Probabilistic model. The Set-theoretic models include Boolean, Case-based reasoning, Fuzzy set and extended Boolean. The Algebraic models include Vector-Space, Generalized Vector-Space, Latent Semantic Indexing and Neural Network. The Probabilistic models include Basic probabilistic, Inference Network and Brief Network. The performance of these models can be evaluated using measures [1] such as precision and recall.

IR system supports the following processes [3]:

- representation of the content of the documents,
- representation of the user's information need
- Comparison of two representations.

A set of documents are indexed in a summarized manner. A user specifies the need for information in the form of query with query formulation process. Both representations are then compared and matched to retrieve relevant information. Based on retrieved documents, a relevance feedback can be provided to reformulate the query for better results.

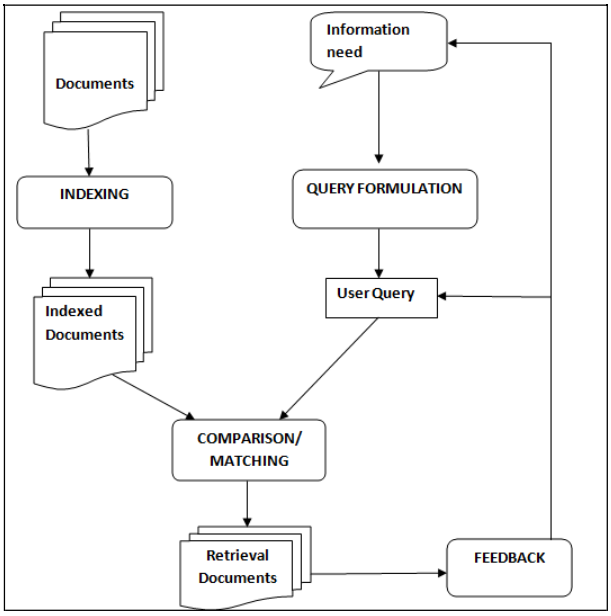

Fig. 1. A general process of Information retrieval

In this paper we discuss each of the IR models and also the performance measures used to evaluate them.

II. IR MODELS

A. Set-Theoretic models

- Boolean

This is the basic model of information retrieval. It deals with using logical functions in the query to retrieve the required data. This model is based on set theory and Boolean algebra which together form a model for determining the data.

Documents that are being searched in the database are sets of terms while Queries, given by the user are Boolean expressions on terms [1]. The terms are combined using AND and OR operators, where AND is intersection or logical product of any term and OR is union or logical sum of any terms.

The approach of Boolean model is as follows: Suppose, Document (D) =Logical conjunction of keywords. Query (Q) =Boolean expression of keywords and record,

$R(D, Q) = D \rightarrow Q$

$D = t1 \wedge t2 \wedge \ldots \wedge tn$

$Q = (t1 \wedge t2) \vee (t3 \wedge \neg t4)$

$D \rightarrow Q$, thus $R(D, Q) = 1$.

This approach was simple and easy to implement. But the biggest drawback of this approach was that there was no concept of ranking and it gives only the exact match [1].

- **Case-based reasoning**

CBR[7] is a simplest model which is build of set of cases and stored in knowledge base known as case bases. Consider cases to be CVs and corpus to be case bases. A case is divided into two parts, one is problem description part and another is problem solution part. The problem description part provides set of CVs to the reasoning model and the problem solution part does processing on CVs and gives output from the reasoning model. This model is also called as problem solving method because the HR composes a search query and at the same time if the problem is encountered then he or she wants to solve it. A problem is solved by retrieving and reusing the same experiences from the corpus i.e. case bases. If the solution to the problem is found then we add CVs i.e. cases to the corpus i.e. case bases for further retrieval and reuse.

CBR consists of four phases:"retrieve", "reuse", "revise", "retain". The "retrieve" phase is used to retrieve CVs from the corpus. The "reuse" phase uses the already existing solutions to solve a new problem. This phase ends by manipulating the IR process or presenting queries to the HR. The "revise" phase is used for building up the corpus i.e. case bases which acts as a backbone of the CBR.A new CV i.e. case is retained, if revise phase requires more than one reuse phase to achieve a satisfactory result. The major drawback of this model is that CBR model fails when a search query does not satisfy the user needs.

- **Fuzzy set**

Fuzzy Set Model is a formal framework used to improve the extensions of Boolean model, concerning both the representation of documents and the query language without redesigning it completely. This model behaves like a set whose elements contain degree of association. Here, the completely belongs or does not belong to a given set. The output of this model is given in terms of true and false where true represents 1 and false represents 0. [6]

According to the extension of Boolean model, each document is represented as a fuzzy set of terms. With respect to the information contained in the document a numeric weight is specified for each term associated with a document.[6] The relation between document and terms is defined as

$$F: D * T \rightarrow [0,1]$$

A document is then represented as fuzzy set of terms, $\{\mu(s)/s\}$, in which $\mu(s) = F(d,s)$ [6]. The fuzzy document representation is based on the definition of a weighted indexing function, in which the weight specifies the index of the search term in the particular documents.

- **Extended Boolean**

The biggest drawback of Boolean Model was that there was no concept of ranking and it gives only the exact match. In order to overcome this problem an Extended Boolean Model was developed. The aim of this model is to implement partial matching i.e. if some of the queried terms are also matched then it gives a relevant document. The model also performs ranking on the similarities between the queries and documents. The major drawback of this model is that it cannot rank the documents in decreasing order of relevance [6].

B. Algebraic models

- **Vector-Space**

The VSM is an algebraic model used for Information Retrieval. The concepts of the vector space model is that it places terms, documents, and queries in a term-document space and computes the similarity score

Working

Step-1: Each document is broken down into a word frequency table.

The tables are called vectors and can be stored as arrays. Step-2: A vocabulary is built from all the words in all documents in the system.

Step-3: The vocabulary needs to be sorted.

Step-4: Each document is represented as a vector based against the vocabulary

Step-5: Queries can be represented as vectors in the same way as documents.

Similarity measures

Documents are compared with the input query and the most similar documents are returned using similarity measure. The most popular similarity measure is the cosine coefficient, which measures the angel between a document vector and query vector.

The cosine measure:

The cosine measure calculates the angle between the vectors. For two vectors d and d' the cosine similarity between d and d' is given by:

(D * D') / | D | * | D' |

Here d X d' is the vector product of d and d', calculated by multiplying corresponding frequencies together.

- **Generalized Vector-Space**

VSM is linearly independent and are orthogonal whereas GVSM is linearly independent, but not pair wise orthogonal [13].

GVSM extend the standard Vector Space Model (VSM) by adding additional types of information, in the representation of documents. Semantic information can boost text retrieval performance with the use of GVSM. It is an alternative to VSM.

GVSM is term to term correlations, where a new space is considered, where each term vector t_i was expressed as a linear combination of 2 n vectors m_r, r = 1..2 n .The similarity measure between a document and a query then became

$$\cos(\overrightarrow{d_k}, \vec{q}) = \frac{\sum_{j=1}^{n} \sum_{i=1}^{n} a'_{ki} q'_j \overrightarrow{t_i} \overrightarrow{t_l}}{\sqrt{\sum_{i=1}^{n} a'_{ki}{}^2 \sum_{j=1}^{n} q'_j{}^2}}$$

Where, the term vectors t_i and t_j need not be known, as long as the correlations between terms t_i and tj are known [19].

- **Latent Semantic Indexing**

LSI [12] retrieves information on the based on meaning of a document. It is a technique in NLP of analyzing relationships between documents and the terms they contain by producing a set of concepts related to the documents and terms. A truncated singular value decomposition (SVD) is used to estimate the structure in word usage across documents.SVD is a least-squares method. LSI is measure for dimensionality reduction. A dimensionality reduction technique takes a set of objects which are in high-dimensional space and represents in a low dimensional space often could be in 2D or 3D space.

LSI tries to capture hidden structure using techniques from linear algebra.

Advantage of LSI is new dimensions are a better representation of documents and queries. LSI analysis recovers the original semantic structure of the space and its original dimensions.

Disadvantage of LSI is Efficiency, if the query has more terms than its representation in the LSI vector space, then inner product similarity scores will take more time to compute in term space, increases the search time.

- **Neural Network**

The Neural Network is *feed-forward* networks. In IR, the output of the entire network is often either a vector representation of the input or some predicted scores [9]. Information Retrieval of Neural Network use deep neural network to rank search result in form of query. Neural models [10] learn representations of language from raw text that can bridge the gap between query and document vocabulary. Neural network ranking have adopted for textual IR applications includes ad-hoc retrieval, question answering etc.

C. Probabilistic models

- **Basic probabilistic**

The main feature of probabilistic model is that it ranks the documents depending upon how relevant the document is to the given query. The probabilistic ranking principle states that "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 that data" [7]. According to the probabilistic ranking principle the documents in the collection are ranked in the order of decreasing probability based on the query entered by the user. The accuracy of the probabilities estimated solely depends upon the information provided to the system [7].

The probabilistic model works as follows: Firstly the user enters a query and a set of documents is then returned by the search engine, from this set the user marks some documents as relevant and some as irrelevant, this provides the user with a preliminary probabilistic description of the ideal answer set. Based on this description the search engine computes new relevance results. By repeating this process many times, the system

comes closer to the real description of the ideal answer set. This is how the processing of the query is done. The probabilistic model makes use of the formula O(R)=P(R)/1-P(R), where R means document is relevant.

The main limitation of this model is that initially the document is in the ideal set or irrelevant. It becomes difficult to compute the probability that the retrieved documents are relevant. This model is very complex since neither the weights nor the ideal set is defined.

- **Inference Network**

In this model the query entered by the user is considered as an assertion, and only those documents are retrieved for

which the assertion is true. In this model a document creates an instance of a term by assigning a certain weight and then depending upon the credit from multiple terms the document score is computed [3].

An inference network is an acyclic dependency graph. In the graph index term variables, query variables and document variables are represented as nodes in the network. The dependence relation between document node and its term nodes is represented by an edge between them. This indicates that observation of the document yields improved belief in its term nodes. Query variables mean that information specified by the user in the form of query has been met. In inference network model the inference network consists of two networks:

1. Document network-The document network consists of all the documents and schemes. Once a document network is built its structure does not change during query processing [5].
2. Query network-it consists of a single node which represents the information needed by the user in the form of a query. The structure of this network may change during query processing as the user is allowed to add more queries to the network to make the search efficient [5].

- **Brief Network**

Brief Network Model [14] uses a clearly defined sample space and therefore it separates the document and the query portions of the network.

Sample space contains indexed terms and each document is indexed by the index term. Each index term is a document referred to as concept and the whole set is called concept space. A user query which contains the terms used to index the query in the set is represented as a concept in the set. Similarly a document which contains terms used to index the document can be represented as a concept in the concept space. In this model the query entered by the user is considered as network node which is associated to the document represented as a binary variable.

III. PERFORMANCE MEASURES

Two main evaluation metrics used in retrieval system are recall and precision.

Table 1. Contigency Matrix

	Relevant documents	**Irrelevant documents**
Documents retrieved	A True positive	C False positive
Documents not retrieved	B False negative	D True negative

variable A and D is the number of correct results and variable B and C indicates the number of incorrect result

Recall

Recall is the ratio of the number of relevant records retrieved to the total number of relevant records in the database. Drawback is that, the total number of relevant items in a database cannot be measured.

$$Recall=(D/C+D)*100\%$$

D is the number of irrelevant records not retrieved and C is the number of irrelevant records retrieved.

IV. IR MODEL BASED APPLICATION

As an application of IR based model, consider the following scenario:-

Recruiting and assigning right candidate for right job consumes significant time and efforts in an organization as the HR has to perform manual search to retrieve relevant CVs. So, solution to this problem is a CV retrieval system based on job description. CVs in text document or pdf form can be accepted. System will apply an IR model and retrieve the CVs that match the job description entered by the HR in the form of query.

V. CONCLUSION

In this paper, we have classified and reviewed the various IR models. The merits and demerits of each model are analyzed. Also, we have discussed an application which can make use of IR technique and make the task more efficient.

VI. ACKNOWLEDGEMENT

This survey paper owes its existence and certainly its quality to a number of people. Hence we take this opportunity to thank our teachers, friends and all those who provided us with their constant support and guidance because of which we could progress further.

REFRENCES

[1] R. Baeza, and B. Ribeiro (2011). Modern InformationRetrieval: Second edition. Addison-Wesley, New York, NY,USA.

[2] Ndengabaganizi Tonny James, Rajkumar Kannan, “A Survey on Information Retrieval Models, Techniques and Applications”, International Journals of Advanced Research in Computer Science and Software Engineering ISSN: 2277-128X (Volume-7, Issue-7), July 2017.

[3] Hai Dong, Farookh Khadeer Hussain, Elizabeth Chang, “A Survey in Traditional Information Retrieval Models”, 2nd International Conference on Digital Ecosystems & Technology, Feb 2008, pp. 397-

402.Thailand:IEEE.http://dx.doi.org/10.1109/DEST.2008.4635214.

[4] Akram Roshdi and Akram Roohparvar,"Review: Information Retrieval Techniques and Applications ".International Journal of Computer Networks and Communications Security Vol 3, No. 9, Sept. 2015, 373–377

[5] Howard R. Turtle And W. Bruce Croft, "A Comparison of Text Retrieval Models", West Publishing Company, St Paul, MN 55164, USA

[6] Mang'are Fridah Nyamisa, Waweru Mwangi, Wilson Cheruiyot, "A Survey of Information Retrieval Techniques", Published: November 28, 2017

[7] Norbert Gronau and Frank Laskowski , "Using Case-Based Reasoning to improve Information Retrieval in Knowledge Management Systems".

[8] Vikram Singh, Balwinder Saini, "An Effective Pre-Processing Algorithm for Information Retrieval Systems", January 2015 DOI: 10.5121/ijdms.2014.6602

[9] K. D. Onal, Y. Zhang, I. S. Altingovde, M. M. Rahman, P. Karagoz, 1010 A. Braylan, B. Dang, H.-L. Chang, H. Kim, Q. Mcnamara, A. Angert, E. Banner, V. Khetan, T. Mcdonnell, A. T. Nguyen, D. Xu, B. C. Wallace, M. Rijke, M. Lease, "Neural information retrieval: At the end of the early years", Inf. Retr. 21 (2-3) (2018) 111–182.

[10] BhaskarMitra, Nick Craswell, "An Introduction to Neural Information Retrieval", Volume 13, Issue 1, 2019

[11] Mang'areFridah Nyamisa, WaweruMwangi, Wilson Cheruiyot, "A Survey of Information Retrieval Techniques", November 28, 2017

[12] Ashwini Deshmukh ,Gayatri Hegde "A Literature Survey on Latent Semantic Indexing" Volume 1, Issue 4 (September 2012)

[13] George Tsatsaronis and Vicky Panagiotopoulou, "A Generalized Vector Space Model for Text Retrieval Based on Semantic Relatedness" .

[14] B. A. Ribeiro-Neto and R. Muntz, "A brief network model for IR," in 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Zurich, 1996, pp. 235-260.

