IJCSE International Journal of Computer Sciences and Engineering Open Access
Survey Paper Vol.-7, Special Issue, 4, Feb 2019 E-ISSN: 2347-2693

A Survey on Information Retrieval Models in Document Mining

R. Meera

Department of Computer Science, Idhaya College for Women, Kumbakonam, India

Corresponding Author: meera2992@gmail.com

Available online at: www.ijcseonline.org

***Abstract*—** Information retrieval is the process of retrieving relevant documents for the given query over a large document collection. As the technology emergence of digital library and electronic information exchange there is a clear need for organizing and accessing the large quantity of information. Information retrieval focus on the study of storing, organizing and retrieving the information from this large collection. This paper focuses on the types of information retrieval, different fundamental retrieval models and also gives brief overview on document processing.

***Keywords*—** Boolean Model, Information Retrieval(IR),Information Retrieval System (IRS), Indexing, Vector SpaceModel (VSM).

I. INTRODUCTION

Information retrieval is defined as finding information of an unstructured nature that satisfies an user information need from within large collection. IR is the process of storing, organizing and retrieving the information for the user need (query) from large document collection [1][2]. IR is sub field of computer science which attracts many researchers to research on methods to represent, store, and access of information [3]. This process initiates with representing the documents and ends with retrieving the relevant document. The intermediate stages include indexing, filtering, searching, matching and ranking the documents. The main goal of information retrieval system (IRS) is to retrieve all the documents which are relevant to a user query, while retrieving as less non-relevant documents as possible. To pursue this goal, the IRS system has to implement the following process.

Indexing the documents: Indexing is the process used to speed up access to desired information. Searching over large document collection is time consuming, so if there is an index for the documents then searching becomes easy as index files are typically smaller than the original documents.
Filtering the documents: In this process, normally unwanted or common words in the documents are removed.Searching and ranking: This process is the core of IRS; there are various techniques available for retrieving documents that match with users need.

There are two types of information retrieval: one is Adhoc retrieval and the other one is routing or filtering retrieval. In Adhoc information retrieval, the documents are relatively static and also indexed before to any user query. Once the query is issued, the documents which are relevant to the query are ranked based the similarity to the query and presented to the user [4]. Where as in document routing or filtering, the queries are static and the document collection constantly changes [4].There are two main measures to measure the quality of information [4].

Precision: It is the ratio of number of relevant documents retrieved to the total number retrieved.
Recall: It is the ratio of number of relevant documents retrieved to the total number of documents in collection that are beloved to be relevant.

The figure 1 shows the architecture of an IRS. IRS mainly has three processes, first one is document processing, in which the documents are indexed. This is an offline process that is the user of an IRS does not involve in this process. Second is query processing, in this process the user need will be represented as query. Third one is returning ranked documents to the user according to query matched against the documents collection.

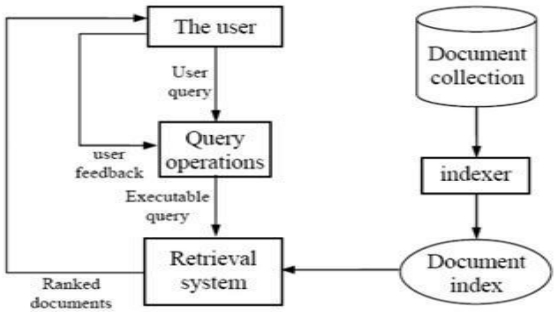

Figure 1: Information retrieval system architecture.

II. IR MODELS

An IR models or retrieval strategies gives a measure of similarity between a query and document. Ranking algorithms are the core of information retrieval system. These algorithms are used to predict which documents are relevant and which are not.

In classic IR model each document is described by a set of representative keywords called index terms. These types of models assign numerical values to distinct relevance between index terms and the query. There are three models available in classic or fundamental IR model they are Boolean model, Vector Space model and Probabilistic model. This paper focuses on Boolean and Vector Space model.

A. Boolean Model:

Boolean model of information retrieval is a classical information retrieval model, which is the first and most adopted one. It is used in many commercial IR systems. Boolean model is based on Boolean algebra involved with And, Or, Not operations to form query. It retrieves the relevant document to the query (exact match). It does not provide any ranking to the documents [6], where other retrieval models provide ranking to the documents. Therefore it is easy and efficient to retrieve the documents.

This model uses tokenization, linguistic model and inverted index in the retrieval process, these concepts will be explained in the next section [5].

Boolean model build a matrix of all the terms in query against all documents. For each term, the Boolean value is assigned to represent whether the term present in the document or not. If term present, value 1 is assigned else zero is assigned.

For example: consider the following documents and query, using Boolean model retrieve the relevant document. Table I shows the terms in these documents and their corresponding Boolean index. Table II shows the documents and its inverted index for the query. Document 1: Anne really loves food.

Document 2: Bob loves to eat pizza.

Document 3: Pizza is a food.

Query: loves and food not pizza

Table 1 Document terms with their Boolean index.

Terms	Document 1	Document 2	Document 3
Anne	1	0	0
Really	1	0	0
Loves	1	1	0
Food	1	0	1
Bob	0	1	0
Eat	0	1	0
Pizza	0	1	1

Table 2 Boolean index for the given query

Terms	Document 1	Document 2	Document 3	Query
Loves	1	1	0	1
Food	1	0	1	1
Pizza	0	1	1	1

By using the above tables Boolean model retrieves the document1as answer for the given query.

The advantages of Boolean retrieval model are,

- Query formulation is easy as quires are expressed as Boolean expressions.
- Retrieves the documents that are exactly matched with query.

The Disadvantages are,

- It is not simple to translate an user information need into a Boolean expression.
- Exact matching may lead to retrieval of too few or too many documents.

B. C. Vector Space model:

Vector space model was suggested by Salton and his colleagues in 1975. VSM computes the measure of similarity by defining vector that represents each document and a query [4]. Document meaning is conveyed through its words. If the words in documents are represented as vectors, then it's easy to compare the documents and query to determine how similar their contents are. Vector space model is characterized by its attempt to rank documents by finding the similarity between the query and each document [4].

In VSM, the value or weigh of each term in document is a non Boolean positive value computed based on inverse document frequency (IDF). To construct a vector for each document, consider the following definitions:

t=number of distinct terms in the document collection.

tf_{ij}=Term frequency, number of occurrence of the term t_j in document d_i.

df_j=document frequency, number of documents which contain t_j.

$idf_j = \log(d/df_j)$ this is the inverse document frequency.

Where d is the total number of documents.

The value of j^{th} entry in the vector corresponding to the document i is calculated using the following equation.

$$D_{ij}=tf_{ij}\times idf_j$$

The similarity between query and document is computed using the following relation.

For example: consider the following documents and the query. Let us see how VSM is used to rank the documents. Q: green blue crocodile.

D1: green corridor with roses.

D2: blue whale and green crocodile.

D3: green corridor and blue sky.

In this collection, there are three documents, so d = 3. If

a term appears in only one of the three documents, its idf is log (d/df)j = log(3/1) = 0.477. Similarly, if a term appears in two of the three documents its idf is log(3/2)= 0.176, and a term which appears in all three documents then its idf is log(3/3) = 0. The idf for the terms in the three documents is given below:

idf_{green}= 0
$idf_{corridor}$= 0.176
idf_{with} = 0.477
idf_{roses} = 0.477
idf_{blue} = 0.176
idf_{whale} = 0.477
idf_{and} = 0.176
$idf_{crocodile}$ = 0.477
idf_{sky}=0.477

Since nine terms appear in the document collection, a nine-dimensional document vector can be constructed. Table 3 shows the vectors corresponding to the documents and the query.

Table 3:Document vectors

docid	And	blue	corridor	crocodile	GrEeN	roses	SkY	whale	WiTh
D1	0	0	0.176	0	0	0.477	0	0	0.477
D2	0.176	0.176	0	0.477	0	0	0	0.477	0
D3	0.176	0.176	0.176	0	0	0	0.477	0	0
Q	0	0.176	0	0.477	0	0	0	0	0

SC(Q,D1)=(0)(0)+ (0.176)(0)+ (0)(0.176)+ (0.477)(0)+ (0)(0)+ (0)(0.477)+ (0)(0)+ (0)(0)+ (0)(0.477)

SC(Q,D1)= 0

SC(Q,D2)= (0.176)(0)+ (0.176)(0.176)+ (0)(0)+ (0.477)(0.477)+ (0)(0)+ (0)(0)+ (0)(0)+ (0)(0.477)+ (0)(0)

SC (Q,D2)=0.03+0.22=0.25

SC(Q,D3)= (0)(0.176)+ (0.176)(0.176)+ (0)(0.176)+ (0)(0.477)+ (0)(0)+ (0)(0)+ (0.477)(0)+ (0)(0)+ (0)(0)

SC (Q, D3) = 0.03

Therefore the ranking of documents would be D2, D3 and D1.

The advantages of vector space model are,

- Its term-weighting scheme improves retrieval performance
- Its partial matching strategy allows retrieval of documents that approximate the query conditions The disadvantages are,
- The assumption of mutual independence between index terms.

III. DOCUMENT PROCESSING

While explaining Boolean model in the previous section, we mentioned that Boolean model involves tokenization, linguistic model and inverted index in the retrieval process. The tokenization, linguistic model are used to process the documents [5]. This document processing can be used to define the basic unit of a document and the character sequences that it comprises are determined. Document processing is also called as process of determining the vocabulary of terms.

Tokenization: it is process of chopping document unit in to number of chunks called as tokens. At time of tokenization, the punctuations within the document unit are removed. These tokens are refereed as terms or words in particular document. These terms are included in the IRS dictionary.

Linguistic Model: In this model the terms related to the particular natural language is processed by removing the common words, for example removing stop words in English language. Also it makes the case folding, that from upper case to lower case or vice versa.

Normalization: It is the process of equivalence classing of terms. After dividing the documents as tokens, if the token of the query matches tokens of the document then it is easy to compare. But if the two character sequences are not same but we would like a match should occur. For example if we search for MCA, we might hope to also match documents containing M.C.A. Therefore term normalization is the process of standardization of tokens so that match occurs even though there is differences in the character sequence of the tokens.

Stemming: Stemming removes word suffixes, perhaps recursively in layer after layer of processing. Stemming reduces the number of unique words, which in turn reduces the storage space required. Hence stemming speeds up the search process. Stemming improves recall by reducing all forms of the word to a base or stemmed form. For example, if a user asks for analysis, they may also want documents whichcontain ,analyze, analyzing, analyzer, analyzes, and analyzed. Therefore, the document processor stems document terms to analy- so that documents which include various forms of analy- will have equal likelihood of being retrieved.

IV. INDEXING TECHNIQUES

Indexing is process, which is used to speed up access to the desired data. An index file consists of records called index entries of the formSearch Key Pointer

There are several popular information retrieval indexing techniques available, which includes order indexing, Hashed indexing and inverted indexing.

- Ordered indexing: In this technique, search keys are usually sorted. Therefore the searching over document using index terms is also in order. This technique efficiently implemented using B tree of order m and B^+ trees [7].
- Hash based indexing: In this technique, search keys are distributed uniformly across the hash table using hash function. The index file is much smaller than the original file and can be searched much faster [7].
- Inversion Indices: Each document can be represented by a list of keywords which describe the contents of the document for retrieval purposes [8]. Fast retrieval can be achieved if we invert on those keywords. The keywords are stored, eg alphabetically; in the index file for each keyword we maintain a list of pointers to the qualifying documents in the postings file. This method is followed by almost all the commercial systems.

V. CONCLUSION

Lastly I would like conclude that, information retrieval is the process of searching and retrieving the information from the document collection. This REVIEW paper dealt with basics of IR systems, retrieval models and indexing.

Information retrieval systems are widely used in organizing and accessing the information that are stored in large document collection. For example, retrieving the information from the Medical databases such as X-rays, CT and MRI scans, Criminal investigation suspects, fingerprints, Personal archives, text and colour images. Retrieving the information from Scientific Databases such as sensor data, weather, geological, environmental data, Office automation, electronic books etc.

REFERENCES

[1] K. Aas and L. Eikvil, "Text Categorisation: A Survey," Technical Report Raport NR 941, Norwegian Computing Center, 1999.

[2] R. Agrawal and R. Srikant, "Fast Algorithms for Mining Association Rules in Large Databases," Proc. 20th Int'l Conf. Very Large Data Bases (VLDB '94), pp. 478-499, 1994.

[3] H. Ahonen, O. Heinonen, M. Klemettinen, and A.I. Verkamo, "Applying Data Mining Techniques for Descriptive Phrase Extraction in Digital Document Collections," Proc. IEEE Int'l Forum on Research and Technology Advances in Digital Libraries (ADL '98), pp. 2-11, 1998.

[4] R.Baeza-Yates and B.Ribeiro-Neto, Modern Information Retrieval. Addison Wesley, 1999.

[5] N. Cancedda, N. Cesa-Bianchi, A. Conconi, and C. Gentile, "Kernel Methods for Document Filtering," TREC, trec.nist.gov/ pubs/trec11/papers/kermit.ps.gz, 2002.

[6] N. Cancedda, E. Gaussier, C. Goutte, and J.-M. Renders, "Word-Sequence Kernels," J. Machine Learning Research, vol. 3, pp. 1059-1082, 2003.

[7] M.F. Caropreso, S. Matwin, and F. Sebastiani, "Statistical Phrases in Automated Text Categorization," Technical Report IEI-B4-07-2000, Instituto di Elaborazionedell'Informazione, 2000.

[8] C. Cortes and V. Vapnik, "Support-Vector Networks," Machine Learning, vol. 20, no. 3, pp. 273-297, 1995.

[9] S.T. Dumais, "Improving the Retrieval of Information from External Sources," Behavior Research Methods, Instruments, and Computers, vol. 23, no. 2, pp. 229-236, 1991.

[10] J. Han and K.C.-C. Chang, "Data Mining for Web Intelligence," Computer, vol. 35, no. 11, pp. 64-70, Nov. 2002.

[11] J. Han, J. Pei, and Y. Yin, "Mining Frequent Patterns without Candidate Generation," Proc. ACM SIGMOD Int'l Conf. Management of Data (SIGMOD '00), pp. 1-12, 2000.

[12] Y. Huang and S. Lin, "Mining Sequential Patterns Using Graph Search Techniques," Proc. 27th Ann. Int'l Computer Software and Applications Conf., pp. 4-9, 2003.

[13] N. Jindal and B. Liu, "Identifying Comparative Sentences in Text Documents," Proc. 29th Ann. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval (SIGIR '06), pp. 244-251, 2006.

[14] T. Joachims, "A Probabilistic Analysis of the Rocchio Algorithm with tfidf for Text Categorization," Proc. 14th Int'l Conf. Machine Learning (ICML '97), pp. 143-151, 1997.

[15] T. Joachims, "Text Categorization with Support Vector Machines: Learning with Many Relevant Features," Proc. European Conf. Machine Learning (ICML '98),, pp. 137-142, 1998.

[16] T. Joachims, "Transductive Inference for Text Classification Using Support Vector Machines," Proc. 16th Int'l Conf. Machine Learning (ICML '99), pp. 200-209, 1999.

[17] W. Lam, M.E. Ruiz, and P. Srinivasan, "Automatic Text Categorization and Its Application to Text Retrieval," IEEE Trans. Knowledge and Data Eng., vol. 11, no. 6, pp. 865-879, Nov./Dec. 1999.42 IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 24, NO. 1, JANUARY 2012

[18] D.D. Lewis, "An Evaluation of Phrasal and Clustered Representations on a Text Categorization Task," Proc. 15th Ann. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval (SIGIR '92), pp. 37-50, 1992.

[19] D.D. Lewis, "Feature Selection and Feature Extraction for Text Categorization," Proc. Workshop Speech and Natural Language, pp. 212-217, 1992.

[20] D.D. Lewis, "Evaluating and Optimizing Automous Text Classification Systems," Proc. 18th Ann. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval (SIGIR '95), pp. 246-254,1995.

