International Journal of Computer Networks and Communications Security

VOL. 3, NO. 9, SEPTEMBER 2015, 373–377
Available online at: www.ijcncs.org
E-ISSN 2308-9830 (Online) / ISSN 2410-0595 (Print)

Review: Information Retrieval Techniques and Applications

Akram Roshdi[1] and Akram Roohparvar[2]

[1] Department of Engineering, Khoy branch, Islamic Azad University, Khoy, IRAN

[2] Department of Engineering, Qom branch, Islamic Azad University, Qom, IRAN

E-mail: [1]akram.roshdi@gmail.com, [2]akram.roohparvar@gmail.com

ABSTRACT

For thousands of years people have realized the importance of archiving and finding information. With the advent of computers, it became possible to store large amounts of information; and finding useful information from such collections became a necessity. The field of Information Retrieval (IR) was born in the 1950s out of this necessity. Over the last forty years, the field has matured considerably. Several IR systems are used on an everyday basis by a wide variety of users. Information retrieval is become a important research area in the field of computer science. Information retrieval (IR) is generally concerned with the searching and retrieving of knowledge-based information from database. In this paper, we represent the various models and techniques for information retrieval. In this Review paper we are describing different indexing methods for reducing search space and different searching techniques for retrieving a information. We are also providing the overview of traditional IR models.

Keywords: *Information Retrieval (IR), Indexing, IR mode,Searching, Vector Space Model (VSM).*

1 INTRODUCTION

Information retrieval is generally considered as a subfield of computer science that deals with the representation, storage, and access of information [1]. Information retrieval is concerned with the organization and retrieval of information from large database collections [2]. 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 (query) [3].this process involves various stages initiate with representing data and ending with returning relevant information to the user. Intermediate stage includes filtering, searching, matching and ranking operations. The main goal of information retrieval system (IRS) is to "finding relevant information or a document that satisfies user information needs". To achieve this goal, IRSs usually implement following processes:

1) In indexing process the documents are represented in summarized content form.
2) In filtering process all the stop words and common words are remove.
3) Searching is the core process of IRS. There are various techniques for retrieving documents that match with users need.

There are two basic measures for assessing the quality of information retrieval [2].

Precision*:* This is the percentage of retrieved documents that are in fact relevant to the query.

Recall: This is the percentage of documents that are relevant to the query and were in fact retrieved.

There are three basic processes an information retrieval system has to support: the representation of the content of the documents, the representation of the user's information need, and the comparison of the two representations. The processes are visualized in Figure 1. In the figure, squared boxes represent data and rounded boxes represent processes.

Representing the documents is usually called the indexing process. The process takes place off-line, that is, the end user of the information retrieval system is not directly involved. The indexing process results in a representation of the document [5].

Users do not search just for fun, they have a need for information. The process of representing their

information need is often referred to as the query formulation process.
The resulting representation is the query [5].

Comparing the two representations is known as the matching process. Retrieval of documents is the result of this process.

The structure of this paper is as follows. A brief introduction of IR models is presented in Section II, followed by indexing method in section III. Followed by searching techniques in Section IV. Followed by IR applications in section V, Finally, Section VI covers conclusions.

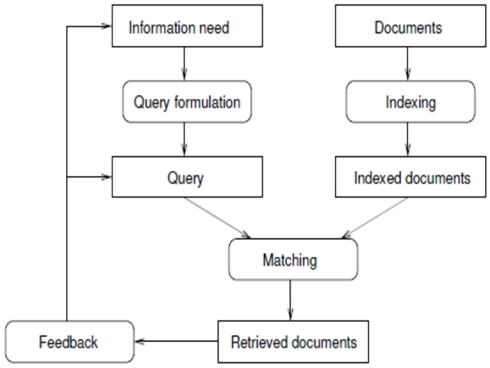

Fig 1. Information retrieval processes

2 IR MODELS

An IR model specifies the details of the document representation, the query representation and the retrieval functionality [3].
The fundamental IR models can be classified into Boolean, vector, probabilistic and inference network model [8] [3]. The rest of this section briefly describes these models.

2.1 Boolean Model

The Boolean model is the _rst model of information retrieval and probably also the most criticised model. The Boolean model is the _rst model of information retrieval and probably also the most criticised model. The model can be explained by thinking of a query term as a unambiguous de_nition of a set of documents. For instance, the query term economic simply de_nes the set of all documents that are indexed with the term economic. Using the operators of George Boole's mathematical logic, query terms and their corresponding sets of documents can be combined to form new sets of documents. The Boolean model allows for the use of operators of Boolean algebra, AND, OR and NOT, for query formulation, but has one major disadvantage: a Boolean system is not able to rank the returned list of documents [4]. In the Boolean model, a document is associated with a set of keywords. Queries are also expressions of keywords separated by AND, OR, or NOT/BUT. The retrieval function in this model treats a document as either relevant or irrelevant [3]. In Figure 2, the retrieved sets are visualised by the shaded areas.

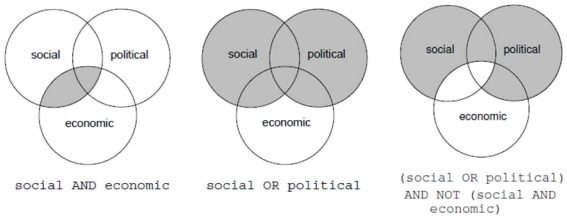

Fig 2.Boolean combinations of sets visualised as Venn diagrams

2.2 Vector Space Model

Gerard Salton and his colleagues suggested a model based on Luhn's similarity criterion that has a stronger theoretical motivation (Salton and McGill 1983). They considered the index representations and the query as vectors embedded in a high dimensional Euclidean space, where each term is assigned a separate dimension. The vector space model can best be characterized by its attempt to rank documents by the similarity between the query and each document [10].In the Vector Space Model(VSM), documents and query are represent as a Vector and the angle between the two vectors are computed using the similarity cosine function. Similarity Cosine function can be defined as:

Where,

$$sim(d_j, q) = \frac{d_j \cdot q}{\|d_j\| \cdot \|q\|} = \frac{\sum_{i=1}^{N} w_{i,j} w_{i,q}}{\sqrt{\sum_{i=1}^{N} w^2{}_{i,j}} \sqrt{\sum_{i=1}^{N} w^2{}_{i,q}}} \quad (1)$$

Documents and queries are represented as vectors.

$$d_j = (w_{1,j}, w_{2,j}, \ldots, w_{t,j})$$

$$q = (w_{1,q}, w_{2,q}, \ldots, w_{t,q})$$

Vector Space Model have been introduce term weight scheme known as if-idf weighting. These weights have a term frequency (tf) factor measuring the frequency of occurrence of the terms in the document or query texts and an inverse document frequency (idf) factor measuring the inverse of the number of documents that contain a query or document term [4].

2.3 Probabilistic Model

Whereas Maron and Kuhns introduced ranking by the probability of relevance, it was Stephen Robertson who turned the idea into a principle. He formulated the probability ranking principle, which he attributed to William Cooper, as follows (Robertson 1977). The most important characteristic of the probabilistic model is its attempt to rank documents by their probability of relevance given a query [9]. Documents and queries are represented by binary vectors ~d and ~q, each vector element indicating whether a document attribute or term occurs in the document or query, or not. Instead of probabilities, the probabilistic model uses odds O(R), where O(R) = P(R)/1 − P(R), R means "document is relevant" and ¯R means "document is not relevant" [4].

2.4 Inference Network Model

In this model, document retrieval is modeled as an inference process in an inference network. [11] Most techniques used by IR systems can be implemented under this model. In the simplest implementation of this model, a document instantiates a term with a certain strength, and the credit from multiple terms is accumulated given a query to compute the equivalent of a numeric score for the document. From an operational perspective, the strength of instantiation of a term for a document can be considered as the *weight* of the term in the document, and document ranking in the simplest form of this model becomes similar to ranking in the vector space model and the probabilistic models described above. The strength of instantiation of a term for a document is not defined by the model, and any formulation can be used.

3 INDEXING TECHNIQUES

There are several popular information retrieval indexing techniques, including inverted indices and signature files.

3.1 Signature File

In signature file method each document yields a bit string („signature") using hashing on its words and superimposed coding. The resulting document signatures are stored sequentially in a separate file called signature file, which is much smaller than the original file, and can be searched much faster [6].

3.2 Inversion Indices

Each document can be represented by a list of keywords which describe the contents of the document for retrieval purposes [6]. 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 [10].

4 SEARCHING TECHNIQUES

There are various searching algorithms, including linear search, binary search, brute force search etc. some general searching algorithms are described below:

1) In linear search algorithm is a method of finding a particular element or keyword from list or array that checks every element in list, one at a time and in sequence. Linear search is a simplest search algorithm. One of the most important drawbacks of linear search is slow searching speed in ordered list. This search is also known as sequential search.

2) Brute force search is a very general problem-solving technique that consists of systematically enumerating all possible candidates for the solution and checking whether each candidate satisfies the problem's statement. Brute force algorithm is simple to implement and it will always find a solution if it exist.

3) Binary search algorithm, finds specified position of the element by using the key value with in a sorted array. In each step, the algorithm compares the search key value with the key value of the middle element of the array. If the keys match, then a matching element has been found and its index, or position, is returned. Otherwise, if the search key is less than the middle element's key, then the algorithm repeats its action on the sub-array to the left of the middle element or, if the search key is greater, on the sub-array to the right.

If the remaining array to be searched is empty, then the key cannot be found in the array and a special "not found" indication is returned.

5 AREA OF IR APPLICATION

Information retrieval (IR) systems were firstly

developed to help manage the huge amount of information. Many universities, corporate, and public Libraries now use IR systems to provide access to books, journals, and other documents. Information retrieval is used today in many applications [7]. General applications of information retrieval system are as follows:

5.1 Digital Library

A digital library is a library in which collections are stored in digital formats and accessible by computers. The digital content may be stored locally, or accessed remotely via computer networks. A digital library is a type of information retrieval system [7].

5.2 Search Engines

A search engine is one of the most the practical applications of information retrieval techniques to large scale text collections. Web search engines are best- known examples, but many others searches exist, like: Desktop search, Enterprise search, Federated search, Mobile search, and Social search [7].

5.3 Media Search

An image retrieval system is a computer system for browsing, searching and retrieving images from a large database of digital images [7].

6 CONCLUSION

At last we conclude that, information retrieval is a process of searching and retrieving the knowledge based information from collection of documents. This REVIEW has dealt with the basics of the information retrieval. In first section we are defining the information retrieval system with their basic measurements. After this we concerns with traditional IR models and also discuss about the different indexing techniques and searching techniques. This paper also includes the area of IR applications.

7 REFERENCES

[1] M.François Sy, S.Ranwez, J.Montmain,"User centered and ontology based information Retrieval system for life sciences", BMC Bioinformatics,2105.

[2] R. Sagayam, S.Srinivasan, S. Roshni, "A Survey of Text Mining: Retrieval, Extraction and Indexing Techniques", IJCER, sep 2012, Vol. 2 Issue. 5, , PP: 1443-1444,.

[3] Anwar A. Alhenshiri, "Web Information Retrieval and Search Engines Techniques",2010,Al- Satil journal,PP: 55-92.

[4] D.Hiemstra,P. de Vries, "Relating the new language models of information retrieval to the traditional retrieval models", published as CTIT technical report TR-CTIT-00-09, May 2000.

[5] Djoerd Hiemstra, "Information Retrieval Models", published in Goker, A., and Davies, J. Information Retrieval: Searching in the 21st Century. John Wiley and Sons, November 2009,Ltd., ISBN-13: 978-0470027622.

[6] Christos Faloutsos, Douglas W. Oard, "A Survey of Information Retrieval and Filtering Methods", CS-TR-3514, Aug 1995. "Algorithms for Information Retrieval – Introduction", Lab module 1.

[7] R. Baeza-Yates and B. Ribeiro-Neto, "Modern Information Retrieval",2009, ACM Press, ISBN: 0-201-39829-X.

[8] S.E. Robertson and K. Sparck Jones. "Relevance weighting of search terms. Journal of the American Society for Information Science", 1976, 27:129–146.

[9] G. Salton and M.J. McGill, "editors. Introduction to Modern Information Retrieval". McGraw-Hill ,1983.

[10] H. Turtle, "Inference Networks for Document Retrieval". Ph.D. thesis, Department of Computer Science,University of Massachusetts, Amherst, MA 01003. Available as COINS Technical Report 90-92, 1990.

[11] C. J. van Rijsbergen. "Information Retrieval. Butterworths", London,1979.

[12] T. Strzalkowski, L. Guthrie, J. Karlgren, J. and et. "Natural language information retrieval: TREC-5 report". In Proceedings of the Fifth Text REtrieval Conference (TREC-5), 1997.

[13] Gerard Salton and M. J. McGill. "Introduction to Modern Information Retrieval". McGraw Hill Book Co.,New York, 1983.

[14] Gerard Salton and Chris Buckley. "Term-weighting approaches in automatic text retrieval". Information Processing and Management, , 1988, 24(5):513–523.

[15] Gerard Salton, editor. "The SMART Retrieval System—Experiments in Automatic Document Retrieval".Prentice Hall Inc., Englewood Cliffs, NJ, 1971.

[16] N. J. Belkin and W. B. Croft." Information filtering and information retrieval: Two sides of the same coin? ",Communications of the ACM, 1992,35(12):29–38.

[17] G.Grefenstette,editor.”Cross-Language Information Retrieval”. Kluwer Academic Publishers, 1998.
[18] J. L. Fagan. “The effectiveness of a nonsyntactic approach to automatic phrase indexing for document
[19] Retrieval”. Journal of the American Society for Information Science, 1989, 40(2):115–139.
[20] K. Sparck Jones. “A statistical interpretation of term specificity and its application in retrieval”. Journal of Documentation, , 1972,28:11–21.