A Survey on Information Retrieval Techniques and Applications

Syed Abolhasan Nezam Doost,
Islamic Science & Culture Academy, Qom, Iran

Abstract—*Information retrieval process deals with representation, storage and search of a collection of data. Data can be in various forms e.g. text files, image files, speech files, music files. Research in this area and the corresponding tools are quite mature. However, with the advancement of technology new challenges are coming up. A continuous effort of the researchers and product developers are needed to handle these challenges.*

Keywords—*information; retrieval; relevance; ranking; probability; fuzzy coefficient; artificial neural network.*

I. INTRODUCTION

Information Retrieval (IR) is the process by which a collection of data is represented, stored and searched as a response to a user request or query [1]. The goal of IR research is to develop models and algorithms for retrieving information from document repositories, in particular, textual information [2]. However, information can be of any type: textual, visual or auditory.

The standard practice for IR is ad-hoc. Here user puts a query and the matched information is retrieved. Matching can be exact based on Boolean logic. As a better alternative, matching can be ranked. Presently, ranking is done by statistical analysis. However, author opines that application of fuzzy logic might provide a better ranking system.

Following figure explains the basic steps of IR process. In the figure, squared boxes represent data and rounded boxes represent processes. There are three basic processes in IR: (i) the representation of the content of the documents, (ii) the representation of the user's information need, and (iii) comparison of the two representations.

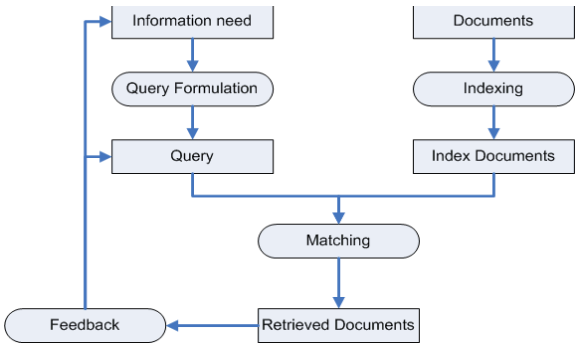

Figure 1 Block diagram of information retrieval process [1]

There are two commonly used metrics or evaluators.

I. Precision: This is the percentage of retrieved documents relevant to the query.

II. Recall: This is the percentage of documents that are relevant to the query and could be retrieved.

Following figure gives a normalized plot of the metrics of a case study.

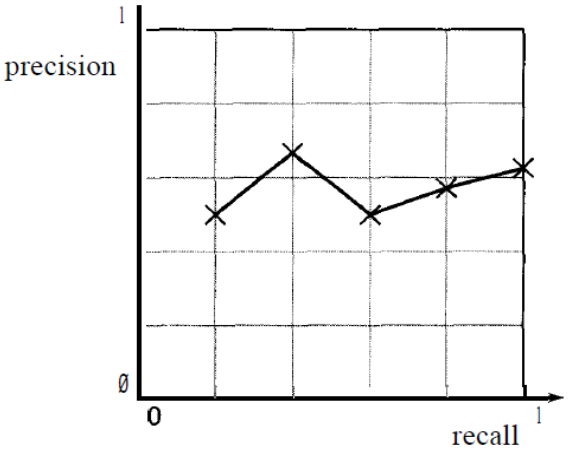

Figure 2: Metric evaluation of a case study [2].

As can be seen from the figure, there is no functional relation between the two metrics. For full evaluation of algorithms and tools more metrics should be developed.

The vector space model is one of the common used models for ad-hoc retrieval. Documents and queries are placed in an n-dimensional hyper space. The most relevant document makes the smallest angle with the query. Following figure gives an example vector space model in 2 dimensions.

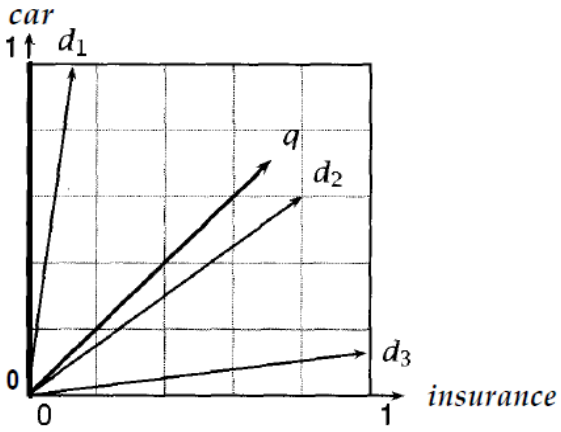

Figure 3: An example of Vector Space Model [2]

Two keywords 'Car' and 'Insurance' were used in the query. Document d_2 gets the highest rank as it makes the smallest angle with query q.

In a text document occurrence of a particular word can be counted. This will give the weight of the word to decide the rank. As an alternative, Poisson's distribution can be used. This gives the probability for a particular count. Following figure gives an example Poisson's distribution.

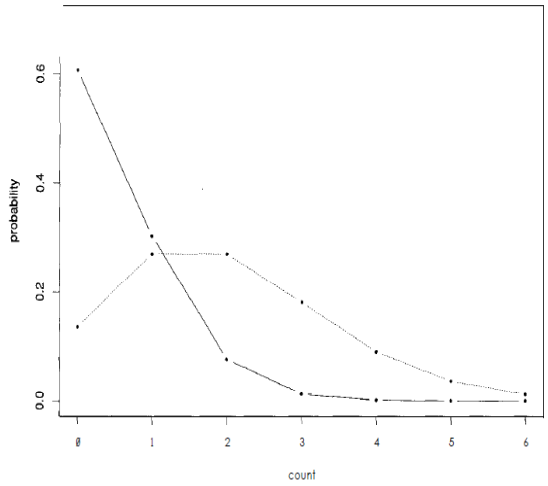

Figure 4: An example Poisson's distribution [2]

II. RELATED WORK

Kando [3] studied the implication of text-level structure for information retrieval of research papers. Following figure gives the evaluation of this case study.

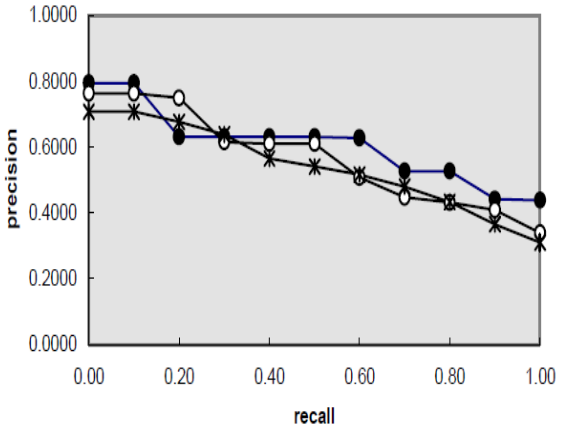

Figure 5: Metrics for a full-length text search [3]

Use of text level structure yields higher precision.

Information retrieval techniques from images are getting their due attention [4]. Standard practice is to annotate images to text and then apply the well established text information retrieval techniques. However, there are limitations of this standard technique. Image visualization is quite subjective. It depends on the viewer's eyesight, mood, and temperament. Instead of text, content based searching proves to be more useful. Learning of individual's perception using artificial neural network is quite promising. Following figure summarizes the major steps in image retrieval technique.

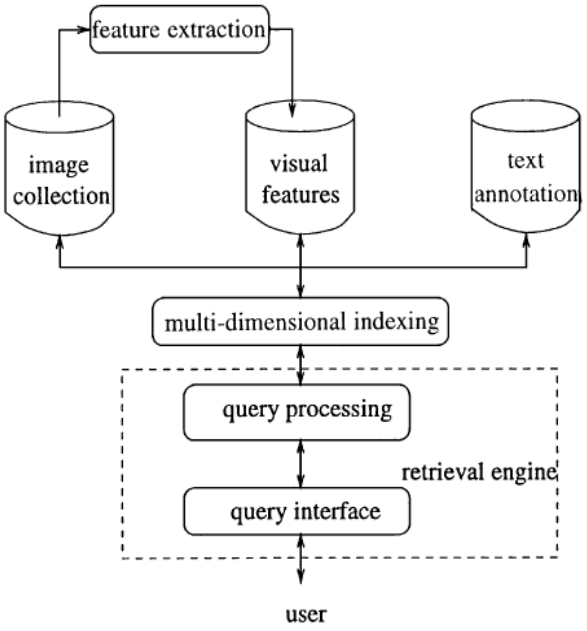

Figure 6: An image retrieval system architecture [4].

Music Information Retrieval (MIR) is very complex [5]. Like images, this information is also very subjective and depends on the mood of a person.

Stahlbock and Voβ studied various IR techniques and applied for container transportation [6]. There are many new challenges that demand attention of the researchers.

Fox reported about their developed tool on Composite Document Expert/Extended/Effective Retrieval (CODER) [7]. Several artificial intelligence algorithms are used to effectively handle wide varieties of documents. Performance comparison or benchmarking of the tool is not available in this paper. The relationship between various components, the experts, the blackboards, the external knowledge bases, and the various resource managers are given in the following diagram.

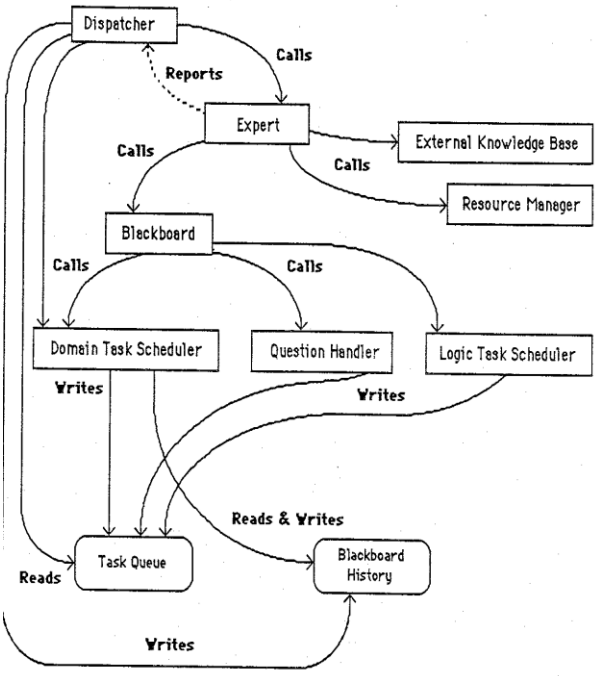

Figure 7: Calling hierarchy for CODER community of experts [7].

The structure is almost loop-free so it simplifies the task of the various components and prevents the dead-lock.

Vrajitoru applied Genetic Algorithm (GA) to optimize query processing [8]. Author further improved the return by modifying one of the steps of the algorithm, crossover. Following table summarizes the results. Two standard collections, (i) CACM with 320 documents, 50 queries and (ii) CISI with 146 documents, 35 queries were used. Collections are tabulated in the first column. Next two columns are with results for GA with improved crossover operation succeeded by two columns for standard GA. The pair of columns give the results where the algorithm is better and significantly better (difference > 5%) than the other algorithm. The last column gives the results for equal performance of both the algorithms.

TABLE I. COMPARISON OF THE OPERATOR BY QUERY PERCENTAGE [8].

Collection	Dissociated	Significant	Classical	Significant	Equality
CACM	60.33%	47.67 %	9.67 %	5.33%	30%
CISI	70%	46.67 %	20.95 %	10.48 %	9.05 %

The improvised algorithm not only performs better in all situations but also on a greater number of queries of CACM as well.

Liu et al. studied a standard dataset LETOR for benchmarking of ranking performance for different algorithms [9]. Following table gives a rank comparison for two different algorithms for a data subset TD2003.

TABLE II. MEAN AVERAGE PRECISION [9]

Algorithm	MAP
Rank Boost	0.212
Ranking SVM	0.256

Results show SVM has a slightly better performance. The study reveals that along-with the standard dataset standard feature for ranking is also needed. The area is open for research.

Application of Spreading Activation (SA) on semantic networks for IR is a very active research area [10]. This works on an already retrieved relevant document to retrieve associated documents. Semantic Networks express knowledge in terms of concepts. Each concept is represented by a node and the hierarchical relationship between concepts is depicted by links. Following figure gives a self explanatory example of SA.

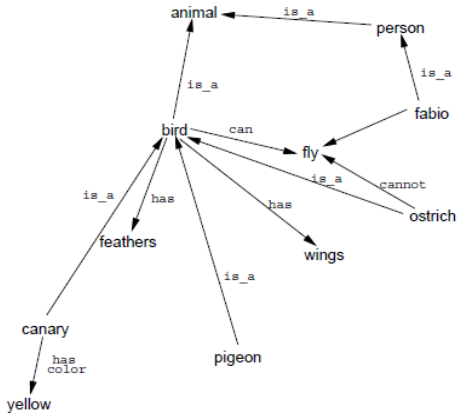

Figure 8: An example of Semantic Network [10].

This type of network can be represented in the general form as shown below. Here each link is weighted and needs a minimum activation. Further feedback between nodes is also possible.

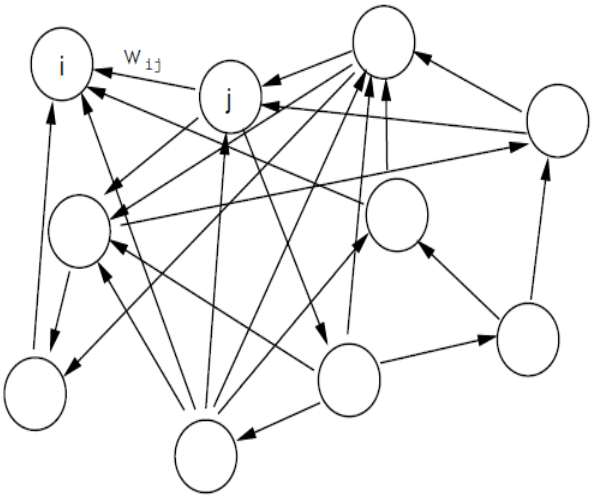

Figure 9: Network structure of Spreading Activation (SA) model [10].

Such type of network can be created for document retrieval.

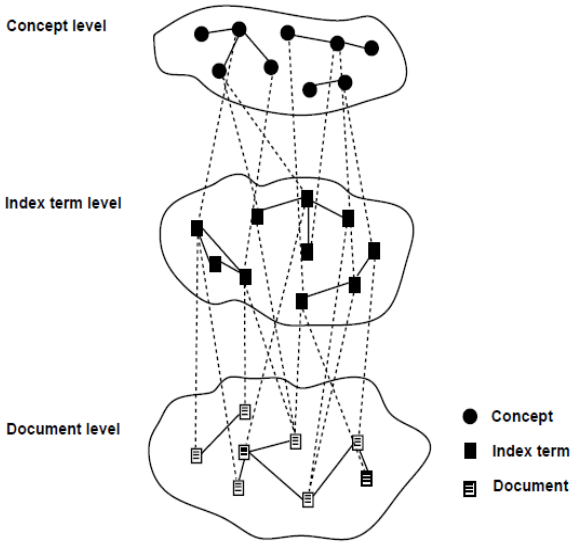

Figure 10: A three level network for IR [10].

The full retrieval process steps are described in the following two figures. The role of knowledge base can be understood when the steps in the two figures are compared.

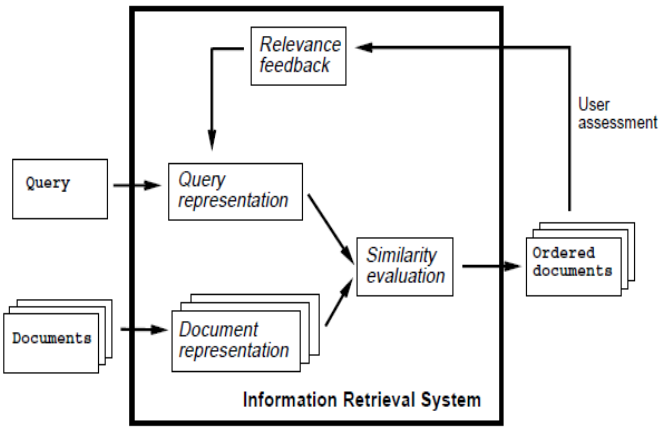

Figure 11: A Classical information retrieval system [10].

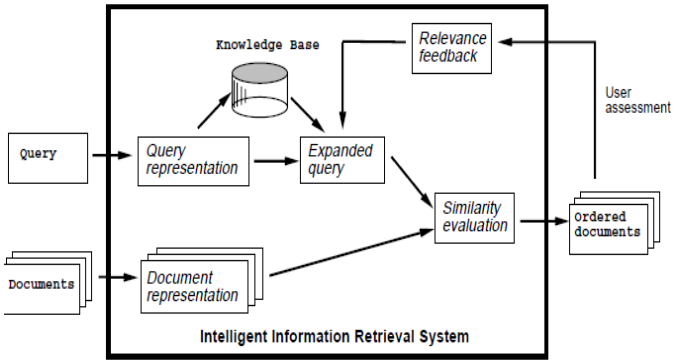

Figure 12: An intelligent information retrieval information system [10].

III. ANALYSIS

A. A Thrust Area for Application Oriented Research

With the advent of internet, uploading and downloading of document is becoming very easy. This is creating ever increasing ocean of documents. Automated retrieval of relevant data continues to be a challenge for this ever growing data.

As internet is getting more and more pervasive more information is available. Very efficient search engines are emerging for access of relevant information. Algorithms, the backbone of these engines, demand their due attention from the researchers and product developers.

B. Product Benchmarking

As strategies and algorithms used in the current search engines or IR products are not freely available it is difficult to find out the novelty of a recently developed algorithm. Superiority can be claimed after positive results from benchmarking of the product prototype. Standard document set for benchmarking are inadequate. Further, relevance of information is quite subjective. A meticulous survey on human population is needed for useful benchmarking.

C. System Diversity

Nowadays IR is not limited desktop computer search. Search is needed more frequently in mobile handsets, embedded systems. A tool that works very efficiently on a desktop is unlikely to perform well on a platform with limiting processing capabilities. Algorithms for such environment should be researched.

D. Speed of Execution

With the advent of electronics instruction processing on hardware system is getting super fast. Time complexity of the search algorithm should be reasonable to avoid serious overhead for the retrieval process. There is always a trade-off between time complexity and space complexity. For systems with large memory, time complexity should be improved at the cost of space complexity

E. Fuzzy vs. Statistics

So far, fuzzy logic is under utilized in this research domain. Algorithms are mainly based on Natural Language Processing (NLP) that depends on statistical analysis. This type of analysis predicts the probability of relevance of a document. It will be more useful if the relevance of a document is given with the fuzzy coefficient.

Role of fuzzy coefficient can be understood from a simple scenario. Suppose you are given two glasses of drinking water. Qualities of water in the two glasses are totally unknown except the following information. The probability that the water is potable is 0.9 for the first glass of water. The fuzzy coefficient that the water is potable is 0.9 for the second glass. Which glass should you select for a safe drink?

The answer is obviously the second glass. There is a chance that the water in the first glass is totally unsafe. The glass might have been picked up randomly from a tray with hundred glasses containing safe drinking water of very good quality. Out of them 10 got contaminated with deadly poison. There is a small chance that the first glass is holding a deadly poison. For the second glass possibility of poison is ruled out. The solutes in the water deviate slightly from the recommendations. Our body can withstand such deviations. It is always safe to drink this type of water.

Relevance of a document to be retrieved can be compared with the safety of drinking water. It is better to accept a document that is almost relevant and rejecting the other that is likely to be relevant. Further statistics usually deals with crisp sets of relevant and irrelevant. In this case, fuzzy set is more appropriate.

F. Use of Artificial Neural Network

In recent research Artificial Neural Network (ANN) is being used quite widely. ANN, a very good learner, is a very good candidate for repeated similar types of searches. ANN can be further used to learn an individual's style and needs. To the best of author's knowledge, such type of study has not been reported so far.

There are some commercial search tools asking users' feedback about relevance of the retrieved information. However, no learning has been observed for these tools.

G. Authenticity of Documents

Present retrieval techniques do not give much information about authenticity of the data. A noise or misinformation analysis can be done to predict the authenticity of the available data.

IV. CONCLUSION

In this survey, several aspects of IR are analyzed. Firstly, IR process is clearly defined. Process steps are explained with the help of diagrams. Two metrics commonly used for evaluation of tools and algorithms are defined and analyzed.

IR is discussed mainly for text information. Other two areas: (i) image and (ii) music where application of IR pops up additional challenges are analyzed. An unconventional area namely, container transportation, where IR has been applied is reviewed.

Applications of recent emerging concepts e.g. artificial intelligence, fuzzy logic, genetic algorithm are briefed.

As the technology advances newer challenges are popping up. The area demands constant attention of the researchers and product developers.

REFERENCES

[1] A. Roshdi and A. Roohparvar, "Review: Information retrieval techniques and applications", International Journal of Computer Networks and Communications Security, vol. 3, no. 9, pp. 373 – 377, September 2015.

[2] C.D. Manning and H. Schutze, Foundations of Statistical Natural Language Processing, 2nd ed., MIT Press, Massachusetts, 2000.

[3] N. Kando, "Text-level structure of research papers: Implications for text based information processing systems", Proc. 19th Annual BCS-IRSG Colloquium on IR Research, Aberdeen, Scotland, April, 1997, pp. 1 – 14.

[4] Y. Rui and T.S. Huang, "Image retrieval: Current techniques, promising directions, and open issues", Journal of Visual Communication and Image Representation, vol. 10, pp. 39 – 62, 1999.

[5] J.S. Downie, "Music information retrieval", in Annual Review of Information Science and Technology, vol. 37, ch. 7, C. Blaise, Ed. Medford, NJ: Information Today, 2003, pp. 295 – 340. Available: http://music-ir.org/downie_mir_artist37.pdf on 31st December, 2016.

[6] R. Stahlbock and S. Voβ, "Operations research at container terminals: a literature update", OR Spectrum, Springer-Verlag, vol. 30, pp. 1 – 52, 2008.

[7] E.A. Fox, "Development of the coder system: A test-bed for artificial intelligence methods in information retrieval", Information Processing and Management, vol. 23, no. 4, pp. 1 – 29, 1987.

[8] D. Vrajitoru, "Crossover improvement for the genetic algorithm in information retrieval", Information Processing and Management, vol. 34, no. 4, pp. 405 – 415, 1998.

[9] T. Liu, J. Xiu, T. Qin, W. Xiong, and H. Li, "LETOR: Benchmark dataset for research on learning to rank for information retrieval", in Proc. SIGIR 2007 workshop on Learning to Rank for Information Retrieval, pp. 3 – 10, 2007.

[10] F. Crestani, "Application of spreading activation techniques in information retrieval", Artificial Intelligence Review, vol.11, no. 6, pp. 453 – 482, December 1997, Kluwer Academic Publishers. Norwell, MA, USA.