International Journal of Information, Control and Computer Sciences
ISSN: 2517-9942
Vol:7, No:12, 2013

A Review on Important Aspects of Information Retrieval

Yogesh Gupta, Ashish Saini, A.K. Saxena

***Abstract*—**Information retrieval has become an important field of study and research under computer science due to explosive growth of information available in the form of full text, hypertext, administrative text, directory, numeric or bibliographic text. The research work is going on various aspects of information retrieval systems so as to improve its efficiency and reliability. This paper presents a comprehensive study, which discusses not only emergence and evolution of information retrieval but also includes different information retrieval models and some important aspects such as document representation, similarity measure and query expansion.

***Keywords*—**Information Retrieval, query expansion, similarity measure, query expansion, vector space model.

I. Introduction

INFORMATION Retrieval (IR) is the science of searching for information within relational databases, documents, text, multimedia files, and the World Wide Web [1]. The idea of searching for information was first mentioned by Vannevar Bush [2]. Mooers [3] and Savino [4] have defined IR as follows:

"Information retrieval is the name of the process or method whereby a prospective user of information is able to convert his need for information into an actual list of citations to documents in storage containing information useful to him."

Fig. 1 shows general IR system architecture [5]. In this figure, the user who needs information issues a query (**user query**) to the **retrieval system** through the **query operations** module. The retrieval module uses the **document index** to retrieve those documents that contain some query terms (such documents are likely to be relevant to the query), compute relevance scores for them, and then rank the retrieved documents according to the scores. The ranked documents are then presented to the user. The **document collection** is also called the **text database**, which is indexed by the **indexer** for efficient retrieval.

The objective of any IR system is to produce a list of relevant documents to the user information need or query provided by the user. The user usually needs relevant documents even if the exact terms s/he used in the provided query were not present in these documents.

Yogesh Gupta, Ashish Saini (Corresponding author) and A.K. Saxena are with the Department of Electrical Engineering, Faculty of Engineering, Dayalbagh Educational Institute, Agra – 282110, Uttar Pradesh, India (e-mail: er.yogeshgupta@gmail.com, ashish7119@gmail.com, aksaxena61@gmail.com).

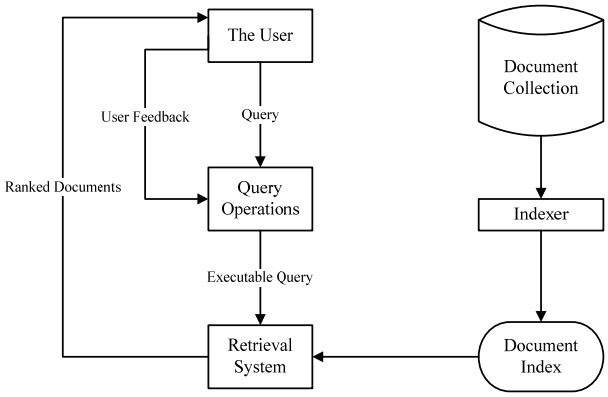

Fig. 1 A general IR system architecture

Therefore, a major difference between information retrieval systems and other kinds of information systems is the uncertainty nature of IR. Table I elaborates on the difference between the IR and data retrieval systems [6].

TABLE I
Difference between Information Retrieval and Data Retrieval

Feature	Information retrieval	Data Retrieval
Data	Free text, unstructured	Database tables, structured
Queries	Keywords, Natural language	SQL, Relational algebras
Results	Approximate matches	Exact matches
Results	Ordered by relevance	Unordered
Accessibility	Non-expert humans	Knowledgeable users or automatic processes

This type of IR is also different from the web IR systems that aim at finding information over the web where billions of web pages are available and fast search engines are required [7]. Table II compares the web IR and the classical IR systems.

TABLE II
Difference between Information Retrieval and Web IR

Feature	Information Retrieval	Web IR
Volume	Large	Huge
Data quality	Clean	Noisy
Data change rate	Infrequent	In flux
Data accessibility	Accessible	Partially accessible
Format diversity	Homogeneous	Widely diverse
Documents	Text	HTML
Number of matches	Small	Large
IR techniques	Content-based	Link-based

II. Historical Perspective

The need to store and retrieve written information has become increasingly important over centuries. The first systematic solution to the problem of finding the desired information from a large information collection was developed about 2,000 years ago by librarians, who kept track of "books" by cataloging them by author and the title. Searching through the catalog to find a book was a marked improvement from the physical search of actual books, but it required the searcher to know the book as well as to know its author and the title. In this field, Soper [8] filed a patent for a device in 1920, where catalogue cards with holes, related to categories, were aligned in front of each other to determine if there were entries in a collection with a particular combination of categories. If light could be seen through the arrangement of cards, a match was found.

In 1945, V. Bush [9] published a ground breaking article titled "As We May Think" that gave birth to the idea of automatic access to large amounts of stored knowledge. In 1950s, this idea materialized into more concrete descriptions of how archives of text could be searched automatically. Several works emerged in mid 1950s that elaborated upon the basic idea of searching text with a computer. One of the most influential methods was described by H.P. Luhn [10], in which he proposed the use of words as indexing units for documents and for measuring word overlap as a criterion for retrieval. IR as a research discipline was starting to emerge at this time with two important developments: how to index documents and how to retrieve them.

After the invention of computers, very soon people realized that they could be used for storing and mechanically retrieving large amounts of information. Mitchell [11] described a project to model the use of a Univac computer to search 1,000,000 records indexed by up to six subject codes, it was estimated that it would take 15 hours to search that many records. Nanus [12] detailed a number of computer-based IR projects run in the 1950s, including one system from General Electric that searched over 30,000 document abstracts.

A. Indexing

In the field of librarianship, the way that items were organized in a collection was a topic that was regularly debated. The classic approach was to use a hierarchical subject classification scheme, such as the Dewey Decimal system, which assigned numerical codes to collection items. However, alternatives were proposed, most notably Taube et al.'s Uniterm system [13], which was essentially a proposal to index items by a list of keywords. As simple an idea as this seems today, this was at the time a radical step. A few years later, Cleverdon [14] conducted a detailed comparison of retrieval effectiveness using Uniterms and the more classic classification techniques. His conclusion that Uniterms were as good as and possibly better than other approaches caused much surprise and his work came in for extensive scrutiny [15]. However, Cleverdon's experimental results were found to be correct and as a result the use of words to index the documents of an IR system became established. Many aspects of Cleverdon's test collection approach to evaluation are still used in both academic research and commercial search testing today.

B. Ranked Retrieval

The style of search used by both the electro-mechanical and computer-based IR systems was so-called Boolean retrieval. A query was a logical combination of terms which resulted in a set of those documents that exactly matched the query. Luhn [16] proposed and Maron et al. [17] tested an alternative approach, where each document in the collection was assigned a score indicating its relevance to a given query. The documents were then sorted and those at the top ranks were returned to the user. The researchers manually assigned keywords to a collection of 200 documents, weighting those assignments based on the importance of the keyword to the document. The scores assigned to the documents were based on a probabilistic approach. The researchers tested their ranked retrieval method, showing that it outperformed Boolean search on this test collection with 39 queries. In the same year as Maron et al.'s work, Luhn suggested "*that the frequency of word occurrence in an article furnishes a useful measurement of word significance*" [16]. This approach later became known as term frequency weighting. This ranked retrieval approach to search was taken up by IR researchers, who over the following decades refined and revised the means by which documents were sorted in relation to a query.

Several key developments in this field happened in 1960s. One of these developments was the formalization of algorithms to rank documents relative to a query. Of particular note was an approach where documents and queries were viewed as vectors within an N dimensional space (N being the number of unique terms in the collection being searched). This was first proposed by Switzer [18]. Later the similarity between a document and query vector was suggested by Salton to be measured as the cosine of the angle between the vectors using the cosine coefficient [19]. Another significant innovation at this time was the introduction of relevance feedback [20]. This was a process to support iterative search, where documents previously retrieved could be marked as relevant in an IR system. A user's query was automatically adjusted using information extracted from the relevant documents. Versions of this process are used in modern search engines, such as the "Related articles" link on Google Scholar. Relevance feedback was also the first use of machine learning in IR.

Other IR enhancements examined in this period included the clustering of documents with similar content, the statistical association of terms with similar semantic meaning, increasing the number of documents matched with a query by expanding the query with lexical variations [21] or with semantically associated words [22]. In this decade, commercial search companies emerged out of the development of bespoke systems built for large companies or government organizations. Bjorner [23] states that one of the first companies dedicated to providing search was Dialog

formed in 1966 from the creation of an IR system for NASA. A striking aspect of this time was the low level of interaction between the commercial and IR research communities. Despite researchers' consistent demonstration that ranked retrieval was a superior technique, almost all commercial searching systems used Boolean search. This situation didn't change until the early to mid-1990s with systems such as WESTLAW's WIN system [24] and the growth of web search engines.

The period of 1970s and 1980s saw many developments built on the advances of the 1960s. One of the key developments of this period was that Luhn's term frequency (*tf*) weights (based on the occurrence of words within a document), were complemented with Spärck Jones's work on word occurrence across the documents of a collection. Her paper on inverse document frequency (*idf*) introduced the idea that the frequency of occurrence of a word in a document collection was inversely proportional to its significance in retrieval less common words tended to refer to more specific concepts, which were more important in retrieval [25]. The idea of combining these two weights (*tf.idf*) was quickly adopted by Salton and Yang [26]. A number of researchers worked to formalize the retrieval process. Salton synthesized the outputs of his group's work on vectors to produce the vector space model [27]. This approach to describing the retrieval process underpinned many research retrieval systems and much research for the coming two decades.

An alternative means of modeling IR systems involved extending Maron's idea of using probability theory. Robertson defined the probability ranking principle [28], which determined how to optimally rank documents based on probabilistic measures with respect to defined evaluation measures. A further paper from Robertson and Spärck Jones [29] along with a derivation of the probabilistic model in Van Rijsbergen's book [22] stimulated much research on this form of modeling. Van Rijsbergen showed that the basic probabilistic model assumed that words in a document occurred independently of each other, which is a somewhat unrealistic assumption. Incorporating term dependency into ranked retrieval started to be examined, which lead to a wide range of research in later years. Building on the developments of the 1970s, variations of *tf.idf* weighting schemes were produced [30] and the formal models of retrieval were extended. The original probabilistic model did not include *tf* weights and a number of researchers worked to incorporate them in an effective and principled way. Advances on the basic vector space model were also developed and probably the most well-known is Latent Semantic Indexing (LSI), where the dimensionality of the vector space of a document collection was reduced though singular-value decomposition [31].

Various models for doing document retrieval were developed and advances were made along all dimensions of the retrieval process. These new models were experimentally proven to be effective on small text collections available to researchers at that time. However, due to lack of availability of large text collections, the question whether these models and techniques would scale to larger corpora remained unanswered. This changed in 1992 with the inception of Text Retrieval Conference (TREC) [32]. Kraft et al. [33] used fuzzy rules to retrieve information.

C. Development of Similarity Measure

Philip Resnik [34] presented measure of semantic similarity measure in taxonomy, based on the notion of information content. In 1998, Lin [35] proposed that, bootstrapping semantics from text is one of the greatest challenges in natural language learning. They defined a word similarity measure based on the distributional pattern of words. Lin also presented an information theoretic definition of similarity that is applicable as long as there is a probabilistic model. Fan [36] presented similarity functions as trees and a classical generational scheme. Fan et al. [37] also presented a different approach to compute similarity measure to improve IR process. Pathak et al. [38] have proposed the idea of combined similarity measure in which they have proposed a linear combination of various similarity measures and then optimize the weight of each similarity measure using Genetic Algorithm (GA). In 2004, Jian Pei et al., [39] proposed a projection-based, sequential pattern growth approach for efficient mining of sequential patterns and Ming Li et al. [40] proposed a metric based on the non-computable notion of Kolmogorov computable distance and called it the similarity metric. Mehran Sahami [41] proposes a novel method for measuring the similarity between short text snippets by leveraging web search results to provide greater context for the short texts. In this paper, a method for measuring the similarity between short text snippets is proposed that captures more of the semantic context of the snippets rather than simply measuring their term-wise similarity. In the same year Chen [42] proposed a web search with double checking model to explore the web as a live corpus. Instead of simple web page counts and complex web page collection, the proposed novel model is a Web Search with Double Checking (WSDC) used to analyze snippets. In 2007, Rudi L. Cilibrasi et al. [43] proposed the words and phrases acquire meaning from the way they are used in society, from their relative semantics to other words and phrases. It is a new theory of similarity between words and phrases based on information distance and Kolmogorov complexity. The method is applicable to all search engines and databases. Authors are introduced some notions underpinning the approach: Kolmogorov complexity, information distance, and the Normalized Google Distance. Hughes et al. [44] proposed a method that presents the application of random walk Markov chain theory for measuring lexical semantic relatedness. Zuber et al. [45] present a novel approach that allows similarities to be asymmetric while still using only information contained in the structure of the ontology. Tuomo et al. [46] used a connection between the cosine measure and the Euclidean distance in association with principal component analysis and grounded searching on the latter then applied the single and complete linkage and Ward clustering to Finnish documents utilizing

their relevance assessment as a new feature**.** Gledson et al. [47] describes a simple web-based similarity measure which relies on page counts only, can be utilized to measure the similarity of entire sets of words in addition to word pairs and can use any web-service enabled search engine distributional similarity measure which uses internet search counts and extends to calculating the similarity within word-groups. Torra et al. [48] presented a method to calculate similarity between words based on dictionaries using Fuzzy graphs. Bollegala et al. [49] proposed a method which exploits the page counts and text snippets returned by a Web search engine. Chen [50] presented a new similarity measure based on the geometric mean averaging operator to handle the similarity problems of generalized fuzzy numbers. Usharani et al. [51] proposed a GA based method for finding similarity of web document based on cosine similarity.

D.Emergence of Query Expansion

Query expansion has a long history in IR, as it has been suggested as early as 1960 by Maron and Kuhns [52]. Early work investigated a range of seminal techniques that have been subsequently improved and extended in various ways, for example, vector feedback [53], [54], term clustering [55]-[57] and comparative analysis of term distributions [58], [59]. Van Rijsbergen [22] proposed a relevance feedback technique to modify the original query by adding some other relevant terms**.** Xu and Croft [60] used the local analysis and the global analysis of documents for query expansion. Cooper and Byrd [61] constructed a visual interface with graphical relations between items by lexical neighborhoods for prompted query refinement. Chen [62] used a GA to learn query terms that better represent a relevant document set provided by the user. On the other hand, in a number of early experiments performed on small scale collections inconclusive results were achieved about the retrieval effectiveness of such techniques, with gain in recall often compensated by the corresponding loss in precision [63], [64]. Horng [65] used a GA to adapt the query term weights in order to get the closest query vector to the optimal one. Li and Agrawal [66] used multi-granularity indexing and query processing for supporting the web query expansion and Wei et al. [67] presented a method to mine term association rules for automatic global query expansion. Chen et al. [68] used association rules to discover the degrees of similarity between terms and constructed a hierarchical-tree structure to pick out query expansion terms. Takagi and Tajima [69] presented a method for query expansion using conceptual fuzzy sets for search engines. It calculates the degrees of similarity between terms to construct a hierarchical tree structure and lets terms with higher degrees of similarity be expansion terms of the structure. Kim et al. [70] also presented a method for query term expansion and reweighting using the term co-occurrence similarity and fuzzy inference techniques. Cui et al. [71] presented a method for probabilistic query expansion using query logs. Billerbeck et al. [72] proposed a method for query expansion using associated queries. Chang et al. [73] presented a query expansion method based on fuzzy rules. In the same year Jin et al. [74] developed a method for query expansion based on the term similarity tree model. Latiri et al. [75] considered the relationship between terms and documents as a fuzzy binary relation, based on the closure of the extended fuzzy Galois connection, and used fuzzy association rules to find out real correlated terms as query expansion terms. Nakauchi et al. [76] created thesaurus and relationships of terms for query expansion. In the same year Safar and Kefi [77] presented a query expansion method based on the domain ontology and the lattice structure. Berardi et al. [78] used association rules to mine query expansion terms and presented how to filter off redundant association rules. Martin et al. [79] presented a method to mine web documents for finding additional query terms. Stojanovic [80] used a conceptual schema to query neighborhood for query expansion. Lin et al. [81] presented a method for mining additional query terms for query expansion. Michel and Annabelle [82] proposed an IR model using the fuzzy proximity degree of term occurrences. Chang et al. [83] presented a new method for query reweighting to deal with document retrieval**.** Grootjen et al. [84] presented a new, hybrid approach that projects an initial query result onto global information, yielding a local conceptual overview. Billerbeck et al. [85] proposed a new method that draws candidate terms from brief document summaries that are held in memory for each document. Chang et al. [86] proposed a new query expansion method for document retrieval based on fuzzy rules. Nowacka et al. [87] proposed a comprehensive fuzzy based model of information retrieval. Fattahi et al. [88] presented a new approach to query expansion in search engines through the use of general non-topical terms and domain-specific semi-topical terms. Cecchini et al. [89] proposed techniques place emphasis on searching for novel material that is related to the search context. Carlos et al. [90] proposed a semi-supervised algorithm to incrementally learn terms that can help bridge the terminology gap existing between the user's information needs and the relevant documents' vocabulary**.** Piotr Wasilewski [91] presented a method for query expansion using semantic modeling of information need**.** Liu et al. [92] proposed two algorithms for query expansions. First is iterative single keyword refinement and second is elimination based convergence. Tayal et al. [93] presented a method for fuzzy weighting of query terms with the help of fuzzy triangular membership function. Latiri et al. [94] proposed automatic query expansion method using association rule mining approach**.**

III. Important Aspects of Information Retrieval

Although there are many aspects in Information Retrieval System but some prime aspects are document representation, similarity measure and query expansion.

A. Document Representation

Traditionally, documents may be available in different forms e.g. full text, hypertext, administrative text, directory, numeric or bibliographic text. It is very difficult to extract relevant information from these forms of documents. Therefore, first these documents should be represented in an

appropriate manner with the help of any IR model. Such IR model provides the fundamental premises and forms the basis for ranking. In general, IR models operate on large and fixed collections of documents (corpus), from which they attempt to find out the useful information that best matches to a query. Yates [95] gives general definition of an IR model as:

Definition: *An IR model is a quadruple [D, Q, F, R* (q_i, d_j)*], where*

***1. D** is a set composed of logical views for the documents in the collection*

***2. Q** is a set composed of logical views for the user information needs expressed as queries*

***3. F** is a framework for modeling document representations, queries and their relationships*

4. R **(q_i, d_j)** *is a ranking function which associates a real number with a query $q_i \in Q$ and a document representation $d_j \in D$. Such ranking defines an ordering among the documents with regard to the query q_i.*

Bing Liu [5] has presented various IR models. The main IR models are described as follows:

1. Boolean Model

The Boolean model [96] was the first model which was adopted by most of the earlier systems and even today some of the commercial systems use this model, which makes use of the concepts of Boolean logic and set theories. The documents and the queries are a collection of terms and each term from the document is indexed. The presence and absence of a term in a document is represented by 1 and 0 respectively. For the term matching of document and query we maintain an inverted index of the terms i.e. for each term we must store a list of documents that contain the term. The terms are tokenized using linguistic models for those terms which can be stemmed down.

Further, the Boolean model often retrieved either too many or too few documents due to the sensitive nature of the Boolean logic that responds rigidly to the absence or presence of a single term. To overcome the problem of output overload i.e., too many documents are retrieved without regard to their degree of potential importance to the user refinements to the system were made to produce ranked outputs by assigning weights to terms based on their "presumed" importance. Other refinement strategies, such as controlling the query formulation process to ease the difficulty of constructing complex Boolean queries, were investigated as well. While some tried to overcome the weaknesses of the Boolean model by building refinements to the existing Boolean model, others approached IR with a different search strategy called the Vector Space model.

2. Vector Space Model

The Vector Space Model, as the name implies, represents documents and queries internally in the form of vectors. In the vector space model all queries and documents are represented as vectors in |V|-dimensional space, where V is the set of all distinct terms in the collection. A document vector contains index terms from the documents that to some extent describe its contents [97]. At the center of the vector space model is the similarity measure, which is used to measure the angle between two vectors. The framework of the vector space model [98] employs a ranking algorithm that tries to rank documents in order of how much of an overlap is between the terminology of the query and each document, where relatively rare terms have comparatively higher weights. Conceptually, documents are ranked on the basis of similarity measure. Some of the advantages of the Vector Space Model are that it is simple and fast model, that it can handle weighted terms, that it produces a ranked list as output and that the indexing process is automated which means a significantly lighter workload for the administrator of the collection. Also, it is easy to modify individual vectors, which is essential for the query expansion technique [97]. The Vector Space Model has few weaknesses. The first weakness is the assumed independency between terms. Due to the locality of many term dependencies, their indiscriminate application to all the documents in the collection might badly affect the retrieval performance [95]. The second weakness is that there are no theoretical justifications to use which similarity coefficients for a particular application and also some of the vector-manipulating operations.

The vector space model continues to be used in a variety of information retrieval areas apart from document retrieval, such as document categorization [99], [100] collaborative filtering [101]**.**

3. Probabilistic Model

The Probabilistic model is similar to vector space model in its representation of documents and queries as vectors, but instead of retrieving documents based on their similarities to the query, the probabilistic model retrieves documents based on their probability of relevance to the query. Rooted in the probabilistic notions introduced by Maron [52], the probabilistic model views the principal function of IR as ranking of documents in the order of decreasing probability of relevance to a user's information need [28]. The basic idea of the probabilistic model is to calculate the term weights, which define the probability of relevance of documents, based on the data about the distribution of query terms in documents that have been assessed for relevance. When term independence is assumed, the probability of relevance for a given document can be calculated by summing its individual term relevance weights, which are the estimations of probabilities that given terms in a query will appear in a relevant document but not in a non-relevant document. The probabilistic model suffers from the same limitation as the vector space model owing to the term independence assumption, an assumption introduced merely for the sake of computational simplicity.

B. Similarity Measure

The IR system needs to calculate the similarity of the query and the particular document in order to decide relevancy of that document with the query. When a document retrieval system is used to query a collection of documents with t

terms, the system computes a vector D ($d_{i1}, d_{i2}, \ldots d_{it}$) of size t for each document. The vectors are filled with the weights and similarly, a vector Q ($W_{q1}, W_{q2}, \ldots W_{qt}$) is constructed for the terms found in the query. There are several typical vector similarity measures as follows:

1. Cosine: - One drawback with using the Inner product is that longer documents, having more terms, will dominate the similarity calculations. Therefore, the vectors need to be normalized. The most common of these is the cosine measure where the cosine of the angle between the query and document vector is given in (1).

$$Cos\ (Q,\ D_i) = \frac{\sum_{j=1}^{t} W_{qj}\ d_{ij}}{\sum_{j=1}^{t} (d_{ij})^2 \sum_{j=1}^{t} (W_{qj})^2} \quad (1)$$

2. Jaccard Coefficient: -The Jaccard coefficient is defined as the size of the intersection divided by the size of the union of the document and query vectors as expressed in (2).

$$Jaccard\ (Q,\ D_i) = \frac{\sum_{j=1}^{t} W_{qj}\ d_{ij}}{\sum_{j=1}^{t} (d_{ij})^2 + \sum_{j=1}^{t} (W_{qj})^2 - \sum_{j=1}^{t} W_{qj}\ d_{ij}} \quad (2)$$

3. Okapi: - Okapi similarity measure can be expressed in (3). This is one of the most popular methods used in the traditional IR field. Unlike VSM, the Okapi method not only considers the frequency of the query terms, but also the average length of the whole collection and the length of the document under evaluation.

$$Okapi\ (Q,\ D_i) = \sum_{T \in Q} W\ \frac{(k_1+1)\ tf}{K + tf} \times \frac{(k_3+1)\ qtf}{k_3 + qtf} \quad (3)$$

Where Q is a query that contains the words T
k1, b, and k3 are constant parameters (k1=1.2 and b=0.75 work well, k3 is 7 or 1000)
K is $k_1((1\text{-}b) + (b \cdot dl / avdl))$
tf is the term frequency of the term with a document, qtf is the term frequency in the query
W is $log\ \frac{(N - n + 0.5)}{(n + 0.5)}$, N is the number of documents, n is the number containing the term
dl and avdl are the document length and average document length.

C. Query Expansion

Query Expansion is one of the promising approaches to deal with word mismatch problem in information retrieval [60]. The basic idea of query expansion is to expand a user query by adding terms that are relevant to the original query terms. Since the expanded query contains more terms, the probability of matching them with terms in relevant documents is therefore increased. Three common types of query expansion are manual, interactive and automatic based on the role of involvement of user in whole process. One argument in favor of Automatic Query Expansion (AQE) is the system that has access to more statistical information on the relative utility of expansion terms and can make a better selection of which terms to add to the user's query.

AQE can be divided into two methods including global analysis and local feedback. The global analysis method relies on a thesaurus, typically constructed from a document corpus. Using the thesaurus, the global analysis method generates a ranked list of terms with respect to the original query terms and the top *n* terms are added to original query [60], [102]. On the other hand, the local feedback method first retrieves *N* documents that are most relevant to the original query, extracts the most important *n* terms from those documents, and subsequently, adds the extracted terns to the original query. As it assumes that top *N* documents are most relevant, it is also known as pseudo relevance feedback method. One of the problems inherent to the global analysis method for query expansion is that a global thesaurus is constructed and employed for expanding user queries. That is, a single weight for each pair of terms is derived from a collection of documents. Typically, the collection contains documents with different themes. For example, a set of information technology related documents may be classified into such themes as databases, artificial intelligence, computer architecture, and operating systems. In this case, the global view of term associations taken by the global analysis method for query expansion may not be adequate since the strengths/weights between two terms may be dissimilar or even totally different across different themes. For example, the terms "Feasibility Study" and "Quality Assurance" may be highly relevant under the theme of software engineering, while they are less relevant or even irrelevant in such themes as database and computer architecture. Thus, with the thematic view of term associations, for a user query "Feasibility Study," the term "Quality Assurance" should be added to original query under the theme of software engineering but should not be added under the theme of computer architecture. When taking the global view of term associations, the global weight between a pair of terms is a compromise of local weights across different themes. Thus, when expanding terms for a query, the global analysis method may select terms that are compromised across different themes rather than highly relevant terms in some of the themes in the document collection; thus, potentially limiting its retrieval effectiveness. In contrast to global analysis method, the local feedback method does not depend on a pre-constructed thesaurus for query expansion. Hence, it does not encounter the same problem as the global analysis does. Moreover, the local feedback method could result in better retrieval if the top *N* documents initially retrieved and used for feedback are in fact relevant to the original query [103].

D. Evaluation of performance of Information Retrieval System

The performance of any IR system can be evaluated by following four parameters.

1) Precision: Precision is a fraction of documents that are relevant among the entire retrieved document.

2) Recall: Recall is a fraction of the documents that are retrieved and relevant among all relevant documents.

3) Precision-Recall Curve: This curve is based upon the value of precision and recall where the x-axis is recall and y-axis is precision. Instead of using precision and recall on at each rank position , the curve is commonly plotted using 11 standard recall level 0%, 10%, 20% ……..100%.

4) F-score: F-score is harmonic mean of precision and recall.

IV. Conclusion

In nineteenth century, a person with an information need had to approach library and used a card catalogue, locate books or documents that hopefully answered his/her need. Because of the relative inconvenience of accessing information in that way, that person was able to seek answers of a small number of questions. The scope of information available to people was limited by the size of their library. But now a days, due to the ubiquity of web-based search, it need hardly be said what the current state of the art is for those with an internet connection, one can instantaneously access hundreds of terabytes of information. Due to continuously growing this information size, the question to get most relevant information pertinent to a query becomes an important issue. The various researchers have proposed different methods and models such as indexing, ranked retrieval, similarity measures and query expansion to resolve this problem in recent years. This survey covers all the important development related to above mentioned aspects of information retrieval.

References

[1] Lauren D, Joseph B (1975) Information Retrieval and Processing, Melville.
[2] Singhal A (2001) Modern Information Retrieval: A Brief Overview. Bulletin of the IEEE Computer Society Technical Committee on Data Engineering.
[3] Mooers CN (1950) Information retrieval viewed as temporal signaling. Proceedings of the International Congress of Mathematicians 1: 572-573.
[4] Savino P, Sebastiani F (1998) Essential bibliography on multimedia information retrieval, categorization and filtering. 2nd European Digital Libraries Conference Tutorial on Multimedia Information.
[5] Liu B (2007) Web Data Mining: Exploring Hyperlinks, Contents and Usage Data. Springer-Verlag, Berlin Heidelberg.
[6] Khatatneh K, Hussain I (2002) Information Retrievals tree v/s Inverted File Word Method for Arabic language. Prince Abdu Allah Bin Ghazi for IT, Al-Balqa Applied University Salt, Jordan.
[7] Christopher DM, Raghavan P, Hinrich S (2008) An Introduction to Information Retrieval. Cambridge University Press.
[8] Soper HE (1920) Means for compiling tabular and statistical data, U.S.
[9] Bush V (1945) As We May Think. Atlantic Monthly 176: 101-108.
[10] Luhn HP (1957) A statistical approach to mechanized encoding and searching of literary information. IBM Journal of Research and Development 1(4): 309–317.
[11] Mitchell HF (1953) The Use of the University AC FAC-Tronic System in the Library Reference Field. American Documentation 4(1): 16-17.
[12] Nanus B (1960) The Use of Electronic Computers for Information Retrieval. Bull Med Library Association 48(3): 278-291.
[13] Taube M, Gull CD, Wachtel IS (1952) Unit terms in coordinate indexing. American Documentation 3(4): 213-218.
[14] Cleverdon CW (1959) The Evaluation of Systems Used in Information Retrieval. Proceedings of the International Conference on Scientific Information 2: 687-698.
[15] Cleverdon CW (1991) The significance of the Cranfield tests on index languages. Proceedings of the 14th annual international ACM SIGIR conference on Research and development in information retrieval, Chicago, Illinois, United States, 3-12.
[16] Luhn HP (1958) The automatic creation of literature abstracts. IBM Journal of Research and Development 2(2): 159-165.
[17] Maron ME, Kuhns JL, Ray LC (1959) Probabilistic indexing: A statistical technique for document identification and retrieval. Thomson Wooldridge Inc, Los Angeles.
[18] Switzer P (1963) Vector Images in Document Retrieval, Harvard University.
[19] Salton G (1968) Automatic Information Organization and Retrieval. McGraw Hill.
[20] Rocchio JJ (1965) Relevance Feedback in Information Retrieval. Harvard University.
[21] Stevens ME, Giuliano VE, Heilprin LB (1964) Statistical association methods for mechanized documentation. Symposium proceedings, Washington.
[22] Rijsbergen CJV (1979) Information Retrieval, Butterworth-Heinemann Ltd.
[23] Bjorner S, Ardito SC (2003) Online Before the Internet, Part 1: Early Pioneers Tell Their Stories. Searcher: The Magazine for Database Professionals 11(6).
[24] Turtle H (1994) Natural language vs. Boolean query evaluation: A comparison of retrieval performance. Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval, 212-220.
[25] Spärck KJ (1972) A statistical interpretation of term specificity and its application in retrieval. Journal of documentation 28(1): 11-21.
[26] Salton G, Yang CS (1973) On the Specification of Term Values in Automatic Indexing, Department of Computer Science, Cornell University, Ithaca, New York.
[27] Salton G, Wong A, Yang CS (1975) A vector space model for automatic indexing. Communications of the ACM 18(11): 613-620.
[28] Robertson SE (1977) The probability ranking principle in IR. Journal of Documentation 33: 294-304.
[29] Robertson SE, Spärck KJ (1976) Relevance weighting of search terms. Journal of the American Society for Information science 27(3): 129-146.
[30] Salton G, Buckley C (1988) Term-weighting approaches in automatic text retrieval. Information processing & management 24(5): 513-523.
[31] Deerwester S, Dumais ST, Furnas GW, Landauer TK, Harshman R (1990) Indexing by latent semantic analysis. Journal of the American society for information Science 41(6): 391-407.
[32] Harman DK (1993) Overview of the first Text REtrieval Conference (TREC-1). Proceedings of the First Text REtrieval Conference (TREC-1), 1–20.
[33] Kraft DH, Martin MJ, Chen J (2003) Sanchez Rules and fuzzy rules in text: concept, extraction and usage. International Journal of Approximate Reasoning 34: 145-161.
[34] Resnik P (1999) Semantic Similarity in Taxonomy: An Information based Measure and its Application to problems of Ambiguity in Natural Language. Journal of Artificial Intelligence Research 11: 95-130.
[35] Lin D (1998) Automatic Retrieval and Clustering of Similar Words. International Committee on Computational Linguistics and the Association for Computational Linguistics, 768-774.
[36] Fan W, Gordon M, Pathak P. (1999) Automatic generation of a matching function by genetic programming for effective information retrieval. America's Conference on Information System, Milwaukee, USA.
[37] Fan W, Gordon MD, Pathak P (2000) Personalization of search engine services for effective retrieval and knowledge management. Proceedings of International Conference on Information Systems (ICIS), Brisbane, Australia.
[38] Pathak P, Gordon M, Fan W (2000) Effective information retrieval using genetic algorithms based matching functions adaption. Proceedings of 33rd Hawaii International Conference on Science, Hawaii, USA.
[39] Pei J, Han J, Mortazavi AB, Wang J, Pinto H, Chen Q, Dayal U, Hsu M (2004) Mining Sequential Patterns by Pattern growth: the Prefix span Approach. IEEE Transactions on Knowledge and Data Engineering 16(11): 1424-1440.
[40] Li M, Chen X, Li X, Ma B, Paul M, et al. (2004) The Similarity Metric. IEEE Transactions on Information Theory 50(12): 3250-3264.

[41] Sahami M, Heilman T (2006) A Web-based Kernel Function for Measuring the Similarity of Short Text Snippets. 15th International Conference on World Wide Web, 377-386.
[42] Chen H, Lin M, Wei Y (2006) Novel Association Measures using Web Search with Double Checking. International Committee on Computational Linguistics and the Association for Computational Linguistics, 1009-1016.
[43] Cilibrasi R, Vitanyi P (2007) The Google similarity distance. IEEE Transactions on Knowledge and Data Engineering 19(3): 370-383.
[44] Hughes T, Ramage D (2007) Lexical Semantic Relatedness with Random Graph Walks. Conference on Empirical Methods in Natural Language Processing Conference on Computational Natural Language Learning, 581-589.
[45] Zuber V, Faltings B (2007) OSS: A Semantic Similarity Function Based on Hierarchical Ontologies. International Joint Conference on Artificial Intelligence, 551-556.
[46] Korenius T, Laurikkala J, Juhola M (2007) On principal component analysis, cosine and Euclidean measures in information retrieval. Information Science 177: 4893-4905.
[47] Gledson A, Keane J (2008) Using Web-Search Results to Measure Word-Group Similarity. 22nd International Conference on Computational Linguistics, 281-288.
[48] Torra V, Narukawa Y (2008) Word Similarity from dictionaries: Inferring Fuzzy measures and Fuzzy graphs. International Journal of Computational Intelligence Systems 1(1): 19-23.
[49] Bollegala D, Matsuo Y, Ishizuka M (2011) A Web Search Engine-based Approach to Measure Semantic Similarity between Words. IEEE Transactions on Knowledge and Data Engineering 23(7): 977-990.
[50] Chen SJ (2011) Fuzzy information retrieval based on a new similarity measure of generalized fuzzy numbers. Intelligent Automation and Soft Computing 17(4): 465-476.
[51] Usharanim J, Iyakutti K (2013) A Genetic Algorithm based on Cosine Similarity for Relevant Document Retrieval. International Journal of Engineering Research and Technology 2(2).
[52] Maron ME, Kuhns JL (1960) On relevance, probabilistic indexing and information Retrieval. Journal of the Association for Computing Machinery 1(7): 216-244.
[53] Ide E (1971) New experiments in relevance feedback. The SMART Retrieval System, Englewood Cliffs, 337–354.
[54] Rocchio JJ (1971) Relevance feedback in information retrieval. The SMART Retrieval System, Prentice-Hall, Englewood Cliffs, 313–323.
[55] Harper GW, Rijsbergen CJV (1978) An evaluation of feedback in document retrieval using co-occurrence data. Journal of Documentation 3: 189-216.
[56] Lesk ME (1969) Word-Word Associations in Document Retrieval Systems. American Documentation 1: 8–36.
[57] Minker J, Wilson GA, Zimmerman BH (1972) An evaluation of query expansion by the addition of clustered terms for a document retrieval system. Information Storage and Retrieval 6: 329-348.
[58] Doszkocs TE (1978) AID, an Associative Interactive Dictionary for Online Searching. Online Revision 2: 163–174.
[59] Porter MF (1982) Implementing a probabilistic information retrieval system. Information Technology: Research and Development, 2: 131-156.
[60] Xu J, Croft WB (1996) Query expansion using local and global document analysis. Proceedings of the 19th annual international ACM SIGIR conference on research and development in information retrieval, Zurich, Switzerland, 4–11.
[61] Cooper JW, Byrd RJ (1998) OBIWAN—a visual interface for prompted query refinement. Proceedings of the 31st Hawaii international conference on system sciences, Hawaii, 2, 277–285.
[62] Chen H (1998) A machine learning approach to inductive query by examples: an experiment using relevance feedback, ID3, genetic algorithms and simulated annealing. Journal of the American Society for Information Science 49(8): 693-705.
[63] Salton G, Buckley C (1990) Improving retrieval performance by relevance feedback. Journal of the American society for information science 41: 288–297.
[64] Harman DK (1992) Relevance feedback and other query modification techniques. Information Retrieval Data Structures and Algorithms, Prentice Hall, Englewood Cliffs, 241-263.
[65] Horng J, Yeh C (2000) Applying genetic algorithms to query optimization in document retrieval. Information Processing and Management 36: 737-759.
[66] Li WS, Agrawal D (2000) Supporting web query expansion efficiently using multi-granularity indexing and query processing. Journal of Data and Knowledge Engineering 35(3): 239-257.
[67] Wei J, Bressan S, Ooi BC (2000) Mining term association rules for automatic global query expansion: Methodology and preliminary results. Proceedings of the first international conference on web information systems engineering, Hong Kong.
[68] Chen H, Yu JX, Furuse K, Ohbo N (2001) Support IR query refinement by partial keyword set. Proceedings of the second international conference on web information systems engineering, Singapore 1: 245–253.
[69] Takagi T, Tajima M (2001) Query expansion using conceptual fuzzy sets for search engine. Proceedings of the 10th IEEE international conference on fuzzy systems, Melbourne, Australia, 1303-1308.
[70] Kim BM, Kim JY, Kim J (2001) Query term expansion and reweighting using term co-occurrence similarity and fuzzy inference. Proceedings of the joint ninth IFSA world congress and 20th NAFIPS international conference, Vancouver, Canada.
[71] Cui H, Wen JR, Nie JY, Ma WY (2002) Probabilistic query expansion using query logs. Proceedings of the 11th international conference on World Wide Web, Honolulu, Hawaii, 325-332.
[72] Billerbeck B, Scholer F, Williams HE, Zobel J (2003) Query expansion using associated queries. Proceedings of the 12th international conference on information and knowledge management, New Orleans, 2-9.
[73] Chang YC, Chen SM, Liau CJ (2003) A new query expansion method based on fuzzy rules. Proceedings of the seventh joint conference on AI, Fuzzy system, and Grey system, Taipei, Taiwan, Republic of China.
[74] Jin Q, Zhao J, Xu B (2003) Query expansion based on term similarity tree model. Proceedings of the 2003 international conference on natural language processing and knowledge engineering, Beijing, China, 400-406.
[75] Latiri CC, Elloumi S, Chevallet JP, Jaoua A (2003) Extension of fuzzy Galois connection for information retrieval using a fuzzy quantifier. Proceedings of IEEE international conference on computer systems and applications, Tunis, Tunisia.
[76] Nakauchi K, Ishikawa Y, Morikawa H, Aoyama T (2003) Peer-to-peer keyword search using keyword relationship. Proceedings of the 3rd IEEE international symposium on cluster computing and the grid, Tokyo, Japan, 359-366.
[77] Safar B, Kefi H (2003) Domain ontology and Galois lattice structure for query refinement. Proceedings of the 15th IEEE international conference on tools with artificial intelligence, Sacramento, California, 597-601.
[78] Berardi M, Lapi M, Leo P, Malerba D, Marinelli C, et al. (2004) A data mining approach to PubMed query refinement. Proceedings of the 15th international workshop on database and expert systems applications, Zaragoza, Spain, 401-405.
[79] Martin M J, Sanches D, Chamorro J, Serrano JM, Vila MA (2004) Mining web documents to find additional query terms using fuzzy association rules. Fuzzy Sets and Systems 148(1): 85-104.
[80] Stojanovic N (2004) On using query neighborhood for better navigation through a product catalog: SMART approach. Proceedings of the 2004 IEEE international conference on e-Technology, e-Commerce and e-Service, Taipei, Taiwan, 405-412.
[81] Lin HC, Wang LH, Chen SM (2005) A new query expansion method for document retrieval by mining additional query terms. Proceedings of the 2005 International conference on business and information, Hong Kong, China.
[82] Michel B, Annabelle M (2004) An Information Retrieval Model using the Fuzzy Proximity Degree of Term occurrences. SAC'05, Santa De, New Mexico, USA.
[83] Chang YC, Chen SM (2006) A New Query Reweighting Method for Document Retrieval Based on Genetic Algorithms. IEEE Transactions on Evolutionary Computation 10(5): 617-622.
[84] Grootjen FA, Weide TP (2006) Conceptual Query Expansion. Data and Knowledge Engineering 56: 174-193.
[85] Billerbeck B, Zobel J (2006) Efficient query expansion with auxiliary data structures. Information Systems 31: 573-584.
[86] Chang YC, Chen SM, Liau CJ (2007) A new query expansion method for document retrieval based on the inference of fuzzy rules. Journal of Chinese Institute of Engineers 30(3): 511-515.
[87] Nowacka K, Zadrozny S, Kacprzyk J (2008) A new fuzzy logic based information retrieval model. Proceeding of IPMU'08, 1749-1756.

[88] Fattahi R, Wilson CS, Cole F (2008) An alternative approach to natural language query expansion in search engines: Text analysis of non-topical terms in Web documents. Information Processing and Management 44: 1503-1516.
[89] Cecchini RL, Carlos ML, Ana GM, Brignole N (2008) Using genetic algorithms to evolve a population of topical queries. Information Processing and Management 44: 1863-1878.
[90] Carlos ML, Ana GM (2009) A semi-supervised incremental algorithm to automatically formulate topical queries. Information Sciences 179: 1881-1892.
[91] Wasilewski P (2011) Query Expansion by Semantic Modeling of Information Need. Proceedings of International Workshop CS&P.
[92] Liu Z, Natarajan S, Chen Y (2011) Query Expansion based on Clustered Results. Proceedings of the VLDB Endowment, 4(6).
[93] Tayal DK, Sabharwal S, Jain A, Mittal K (2012) Intelligent query expansion for the queries including numerical terms. Proceedings of International Journal of Computer Applications, 35-39.
[94] Latiri C, Haddad H, Hamrouni T (2012) Towards an effective automatic query expansion process using an association rule mining approach. Journal of Intelligent Information System, 209-247.
[95] Yates RB, Berthier R (1999) Modern Information retrieval, Addisson Wesley.
[96] Cooper WS (1988) Getting beyond Boole. Information Processing and Management 24: 243-225.
[97] Salton G (1998) Automatic text processing: the transformation, analysis, and retrieval of information by computer, Addison-Wesley.
[98] Witten I, Moffat A, Bell T (1999) Managing Gigabytes: Compressing and Indexing Documents and Images, Morgan Kaufmann.
[99] Joachims T (1997) A Probabilistic Analysis of the Rocchio Algorithm with TF-IDF for Text Categorization. In Proceedings of the 14th International Conference on Machine Learning, Nashville, Tennessee, USA.
[100] Hull TC (1994) On the mathematics of flat origamis. Congressus Numerantium 100: 215-224.
[101] Soboroff I, Nicholas C (2000) Collaborative Filtering and the Generalized Vector Space Model**.** Proceedings of the 23rd Annual International Conference on Research and Development in Information Retrieval, Athens, Greece.
[102] Qiu Y, Frei HP (1993) Concept based query expansion. Proceedings of the 16th annual international ACM SIGIR conference on Research and development in information retrieval, USA, ACM Press, 160-169.
[103] Xu J (1997) Solving the word mismatch problem through text analysis, Ph.D. Thesis, University of Massachusetts, Department of Computer Science, USA.