International Journal of Engineering Research and General Science Volume 3, Issue 4, Part-2, July-August, 2015
ISSN 2091-2730

Emerging trends and methods used in the field of information retrieval

Sameera Kawatkar

Kawatkar93@gmail.com

ABSTRACT- With the emerging trends in science and technology, the importance of acquiring and finding information from different sources provided by the technology is increasing tremendously. It was possible to store huge amounts of data & information as the computers became more and more advance. All the information available on the internet is made of several documents. Finding useful and relevant information from such documents was required These documents are not always structured and organized in an efficient manner. As a result, the concept of information Retrieval (IR) emerged in the early 1950. Information retrieval finds the most relevant documents according to the inputs given up by the user. Throughout these years this concept of information retrieval has grown & advanced greatly providing easy & efficient access to many users over the internet. This article gives a brief structure of the key technologies used in the field of Information Retrieval & various ways of handling with it.

KEYWORDS: information retrieval, Boolean systems, NLP, cluster hypothesis,

MODELS AND IMPLEMENTATION OF INFORMATION RETRIEVAL SYSTEMS

In the earlier days, IR systems were regarded as the Boolean systems which allowed its users to specify their information need by a complex combination of Boolean operations like ANDs, ORs and NOTs. However there are several shortcomings of the Boolean system. The main drawback is that there is no indication of document ranking, which makes it difficult for the user to select a good search request. Also the concept of relevance ranking is not so critical in a Boolean system. As a result of these drawbacks given up by the Boolean systems, the method of ranked retrieval is now used on a commonly basis. The ranking of the documents is done by information retrieval systems considering the relevance and usefulness of a particular document for which the user had demanded a query. The information retrieval systems assign a numeric score to every document and rank them by this score. Various models have been proposed for this process. The three most used models in IR research are the vector space model, the probabilistic models, and the inference network model.

1) Vector Space Model

Vector space model is a model in information retrieval for representing the text documents. It is formally used for information retrieval, relevance ranking as well as for the filtering of the unwanted information from the required one.
In the vector space model, a document is represented as a *vector* & the text is usually represented by a vector of *terms*. These terms are nothing but typically words and phrases in a particular document. Each dimension in the vector space model corresponds to a separate term. If selected terms are words, then every word becomes an independent dimension in a vector space model. Any text can be represented by a vector in this model. If a particular term belongs to a text, then a non-zero value is given to it in the text-vector along with the dimensions corresponding to the term. Most vector based systems operate in the positive quadrant of the vector space, i.e., no term is assigned a negative value.

For assigning a numeric score to a document for a query, the vector space model measures the similarity between the query vector and the corresponding document vector. The similarity between two vectors is not inherent in the model. Typically, the angle between two vectors is used as a measure of divergence between the vectors, and cosine of the angle is used as the numeric similarity. The inner-product (or dot-product) between two vectors is also commonly used as a similarity measure as an alternative to the cosine angle.

2) Probabilistic Models

In this type of IR models the general principle which is used is that the documents in a collection should be ranked by decreasing probability of their relevance to a specific query given by the user. This is commonly regarded as the "probabilistic ranking principle (PRP)". These Probabilistic IR models estimate the probability of relevance of documents for a query, since the exact probabilities are not available in the Information retrieval systems. This estimation is considered as the key part of the model. Most probabilistic models differ from one another due to this estimation. Maron and Kuhns proposed the initial idea of probabilistic retrieval in a paper published in 1960. Following that, many probabilistic models have been proposed, each based on a different probability estimation technique.

The probability ranking principle can be implemented as follows:

Let *x* be a document in the collection.
Let *R* represent relevance of a document w.r.t. given (fixed) query and let *NR* represent non-relevance.
We need to find p(*R*|*x)* - probability that a retrieved document *x*
is relevant.

$$p(R \mid x) = \frac{p(x \mid R)\,p(R)}{p(x)}$$

$$p(NR \mid x) = \frac{p(x \mid NR)\,p(NR)}{p(x)}$$

Where,
p(*R)*,p(*NR*) : prior probability of retrieving a (non) relevant document.
p(*x*|*R)*, *p(x*|*NR) :* probability that if a relevant (non-relevant) document is retrieved, it is *x.*

Now, according to Ranking Principle (Bayes' Decision Rule):
If p(*R*|*x)* > p(*NR*|*x*) then *x* is relevant, otherwise *x* is not relevant

3) Inference Network Model

In this model of the Information retrieval system, document retrieval is modeled as an inference process in an inference network. Almost all the techniques used by IR systems can be implemented using this model. For the implementation of this model, a document instantiates a term with a certain strength, and the credit from multiple terms is given to a query to compute the equivalent of a numeric score for the document. 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.

QUERY MODIFICATION

In the beginning of IR, researchers concluded that it was often too hard for users to determine effective search requests. As a result, it was thought that adding different synonyms of query words to the query may improve the overall search effectiveness. Early research in the IR solely relied on a thesaurus to find synonyms for the given query. But to obtain an effective general purpose thesaurus is too costly. Due to this hindrance, researchers therefore developed techniques to automatically generate thesauri for use in query modification. There are many automatic methods which are based on analyzing word co-occurrence in the documents. Most query augmentation techniques based on automatically generated thesaurii had very limited success in improving search effectiveness. This was due to the lack of query context in the augmentation process. Not all words related to a query word are meaningful in context of the query.

In 1965 Rocchio proposed a method using relevance feedback for query modification. Relevance feedback was basically designed by the fact that it is easy for users to judge some documents as relevant or non-relevant for their respective query.

By using such relevance judgments, the system is then automatically able to generate a better query for further searching process. The user is told to judge the relevance of the top few documents retrieved by the system. Later, based on these judgments, the system modifies the query and thereby issues a completely new query for finding more relevant documents from the collection. Relevance feedback has thus proved to work quite effectively across test collections.

Many New techniques to do meaningful query expansion in absence of any user feedback were developed in 1990. Most efficient and commonly used of all these is the “pseudo-feedback”, which is regarded as a variant of relevance feedback. Considering that the top few documents retrieved by an IR system mostly fall under the general query topic, we can thus select the most appropriate related terms from these documents which will yield useful new terms irrespective of document relevance. In pseudo-feedback method, the starting few documents in the information retrieval system that are retrieved for the initial user query are said to be “relevant”, and thus performs relevance feedback to generate a new query. As a result, this expanded new query can then be used to rank the documents for presentation to the user. This concept of Pseudo feedback has therefore shown to be a very effective technique, especially for short user queries.

OTHER TECHNIQUES AND APPLICATIONS:

Along with the typical information retrieval models, many other techniques have been developed and have thereby succeeded in their own distinct ways. Following are few of them:

1) **Cluster hypothesis:** It basically deals with separating the information into different appropriate clusters. After analyzing them, we can then conclude that the documents which fall under same cluster have a similar relevance for a given query. Document clustering techniques are still regarded as huge and an active area of research recently.
   If we consider in terms of classification, we can say that if the points fall in the same cluster, they are most likely to be of the same class. There can also be multiple clusters forming a single class.
2) **Natural Language Processing** (NLP): It is has also considered as a tool to enhance retrieval effectiveness.

Natural language processing (NLP) deals with the application of computational models to text or speech data. Automatic (machine) translation between languages; dialogue systems, are the areas that come under NLP. These applications allow a human to interact with a machine using a natural language; and information extraction. Here the goal is to transform unstructured text into structured (database) representations which can further be searched and browsed in flexible ways by the different users handling it. These applications of NLP technologies are therefore having a dramatic impact on the way people interact with computers, with each other through the use of different languages, and also on the way people access the vast amount of information and data over the internet.

CONCLUSION

From the above described advanced techniques and developments in the field of information retrieval , we can thus say that it has improved tremendously in a positive manner, and made information discovery more easier and thereby faster. The task of finding information from the documents can be effectively done using these statistical techniques. These techniques developed in the field of information retrieval are being used in many other everyday used areas as well. For example; junk-email filters, web search engines, news clipping services. With the amount of information made available on the internet increasing exponentially, information retrieval is definitely going to be a boon for huge number of users in the future.

REFERENCES:

1. www.google.com
2. www.wikipedia.com
3. www.researchbible.com
4. Books from various subjects