REVIEW OF RESEARCH
ISSN: 2249-894X
IMPACT FACTOR : 5.2331(UIF)
VOLUME - 7 | ISSUE - 3 | DECEMBER – 2017

DIGITAL AGE, INFORMATION EXPLOSION AND INFORMATION RETRIEVAL: A REVIEW

Neeraj Tanwar
Research Scholar , Department of Library and Information Science,
Maharshi Dayanand University, Rohtak.

ABSTRACT:

In today's age of information explosion disorganized and unsystematic use of information can cause poor judgement which can reduce performance of professionals in their work areas. This article focuses on some issues related to effective information retrieval and discusses the different techniques for information retrieval.

KEYWORDS: *Information explosion, information retrieval, retrieval models.*

INTRODUCTION:

Information has grown exponentially in the last few decades. This information explosion is the result of free information market backed with the progress in information and communication technology. The information is readily available in huge masses in our society, and it is expected to increase with same pace in near future. Earlier, more information was considered as a very good thing. As information has played a vital role in all round development of our society. It was the main driving force behind the major developments in the society. But today, increasing information is seen as a problem. Apart from this increase in information, relevancy of information and available documents has also declined comparatively. We cannot readily assimilate all the information available to us. This phenomenon is referred to as "Technostress" (Carlson, 2003). Disorganized and unsystematic use of this information causes poor judgement which in turn can reduce intellectual performance.

Recent advancements in information and communication technologies (ICT) address the problems which are growing from information explosion. Many aids, such as intelligent agents, ranking algorithms, cluster analyses techniques, data mining and web mining techniques, web graph algorithms, and personalization and collaborative techniques, are available for retrieving the relevant and organized information from the information available. These aids can be integrated into the search engines which can support different retrieval techniques. Although these techniques help lessening the problem but they do not solve the problem completely.

Growing need of users for specialized information and information explosion in society have directed the information scientists towards development of more efficient information retrieval techniques. Large volumes of data is stored and managed with the creation of databases and edification of indexing techniques. Therefore, for the last few decades, information scientists have focused on design and development of more powerful and efficient search and retrieval systems for information needs of users(Naik & Rao, 2011).

INFORMATION SEARCH AND RETRIEVAL

Information searching and retrieval is the significant feature of library systems. The search facilitates quick retrieval of information by helping users to select relevant information. Search service in libraries covers searching information in multiple formats such as text, images, audio, video etc. Search service is better helpful to users when implemented with a flexible user interface, for metadata support and cross database searching without query modification, and a powerful retrieval engine (Naik & Rao, 2011).

Information Retrieval is the process of storing information in database, searching that stored information, and retrieving information according to the need and request of users. With the increasing use of internet and communication technologies, information retrieval has become very crucial. Users, either individual or group, have desire to find and get information to satisfy their conscious and unconscious needs. To find information required by them, users enquire the retrieval systems in form of queries and in turn system returns a set of matching results. These results are evaluated by users in terms of relevancy and level of satisfaction. The retrieval formats should be flexible according to users' requests. It should allow users to manipulate search process by modifying search strategies, editing results and choosing preferred delivery formats. Many libraries provide users with different alternatives of search based on metadata by searching across the databases of multiple libraries simultaneously in a single search. Information retrieval systems are evaluated based on the effectiveness of results.

METHODS OF INFORMATION RETRIEVAL

Classical information retrieval systems followed the monolingual approach i.e. both the query and the documents retrieved are to be in the same language. But in new era information needs of users have changed which in turn has changed the retrieval methods. Modern information retrieval systems must take into account all the documents relevant to user need regardless of the language being used. To retrieve documents in multiple languages, various translation methods are required. Translation can be done in three forms i.e. Query Translation, Document Translation or both. Query translation is the process of translation of the query to the target language whereas document translation involves translating retrieved documents to the source language or language used by the user.

There are three main tools used for translation i.e. dictionary translation, machine translation and corpus based translation. In *Dictionary based translation*, a bilingual dictionary of source language and its translation in target language is used. Use of dictionary tool removes ambiguity and allows assigning weights to the documents for specifying level of relevancy. In *Machine Translation*, a machine translation system is used to translate either the query or the documents. A *parallel corpus* is a collection of texts in one language and their translations into a set of languages. Generally, it contains data from two languages. Although the creation of parallel corpora is very complex and costly but it gives better performance than dictionary based translation (Ren & Bracewell, 2009).

Dictionary and corpus based translations are used in query translation whereas for document translation machine translation is used mostly. The main issues in both corpus and dictionary based translation is the quality and coverage. It is very difficult to find corpus and dictionaries which covers all the related words which are enough for utility in translation. In machine translation main issue is computational expensiveness. Machine translation is unfeasible where large collection of documents is involved.

Taxonomy of Information Retrieval Models

Every information retrieval system uses a suitable representation for effective retrieval the relevant documents. These representations are based on the user needs and lead to a framework for models. For example documents are represented in form of words, called as indexing terms, in text based information retrieval. Every retrieval strategy combines with a specific model of document representation. Information retrieval models are categorized on the two basis, i.e., mathematical basis and properties of the model. Figure-1 shows the different types of models:

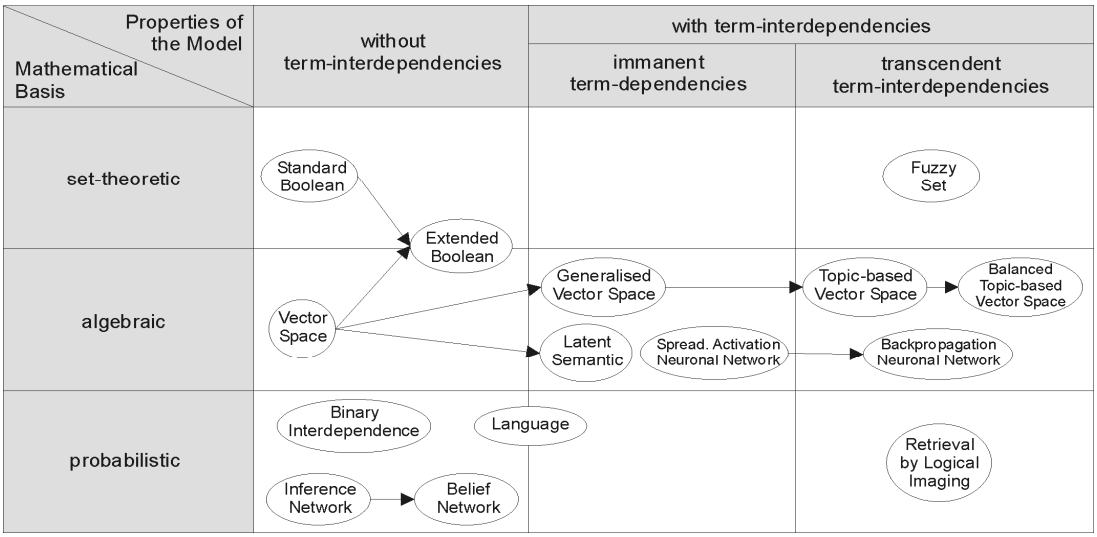

Figure-1: Information Retrieval Models (Source: Wikipedia)

Classification on Mathematical Basis

In ***Set-theoretic models***, documents are represented as set of words or indexing terms. Documents are retrieved by using the set-theoretic operations. Following are some set-theoretic models:

- **Standard Boolean Model:** The very first and most adopted model for information retrieval is the Boolean Model of Information Retrieval. Many information retrieval systems still use this model for document retrieval. Boolean model is based on the Boolean logic and classical set theory operations. Both the user's query and documents are considered as a set of terms. It matches the query terms with the documents using Boolean logic and retrieves the documents containing the query terms. One of the disadvantages of Boolean model is that all the terms are equally weighted i.e. all the documents containing query terms are considered as equally relevant.
- **Extended Boolean Model:** This model overcomes the drawbacks of the Boolean model. It combines the features of both Vector Space model and Boolean algebra. It provides ranks to the similarity between query terms and documents. Thus it is considered as a generalization of Boolean model and Vector Space model. Research has shown that Extended Boolean model is more effective than the Boolean query processing.
- **Fuzzy Retrieval:** Fuzzy retrieval model consists of the properties of the Fuzzy Set Theory and along with the properties of Extended Boolean model. There are two classical fuzzy retrieval models: Mixed Min and Max (MMM) and the Paice model.

  In ***Algebraic models***, documents and queries are represented as vectors, matrices, or records. Query vector is matched with the document vector for similarity to retrieve information required by users and the result is represented as a scalar value. Common algebraic models are listed here:
- **Vector Space Model:** It is also called as *Term Vector Model.* It is used in indexing, filtering, relevancy ranking and retrieval of information. It represents the text documents as vectors of identifiers such as indexing terms. Terms are the single words, keywords or phrases. There are two types of vectors of terms used in this model i.e. document vector and query vector. If a term exists in the document, a non-zero value is assigned to it in the document vector. For computing these vector values different methods are used.
- **Generalized Vector Space Model:** It provides an extension to the vector space model.(Wong, Ziarko, Raghavan, & Wong, 1987) in their Generalised Vector Space Model addressed the issue of pair wise orthogonality in Vector Space Model. Unit vectors are considered pair wise orthogonal if they are dissimilar. Generalised Vector Space Model introduces term correlation rather than pair wise

orthogonality between the document vectors and query vectors. Generalised Vector Space Model introduces term to term correlations, which expresses a strong disapproval of the pair wise orthogonality assumption of vector space model. Each term vectort_i is expressed in 2^n linear combinations of vector m_r where r varies from 1 to 2^n.

- **(Enhanced) Topic-based Vector Space Model:** The Topic-Based Vector Space Model (TVSM) is an extension to the vector space model. It also removes the constraint of orthogonality between the term vectors. In case of Natural Languages, orthogonality causes problems with the related terms such as synonyms and similar words. TVSM allows the use of thesaurus. In TVSM the fundamental topics are represented using positive real numbers i.e. R+ in a d dimensional space represented using positive natural numbers i.e. N+. The term vector *t* specifies the weights for R. Relevant and more important terms have a higher weight than the stop-words and the irrelevant terms. Terms in the documents are represented as the sum of the term vectors. The scalar product of the document vectors defines the similarity between the two documents.
- **Latent Semantic Indexing (Latent Semantic Analysis):** This model uses mathematical technique called as Singular Value Decomposition (SVD) for indexing and retrieval. In this technique the unstructured pool of text is analysed for identifying patterns in relationships between terms and concepts. It identifies the hidden semantics of the words used in text and returns the results that have conceptually similar meaning to the query terms. This method is also called as Latent Semantic Analysis (LSA).

Probabilistic models use probabilistic inference methods for retrieval of documents. For finding relevant documents probability method is used to compute the similarity. Some of the probabilistic models are discussed here below:

- **Binary Independence Model:** In the Binary Independence Model the documents are represented in the form of binary vectors. It assumes that while storing documents, a binary vector is recorded which show that either the terms are present or absent in a particular document. These terms are independently distributed among the relevant and irrelevant documents i.e. it is assumed that there is no association between the terms. For each term in the document or query, term vector consist of one Boolean element. The term vector is represented as an ordered set of these Boolean elements. Queries are also represented in the same way.
- **Probabilistic Relevance Model:** This model was given by Robertson and Jones. It gives a method or ranking the searched documents according to their relevance to a given query. It is used by search engines and web search engines for ranking. This model estimates the probability of relevancy between the query and the documents. It also assumes that for a query *q* there exists an ideal answer set *R* of documents that have maximum probability of relevance i.e. all the documents present in set *R* are relevant to the query and that all the irrelevant documents are not present in the answer set *R*.
- **Uncertain Inference:** This model was given by C.J. van Rijsbergen. He gave a method to formally define the relationship or query and document along with the measure of uncertainty. He proposed that the probability of logical implication of a document *d* to the query *q* be the measure of uncertainty among them i.e. *P(d→q).*

  A user's query is assumed to be a set of assertions about the document. The system infers whether the query assertions are true or not and retrieves the documents if the assertion is true. The *(d→q)* is uncertain because both d and q are generated by users and hence are error prone. So instead of retrieving exactly matching documents, the documents are ranked based on their credibility regarding to the query.
- **Language Models:** A language model is a statistical model which uses the probability distribution over a sequence of words. Language modelling is used in natural language processing, speech recognition, handwriting recognition, machine translation, information retrieval and many other applications. Some of the language models are:
  - Unigram Models
  - N-gram Models
  - Exponential Language Models
  - Neural Language Models

Classification Based on Models' Properties

- ***Without Term-Interdependencies:*** In these models, all the different terms are treated independent to each other. These models assume that there is no association between the terms. Term independences is represented in the vector space models and in probabilistic models through orthogonality and independency assumption respectively.
- ***With Immanent Term Interdependencies*****:** In these models association between terms is represented. Model itself define the interdependency between the terms by using occurrence of terms in the documents.
- ***With Transcendent Term Interdependencies***: In these models, interdependencies between terms is represented but these models do not define those interdependencies. The degree of interdependency is defined by an external source such as an algorithm etc.

Evaluation of Information Retrieval Systems

Evaluation is the process of measurement of the performance of the system that to how much extent a system fulfils the users' requirements and to how much extent the retrieval results satisfied the users' information needs. Evaluation metrics can be classified into two categories:

- **Online metrics**
  - *Session abandonment rate*: It is the ratio of search sessions in which no result was produced.
  - *Click-through rate*: It is the ratio of number of users who clicked on a particular link to the total number of users who visited a webpage, email, or advertisement.
  - *Session success rate*: It is the ratio of sessions in which users' search was successful.
  - *Zero result rate (ZRR)*: It is the ratio of Search Engine Results Page (SERP) which returned with zero results.

- **Offline metrics**
  - *Precision*: The ratio of the total number of relevant documents in a search result to the total documents retrieved is called as the Precision.
  - *Recall*: It is the ratio of the retrieved documents that are relevant to the query to the total relevant documents available.
  - *Fall-out*: The ratio of non-relevant retrieved documents to the all available non-relevant documents is called as Fall-out ratio.
  - *F-score / F-measure*: The weighted harmonic mean of precision and recall is called as F-score.
  - *Average precision*: Precision and recall are single-value metrics but for retrieving a ranked sequence of documents, it is desirable to also consider the order in which the returned documents are presented. Average precision computes the average value of *p(r)* over the interval from *r=0* to *r=1.*
  - *Precision at K*: Precision at k documents shows the number of results relevant to the user's query on the first *k* pages of search result.
  - *R-Precision*: R is used as the cut-off to find out the relevancy fraction. The system looks for the top R documents retrieved that are relevant r and gives a relevancy fraction.
  - *Mean average precision*: For a set of queries, the mean of the average precision score of each query is called as the Mean Average Precision.
  - *Discounted cumulative gain*: It is used for the quality ranking of the search engine results. It is used to measure effectiveness of the search algorithms of a web search engine.
  - Other measures
    - Mean reciprocal rank
    - Spearman's rank correlation coefficient
    - bpref - a summation-based measure of how many relevant documents are ranked before irrelevant documents

- GMAP - geometric mean of (per-topic) average precision
- Measures based on marginal relevance and document diversity
- Measures of both relevance and credibility
- Visualization: It includes
- Graphs which chart precision on one axis and recall on the other
- Histograms of average precision over various topics
- Receiver operating characteristic (ROC curve)
- Confusion matrix

User Empowerment

One of the most important questions today about such vast information available in the society is the retrieval of relevant information. Availability of huge amount of information does not mean that any information required by user is available. (Carlson, 2003)said that there has been a substantial growth in database of most common search engines and at the same time number of searchable pages have also increased proportionally.

In this scenario, users' skill of information retrieval are limited and users can miss the most important and relevant documents available in the databases. Therefore it is necessary for users to understand the general working of information retrieval systems. As discussed before, different information retrieval models work differently to optimize the search results. Search engines greatly augment to the skills of user for information retrieval empowering users to search in more efficient and effective manner. Along with searching skills, users must be able to modify their queries to get better results according to their need and expectations.

In addition to searching skills, users also need to enhance their skills to evaluate the results retrieved by the information retrieval systems. Information professionals have developed some measures that can be used to evaluate the relevancy of retrieved information resources. Some of these measure are required format of documents, scope or coverage of document for particular information, relationship between other resources to find overlapping content, author's or creator's background to check authenticity of the content and cost etc. Users can use these measures to evaluate the retrieved information and resources.

Information literacy programs for library users and library professionals should be organized by the libraries to enhance the information retrieval skills among them. According to American Library Association "*Information Literacy is a set of abilities requiring individuals to recognize when information is needed and have the ability to locate, evaluate, and use effectively the needed information*". Information literate users are able to recognize the need and context of information and can locate and find the required information effectively and can use the gathered information efficiently to solve the purpose.

Information Retrieval and Libraries

A digital library serves as a rich source of information for the society with its organized and digitized data resources. One of the most important component of digital libraries is information retrieval system which is allows storage, processing, retrieval and maintenance of information(Sharma, 2007). Success of a digital library rely on an effective information retrieval system which provides easy access and quality retrieval of information required by users. It helps users in finding information required by them. The success of an information system can be measured by how much it reduces the user efforts in finding their required information.

Digital libraries present new environment and challenges for implementation of information retrieval systems because of nature of information contained by digital libraries and tasks performed by these libraries. Digital libraries consist of different type of data such as text, images, audio and video information. Information retrieval in such heterogeneous environment is not studied that much. More studies have been conducted on textual information retrieval. Another issue with retrieval in digital libraries is the distributed environment. Digital libraries provide a distributed structure. While searching and retrieval in distributed environment the problem of merging the results from different sites is an issue because different sites might use different datasets, data and metadata(Pal & Pal, 2007).

CONCLUSION

In a country like India having a large readers'/users' community, there have always been challenges before the librarians and other library staff to satisfy the needs of users. There is a lot of scope for the librarians to take advantage of new technologies to improve their skills in information management. Library users are gradually moving towards the Digital library systems for more because of characteristics such as location independence, accuracy in retrieval and for saving their time. Today users want information delivered on their computer systems. Issues and challenges such as heterogeneous information retrieval and information retrieval in distributed environment also need to be tackled. Digital libraries provide excellent opportunities not only to the users but also for library professionals and other professionals such as authors, publishers and information retrieval systems are the most important part of the development of digital libraries.

REFERENCES

Carlson, C. N. (2003). Information Overload , Retrieval Strategies and Internet User Empowerment. In *The Good, the Bad and the Irrelevant (COST 269)* (Vol. 1, pp. 169–173). Helsinki (Finland): Media Lab UIAH. Retrieved from http://eprints.rclis.org/5432/1/Information_Overload.pdf%0A

Naik, N. R., & Rao, A. M. (2011). Information Search and Retrieval System in Libraries. In *8th International CALIBER - 2011* (pp. 16–27). Goa: INFLIBNET Centre, Ahmedabad. Retrieved from http://hdl.handle.net/1944/1596

Pal, J. K., & Pal, F. (2007). Search Algorithms – an Aid To Information Retrieval in Digital Libraries. In *5th International CALIBER -2007* (pp. 401–414). Chandigarh (India): INFLIBNET Centre, Ahmedabad. Retrieved from http://hdl.handle.net/1944/583

Ren, F., & Bracewell, D. B. (2009). Advanced Information Retrieval. *Electronic Notes in Theoretical Computer Science*, *225*(C), 303–317. http://doi.org/10.1016/j.entcs.2008.12.082

Sharma, J. C. (2007). Digital Information Management and Retrieval System. In *5th Convention PLANNER - 2007, Gauhati University, Guwahati* (pp. 375–377). Guwahati, Assam: INFLIBNET Centre, Ahmedabad. Retrieved from http://hdl.handle.net/1944/1364

Wong, S. K. M., Ziarko, W., Raghavan, V. V., & Wong, P. C. N. (1987). On modeling of information retrieval concepts in vector spaces. *ACM Transactions on Database Systems*, *12*(2), 299–321. http://doi.org/10.1145/22952.22957