# Information Search and Retrieval System in Libraries

**N Rupsing Naik** **A Madhava Rao**

## Abstract

A digital library comprises diverse collections of digital objects representing text, sound, maps, videos, photos, etc. and a working environment, technology and services. The main objective of any Digital Library (DL) is to fulfil the needs of its users. A general problem for a user is information search and retrieval in the Internet world. This paper discusses the information search and retrieval system its models and uses in digital libraries.

**Keywords:** Information Retrieval System, Information Retrieval Model, Digital Library

## 1. Introduction

The explosion of literature in the form of micro documents on the one hand and the growing number of users demanding more specialised literature/information on the other hand have led to information scientists to develop an efficient information retrieval systems for the realisation of retrieval suspend on advances in technology and its associated techniques. Database creation of the library resources and the sophistication in indexing techniques has eased the problems of storing and handling of large volumes of data. At the best, it enhanced the retrieval of the items. Therefore, the focus of the information scientist for the recent past few decades is on the design and development of more powerful information search and retrieval systems.

## 2. Information Systems

Computer based information system is categorised into:

i. Information Storage and Retrieval systems
ii. Database Management systems
iii. Management Information Systems
iv. Decision support systems

All these systems exhibit similarities to some extent in the area of information processing but differ in their functionalities.

## 3. Objectives of the Information Retrieval System

According to M.L. Pao (1980), "user's input is an important consideration to be incorporated in setting the overall objective of the system for service point of view", as follows:

i. Information content of information resources collected
ii. Utility of information resources

iii. Users

iv. Documentary resources

v. Performance resources

vi. Economics

Information retrieval (IR) is the main purpose of any library. The librarian is a nodal point in the IR process and in the traditional library.

## 4. Information Search and Retrieval in Digital Libraries

Information search and retrieval of an object from digital library software is a vital feature of the system. The search enables quick retrieval of information. Search services help users to select relevant information from digital library. Digital library's service provides fast access to exact information which is looking for. The success of a search service in digital library relies on the implementation of a powerful retrieval engine and a flexible user interface for metadata support. The Search interface allows users to do "across database" searching without having to modify a query. Search service also covers searching beyond text to multiple media formats, including images, sound and video. The retrieval formats should be flexible and should provide users to manipulate the search process and results by retrieving search history, adjusting search strategies, editing and sorting search results and choosing preferable delivery formats. Users should also be able to get statistical analysis of the searches they have carried out.

Many digital libraries provide different search options to users based on the metadata fields along with the facilities for federated search across a number of digital libraries. Most digital libraries offer search by Boolean operator, keyword, and phrase and field searches.

In the case of information retrieval, evaluation is often focused on the effectiveness of a result set in a specific search. Browsing and searching are two major paradigms for exploring digital libraries. Boolean, proximity and truncation searching are commonly used in digital libraries. They are often provided as separate services. Searching is popular because it is useful when appropriate search keyword are unavailable to users. Table 1 presents the difference between data retrieval and information retrieval.

**Table 1: Data Retrieval vs. Information Retrieval [Rijsbergen, 1979]**

| | **Data Retrieval (DR)** | **Information Retrieval (IR)** |
|---|---|---|
| Matching | Exact match | Partial match, best match |
| Inference | Deduction | Induction |
| Model | Deterministic | Probabilistic |
| Classification | Monothetic | Polythetic |
| Query language | Artificial | Natural |
| Query specification | Complete | Incomplete |
| Items wanted | Matching | Relevant |
| Error response | Sensitive | Insensitive |

The use of Information Retrieval is motivated by an information need. This information need can be explicitly or implicitly verbalized. In a 'real world' setting the person seeking information (i.e. the user) formulates such a question and poses it to an expert. The expert calls upon his internal representation of the knowledge space and external documents and formulates answers. From the answers received the user extracts relevant points and gives feedback to the expert.

This "conversational loop" can also be found in the use of an Information Retrieval System (IRS). As the IRS is not capable to understand the information need, thus, an abstraction matching in the IRS is needed. This abstraction is called query. Analogously to the expert, the IRS formulates an answer based upon the internal representation of the knowledge space and external documents. The answer is composed of documents perceived relevant or links to such documents. The user extracts those documents that are indeed relevant. In some systems relevance feedback can be given. These two forms of the conversational loop are detailed in Figure 2. The following sections will discuss in detail about information retrieval and browsing aspects:

### 4.1 Retrieval vs. Browsing

The sequence of action taken to satisfy the information need as described above is called retrieval or searching. Retrieval is used in the case of an explicit information need. The explicit need can be formulated into a query. Searching usually results in lists of results. Sometimes the information need is non-explicit and no query can be formulated. In this case the information need can be satisfied by browsing through the documents of a collection. Browsing is also necessary to find relevant documents from the results of a retrieval process. Figure 3 illustrates the relation of these two concepts.

Both actions described above are pulling actions interpreting to the user makes request of information interactively. Searching for information in such a manner is called ad hoc searching. Alternatively the information may be provided in an automatic and permanent fashion for information by the user. This is called information push and is used in the case of a varying document collection and an unchanging information need. In the same context as information push, filtering can be applied. Again the document collection is varying while the information need stays the same, but in contrast information push is an active task of the user, while in the case of filtering the information is passively provided by software agents or the like.

Browsing can be discerned into flat browsing, structure guided browsing and hypertext browsing. In the case of "flat browsing" the document collection is organized in a flat structure, like a list of search results. The term "structure guided" browsing describes the browsing of collections in structured and hierarchical manner. The "king" of browsing allowing the highest flexibility is "hypertext browsing". In this case, documents are multiplied connected by hyperlinks. The Information Retrieval Model (IRM) needs introduction to the digital library for describing relevancy in the search results. Such a model describes the fundamental premises forming the basis for a ranking. Over the years, different IRM have been proposed. The following sections will give an overview of those models which are relevant in the later part of this paper. The most important Information Retrieval Model is depicted in figure 1 through 3.

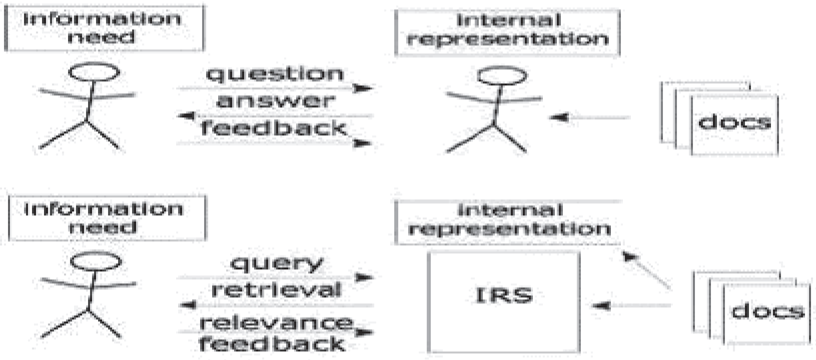

**Figure 1: Conversational loop, after [Source: Gutl, 2005]**

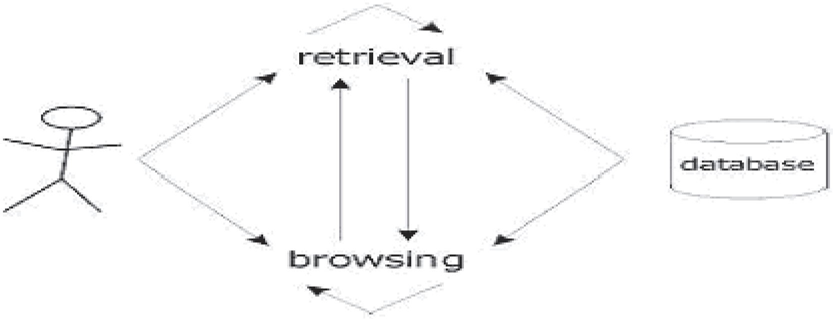

**Figure 2: Interaction with the retrieval system through different tasks, after [Source: Baeza-Yates et al., 1999]**

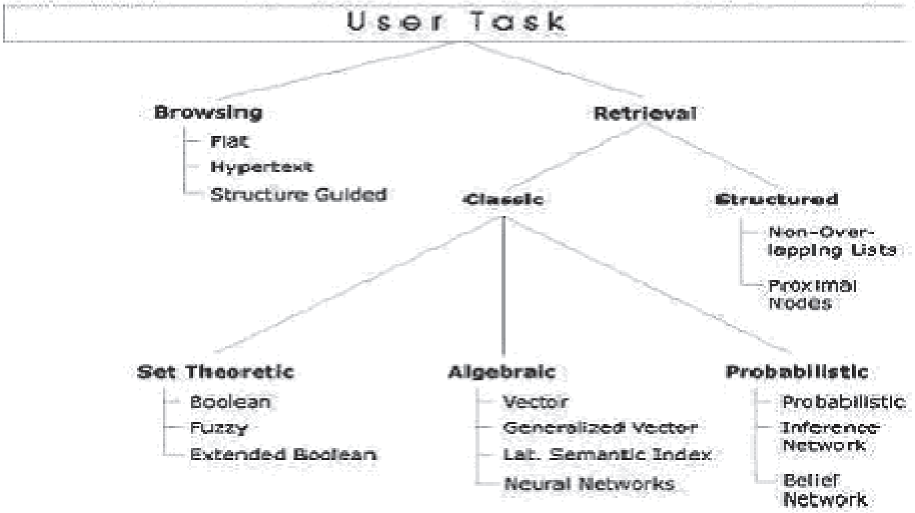

**Figure 3: A taxonomy of information retrieval models, after [Source: Baeza-Yates et al., 1999]**

In order to build a model, a representation for the documents and its need for the users have to be found. These representations lead to the framework in which they can be modeled. For example, in the vector space model, the framework is composed of a t-dimensional vector space and standard linear algebra operations on vectors.

In text based IR the information need and the documents are represented with words. In case of the documents these are called index terms. An index term is a word whose semantics matches the document's main themes. These index terms vary in relevance, as the more frequent words in the document collection are less relevant for the retrieval process.

### 4.1.1 Boolean Model

This simple retrieval model, which was adopted by many early commercial bibliographic systems, is based on set theory and Boolean algebra. The advantages are the intuitiveness of the concept of a set and the precise semantics of Boolean expressions, which form the queries. The major drawback of the model is that the retrieval is based on a binary decision leaving relevancy of the search result. Moreover it is often not simple to translate an information need into a Boolean expression.

In this model, index terms are either considered present or absent in a document, resulting in binary weights. Due to this binary value for relevancy, no partial match to the query is defined. So, if a document includes only one index term it is considered not relevant. As a result, the Boolean model often retrieves very less or too many documents.

### 4.1.2 Vector Model

The vector model heeds the fact that the use of binary weights is limiting and proposes a framework that allows a partial match. Non binary term weights are used to compute a degree of similarity. The resulting set of documents retrieved is returned in decreasing order of this degree of similarity.

Though various term weighting techniques exist, only the main idea of the most effective techniques shall be discussed. The basic idea is to separate a document collection into two parts to satisfy the information need. One of the parts is composed of the objects related to the information need while the other is not.

To accomplish this separation clustering techniques are utilized. Two sets of features are used to discern sets of related elements from those that are not. The first set of features describes the intra-cluster similarity while the second describes inter-cluster dissimilarity.

The improved retrieval performance resulting from the term-weighting scheme is one of the main advantages of the vector model. The second main advantage is the partial matching strategy and the fact that the cosine ranking function sorts the documents according to the degree of similarity to the query. Efficient implementations for the vector model are possible. Finally, the vector model allows easy relevance feedback.

Disadvantages of the vector model include the fact that the term "independency" is not fully given. In fact, real term independency might hurt the retrieval process. Moreover, syntactic information remains unconsidered.

### 4.1.3 Generalized Vector Space Model

The term "independence" in the classic vector space model is addressed by the generalized vector space model. The independence is interpreted as pair-wise orthogonality among the index term vectors that forms the vector space. Wong et al., (1985) proposed an alternate view which leads to the generalized vector space model. In this model, the index term vectors are assumed linearly independent but are not pair wise orthogonal. As such, they are not as like in the classic vector space model where the vectors compose the basis of the space. They are composed of smaller components derived from the collection. These pair wise orthogonal so called min-term vectors compose the bases of the space.

The main advantage of the generalized vector space model is the dependence of the index term. However, this dependence is still a controversial issue. Thus, the advantage of the generalized vector space model in practical situations is not yet proved. The main drawback is the high cost of computation due to the fact that the number of min-terms might be proportional to the number of documents in the collection.

### 4.1.4 Probabilistic Model

The third of the classic IR models, the probabilistic model, attempts to capture the IR problem within a probabilistic framework. To that end, a set of documents is defined for each user query, which contains only the relevant documents. This set is referred to as the ideal answer set. If the properties of this ideal answer would be known, the documents could be retrieved. The querying process can be described as the process of specifying these initially unknown properties.

The main advantage of the probabilistic model is the ranking by the probability to be relevant. The main disadvantages include the need to guess the initial separation into relevant and non-relevant documents. Moreover, the weights for the index terms are binary. Finally, as in the classic vector space model, the index terms are assumed to be independent. Although experiment exist which show better performance of the probabilistic model when compared to the vector space model, it can be expected that it is outperformed by the latter with general collections.

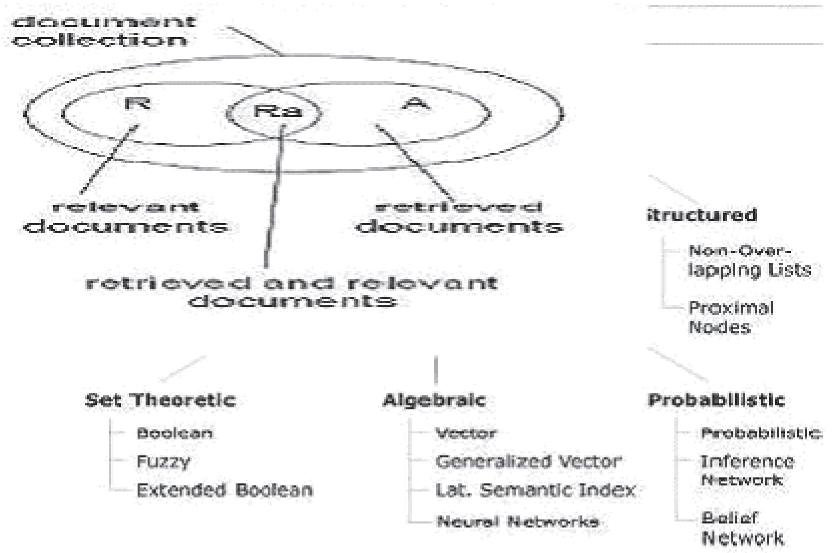

**Figure 4: Precision and recall for a given example information request after [Source: Baeza-Yates et al., 1999].**

## 5. Information Retrieval Concepts in Digital Libraries

Digital libraries rely on effective retrieval methods with easy access to the information. Thus, the success of digital libraries is depends on the quality of retrieval. Accordingly, research in the IR has traditionally been important in the research pertaining to the digital libraries.

One important topic in this context is the search in distributed collections. As the document collection is managed by different organizational units or even the document collections of different digital libraries are used, the interfaces and functionality of these systems are supposedly inhomogeneous. According to Gutl (2004) part of the difficulties in this context are the variable IR methods and interfaces, like representation and relevance ranking, as well as the scope of the search results. Another difficulty is the merging of the individual results and the inter-system ranking. The second important topic in this context is the multimedia content in document collection. Baeza-Yates et al., (1999) states the need for suiting query languages for different media, like a visual query for image search. In a multimedia environment, the task is even more complex, as the combined search in different media, as Gutl (2004) points out.

### 5.1 Browsing Features

The browsing feature is important for retrieval of neither information that is nor already known in part. Browsing enables people to look through a digital library and discover things that they had no previous knowledge of. Users should be able to browse digital objects by:

i. Author / Creator / Contributor

ii. Title of the document / article / book

iii. Issue Date / Date of Publication

iv. Collection

v. Communities

vi. Subject browsing

vii. Publisher wise browsing

viii. Table of contents browsing

ix. Multi-dimensional browsing

### 5.2 Searching Features:

Searching in any digital library is one of the important aspects. Hence, it is necessary to know what type of search features are supported by the software. While, evaluating the software it is necessary to do functional testing of the software, i.e., determining the extent to which a digital library, in whole or in part, is able to perform desired operations.

i. Full text searching

ii. Boolean (AND, OR, NOT) searching

iii. Basic Search

iv. Advanced search

v. Truncation/Wild card searching

vi. Exact words/phrases searching

vii. Proximity searching

viii. Stemming search

ix. Fuzzy search

x. Phonetic search

xi. Case sensitive

xii. Case insensitive

xiii. Boosting the term

xiv. Range searching

xv. Expand search

xvi. Lateral search

xvii. Multilingual search

xviii. Refine search

### 5.3 Quality of Search Results

Generally IR systems have to cope with a vaguely described information need of the user; the results of an IR process are not exactly matches to this information needs, but are ranked by relevance. The evaluation of the precision of this answer is called information retrieval evaluation. Besides, the performance measures generally important for a software system, like response time, space, and the like, this retrieval performance is the key to an IR system.

Information retrieval evaluation is performed by querying a standardized reference collection. These reference collections consist of a set of documents, a set of example information needs, and corresponding sets of relevant documents. The relevant documents for example information requests are determined by experts.

For a given retrieval strategy, the documents retrieved are compared to the set of relevant documents determined by experts. The similarity between these two sets is quantified by the test collection's evaluation measure and leads to the goodness of the tested retrieval strategy.

### 5.4 Recall and Precision

Recall and precision are basic evaluation parameters. The drawback of recall and precision is the fact that the examination of all documents of the answer set is assumed. This, however, is contrary to the usual fact that the user is presented only a part of the retrieved documents, ranked by the degree of relevance. Thus, the recall and precision values vary as the user proceeds through his examination of the retrieved documents. In order to record this development, a precision versus recall curve is plotted.

Usually recall and precision values resulting from averaging various queries are used to compare the performance of IR systems. Such average recall versus precision plots trade the advantage of a better overview on the overall retrieval performance for that fact that single mavericks remain undetected.

In various situations, such as, to discover the superiority between the single value summary of the recall and precision plots. One possibility is to calculate the Average Precision at Seen Relevant Documents. To that end the precision figures obtained are averaged after each new relevant document which is observed.

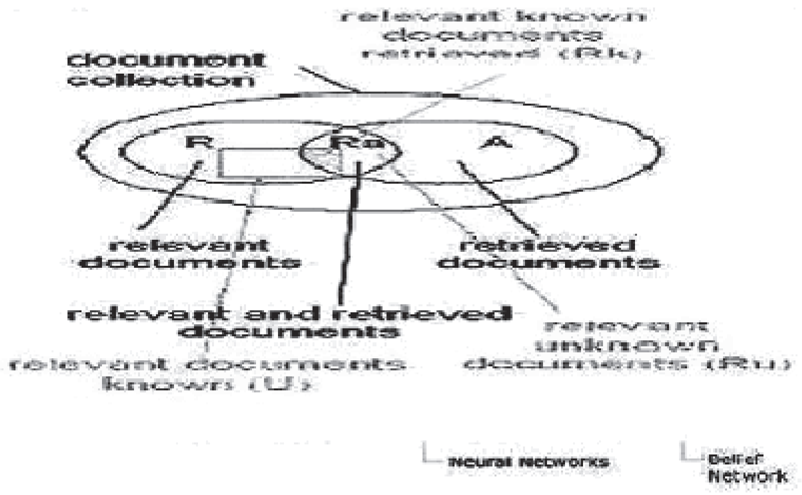
relevant known documents retrieved (Rk)
document collection
R
Ra
A
relevant documents
retrieved documents
relevant and retrieved documents
relevant unknown documents (Ru)
relevant documents known (U)
Neural Networks
Belief Network

**Figure 5: Coverage and novelty ratios for a given example information request after [Source: Baeza- Yates et al., 1999].**

## 6. Conclusion

Digital libraries present still another new environment for information retrieval, presenting new and different challenges and an expanded research agenda. Some of these challenges arise from the nature of the content in digital libraries, others from the nature of the tasks performed and the characteristics of the users of digital libraries.

Like the Web, digital libraries incorporate mixed data types. The data may be structured, semi structured, or unstructured; and incorporate text, images, video, and audio information. Information retrieval from this mix of structure and formats is relatively unstudied, since research has usually been based on an assumption of a homogenous collection, and metadata, where available, has been treated as unstructured text. How do we incorporate evidence from these multiple sources to create an ordered list? Since digital libraries are by definition often distributed or federated systems, another level of complexity is added by the need to make retrieval from multiple sites and multiple collections transparent to the user. Given multiple sites, we need to give priority for search to sites with the highest probability of success. Searching on multiple sites leads to a data fusion problem as the system must integrate and rank information from different datasets, with different data and metadata. Furthermore the importance and challenges of distributed document collections have been analyzed. Though the Web may appear as a distributed digital library at the first glance, there are numerous differences. Information Retrieval system which presents the basic layer applied in conceptualization processes and discussed the models of the IR system.

## References

1. GUINCHAT, Claire & MICHEL Menou. (1983). "General introduction to the techniques of information and documentation work" Paris: UNESCO.

2. PAO, Miranda Lee. (1983)."Concepts of information retrieval System" Englewood, Colo: Libraries Unlimited.

3. SALTON, Gerard & MICHEAL J. Mc Gill. (1983). "Introduction to modern information retrieval" Auckland: McGraw-Hill International Book Co.

4. VICKERY, Brain C. & ALAN Vickery. (1987). " Information science in theory and practice" London: Butterworths.

5. ENCYCLOPEDIA BRITANNICA. Encyclopedia Britannica Online http://www.britannica.com

6. WIKIPEDIA: Wikipedia, the free encyclopaedia. Wikimedia Foundation, Inc. http://www.wikipedia.org

7. ARMS, W.Y.: Digital Libraries. MIT Press, 2000 8, 10, 11, 12

8. BUSH, V. (1945). As We May Think. In: Atlantic Monthly, July, p. 101–108. http://www.theatlantic.com/unbound/flashbks/computer/bushf.htm

9. TOCHTERMANN, K. (March 2002). Personalisierung im Kontext von Digitalen Bibliotheken und issensmanagement. Professorial dissertation, University of Technology Graz.

10. OXFORD ENGLISH DICTIONARY: Compact Oxford English Dictionary. Oxford University Press. 2005. http://www.askoxford.com

11. GUTL, CH.; GARC´I A-BARRIOS, V.M. (2005). The Application of Concepts for Learning and Teaching. In: Proceedings of 8th International Conference on Interactive Computer Aided Learning (ICL 2005), Villach, Austria.

12. BORGMAN, CH. (2000, 11). From Gutenberg to the Global Information Infrastructure. MIT Press.

13. DREHER, H. ; KROTTMAIER, H. ; MAURER, H. (2004). What we Expect from Digital Libraries. In: Journal of Universal Computer Science 10, September, No. 9, p. 1110 – 1122. http://www.jucs.org

14. KROTTMAIER, H.: Aspects of Modern Electronic Publishing Systems. Graz, Austria, University of Technology Graz, dissertation. http://www.iicm.edu.

**15.** RIJSBERGEN, C.J. Van: Modelling Adaptive Information Retrieval / Department of Computing Science, University of Glasgow.

**16.** BAEZA-YATES, R.; RIBEIRO-NETO, B. (1999). Modern Information Retrieval. Addisson Wesley.

**17.** WONG, S.; ZIARKO, W.; WONG, P. (1985). Generalized vector space model in information retrieval. In: Proceedings of the 8th ACM SIGIR conference on Research and Development in Information retrieval, p. 18 – 25 22

## About Authors

**Sri. N. Rupsing Naik**, University Librarian i/c., University Library, JNT University Hyderabad, Kukatpally – 500 085. A.P.Indi**a**
E-mail: nunsavath2007@rediffmail.com

**Mr. A. Madhava Rao**, Research Scholar, Dept. of Lib. Sc. Dr. B.R. Ambedkar Open University, Hyderabad. A.P.
E-mail: amadhava.rao@gmail.com