SEARCH ENGINE TECHNOLOGY: DIGITAL LIBRARY PERSPECTIVE

*S.M.Shafi

ABSTRACT

The paper introduces the navigation of digital libraries and reviews most prominent problems confronting retrieval engine like multiple content sources, distributed documents problem, multiple language problem, inaccurate queries problem and availability of classification schemes problem etc .It traces various solutions in terms of technologies developed for search engine retrieval like Meta data extraction by various methods viz automatic extraction of words, structural analysis, normalization of existing Meta data, using external reference data and citation analysis etc. It explores equal ranking elaborating test runs with representative queries etc besides dealing with proper treatment of queries which covers orthographic, morphological and vocabulary variation .Discussion on data aggregation resorting to distributed documents is given along with a glimpse of future direction.

KEYWORDS: Digital Library: problems; Digital library: research; Search engine technology.

INTRODUCTION

The organization of digitized information and subsequent evolution of Digital library has posed many problems especially for effective retrieval which resulted in different generations of digital libraries during past few years ultimately culminating into efficient and intelligent retrieval tools for exploring the web and digital libraries for delivery, dissemination in an interoperable environment. The paper

* S.M.Shafi Professor, Department of Library of the information science. The University of Kashmir Srinagar (J&K) 190006 India.

looks into the significant problems facing the digital libraries and how these are addressed at different fronts which may ultimately lead to a new generation of digital libraries.

BACKGROUND

The problem of locating items in libraries is frequently referred to as "search," although that word tends to imply that one knows in advance what one is looking for, and possesses handles, indicators or index terms to serve as finding aids. This narrow view ignores the activity of browsing or even the higher-level function of becoming acquainted in general with a library's holdings. Browsing in a traditional library is a physical activity which involves scanning shelves on which related works have been placed in proximity, and occasionally withdrawing them from the shelves for examination. Browsing in a digital library is a logical activity mediated by a computer. It does not require physical proximity in any sense; indeed, two consecutive items examined may be stored on different continents. The question, then, is how can a library user (not to say the library staff) become familiar with the whole of recorded human information in a way that makes it accessible and useful? Experts may adopt the term "navigation" to mean moving about in a digital collection. Search is a directed form of navigation in which the goal is defined in advance with reasonable clarity. The result of a search may be an item, a collection of items, or any part of an item, even down to a single glyph. Tools must be provided that enable users to move about at varying levels of granularity within the corpus. The usual requirement for a search is that the user is looking for a specific piece of information or a summary of what is available about a certain topic. A common case is that the user wants the answer to a specific question, such as how old is Chinese civilization? Only rarely does such a question translate naturally into a keyword query. Such retrieval is indirect in the sense that the user wants to learn A, but formulates a query B, to which he receives a set of retrieved documents that must be scanned to determine whether the answer to A is among them. It would be far better simply to allow the user to ask question A instead of requiring him to convert it to some query language. Besides, the existence of Web searchers proves that text can be searched without being indexed or cataloged. At

least on a microscopic level, documents can be located purely by their content. Many documents consist of text plus other information such as mathematical and chemical equations, tables and drawings those themselves cannot be searched directly but can often be located by the presence of related text. Purely non-textual matter is very different. Although substantial progress is being made on video searching (through the use of extensive captioning cues, speech recognition and other aids), content searching of music and visual materials is non-existent or in its infancy. The problem is further complicated by the existence of work that combines media in various ways.

A user who is looking for general information on a particular topic is constrained in traditional libraries to go to an encyclopedia (which may have no entry or an outdated one on the topic of interest) or to refer to books that are generally about the subject under consideration. The time necessary for the user to obtain an overview at the appropriate level may be large because of the volume of repetitive material obtained. Programs are needed that are able to scan hits with the particular query in mind and produce abstracts, summaries, translations or analyses of the retrieved material. **(Reddy, 1999)** These and associated problems of navigation and subsequent solutions under way in research and development are pinpointed in the paper and how these are setting a trend in research and retrieval of digital libraries.

PROMINENT PROBLEMS

The major problems analyzed from literature search, observation and experimentation on navigating web and digital libraries are revealed below:

I. MULTIPLE CONTENT SOURCES PROBLEM

A digital Library provides a combined search on different collections at a time. The format varies between collections like web pages, journals, proceedings, databases, archives, pre-print etc and availability of structure, external reference data and Meta data variation and kind of content variation. Thus it becomes difficult to provide equal ranking among documents to the resultant set arising from different content sources.

II. META DATA PROBLEM

Meta data is a key to digital documents especially for data discovery but it is presently difficult to understand as who has generated meta data, whether it is automatically generated or author /editor is responsible .It is possible that digital library creators may do the exercise. one can't however predict what element of metadata are available (viz author, title, keywords etc) and above all the original purpose of tagging meta data is not known. Sometimes it is meant for providing a quick summary or gives a feel of condensed description. It may also serve as a source of Normalization of content for search but all this is not clear to patrons or clients of a digital library.

III. DISTRIBUTED DOCUMENTS PROBLEM

The documents are often hypertext based i.e. their parts/sections/references are distributed over a site with links between them. The document can't be retrieved fully as we are accustomed in the physical library with ease of consultancy and usage. Here the feel is of document but touch of a part of a section with poor or sometimes missing or non available hyperlinks etc.

IV. MULTIPLE LANGUAGE PROBLEM

Most library items, particularly in non-English-speaking countries, are not in English. The central translingual library question is how users may navigate through materials in foreign languages and make effective use of them. Translingual search is currently a research problem for which obvious solutions do not work. A keyword search cannot be made multilingual merely by translating the keywords one at a time. The number of possible translations of each word may be very large, so an explosion in the number of hits may result. This approach also takes no account of idiomatic uses, untranslatable words such as particles, and numerous other language-related phenomena. An interim solution is the use of translation assistants-programs that offer dictionary entries or partial or suggested translations of text portions. These show great promise for users who are at least partially familiar with the language of the retrieved document. Besides much work is on the way for analyzing handwriting at many research institutes and

OCR-Ring of many scripts especially non- roman ones which are now well documented. **(Razdan, Femiani & Rowe, 2003)**

V. CLASSIFICATION SCHEMES

Libraries have used different classification schemes which may be of some use in searching but underlying classification taxonomies are not standardized across collection and thus pose problems with digital libraries.

VI. INACCURATE QUERIES PROBLEM

Users lack domain specific knowledge and don't have proper terminology in hand. They are not in a position to include all potential synonyms and variations in the query. The clients have a problem but they are not in a position to comprehend or paraphrase it. In other words, how it is phrased in documents is problematic area in digital libraries and may pose many pitfalls in digital library navigation. There may be also a problem of homonyms. As an illustration, one interested to peruse his research in "Chemical Bond" will lead to funny results such as he may be referred to "James Bond" as well , creating noise and non relevant documents. On the base ground it is extremely difficult to provide a perfectly relevant result sets at the first response. However intelligent suggestions for refinement or expansion can be done by adding specific keywords.

TECHNOGIES UNDERWAY

Various technologies underway to solve the problems. These are outlined in the following paragraphs:

I. META DATA EXTRACTION

Meta data extraction can be achieved through different techniques like:

a. Automatic extraction of keywords
b. Structural Analysis
c. Normalization of exiting metadata
d. Using external reference data and citation analysis.

The automatic extraction of keywords is achieved by analyzing the title/text by speech tagging and normalization, extraction of specific syntactic patterns, and finally identification of new terminology. These

steps are more or less similar in designing Depth schedules **(Gopinath, 1985)** by DRTC (ISI), Bangalore, India especially for identification of isolates , quasi_ isolates etc and some exploratory research done in facet analysis visa-vis subject indexing of the web documents **(Shabahat Husain, 1995)**. This can be illustrated from the following simple text analysis:

Suffering from chronic prostitis the patient was

Vpart Prep Adj N Det N Vcop

Vpart prep **Adj** **N** Det N Vcop

treated

Vpart

t

$$\log \frac{P(\text{"chronic prostritis"})}{P(\text{"chronic"}) * P(\text{"prostritis"})}$$

Chronic Prostritis ⟶ Identification of new terminology

This is further worked out by Linguistic analysis where identification of key concepts are worked out by word and inverted index (stemming, suffuxes, morphological analysis, Boolean proximity, range, fuzzy search), phrasal analysis (like noun phrases, verb phrases etc), sentence level analysis (context free grammar, transformational grammar), semantic analysis (semantic grammar, case based reasoning)

II. STRUCTUARAL ANLYSIS

Structural analysis is a process wherein the text is analysed into various elements like journal title, paper title, author, affiliation and blocks of the text is determined taking various features into consideration based on classificatory ideas, ideology of grouping like Abstract, introduction, methodology, materials, results and discussion etc and finally structure of grammar methodology is applied to make it more result oriented.

III. NORMALIZATION OF EXISTING METADATA

The use of textual context of citation to obtain good descriptors of it and infer relative importance of the literature. It is possible through citation graph and use of external reference data. One can infer relatedness when these are cited by same papers. The bibliographic coupling and co-citation are also useful motivations in this context.Statistical analysis has been experimented at the lab. **(Chen,2002)** where following functions have been worked out for navigating in Digital libraries :

*similarity function: jaccard, cosine
*weighting Heuristics.
*Bi-gram, Tri-gram and N-gram
*Finite data automata (FSA)
*dictionaries and thesauri
*DLI project Illinois worked on co-occurrence analysis in terms of Heuristic term weighting: weighted occurrence analysis.

IV. EQUAL RANKING

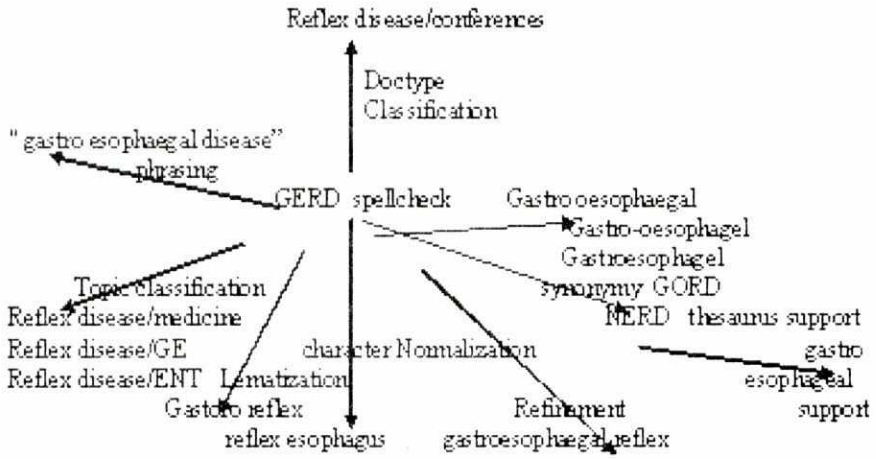
Equal ranking is a useful resource for retrieval engine. One can test runs of a representative queries based on content source like full text, abstracts, web data and indexed on citation index etc. A static rank likes Low, Medium and high boost can be given on content source which helps in retrieval in digital library environment.

V. PROPER TREATMENT OF QUERIES

Dealing with various linguistic and conceptual categorizations of queries like treating them from orthographic, morphological and vocabulary variations help in efficient retrieval. It also deals with

special interest queries to confine them on user home pages, forming definition or narrowing down to articles, patents or specifications. Much wok has been done in the field and development and use of various tools are proving assets in its implementation. This can be illustrated from from the following illustration

Fig. 1. Showing proper treatment of queries (adopted from Jrrgen Oesterle)



VI. CLUSTERING AND CATEGORIZATION (CHEN, 2002)

The clustering has been done on the following grounds:

*Hierarchal clustering: single link, multi link, wards

* Statistical clustering: multidimensional scaling, factor analysis

* Neural network clustering: self organizing map (SOM)

(It results in document clustering, cluster labeling, optimization and parallelization)

*Ontologies: directories, classification schemes.

VII. VISUALIZATION/HCI (CHEN, 2002)

*structures: trees, hierarchies, networks

*Dimensions

1 D (alphabetic listing of categories)

2D(semantic map listing of categories),
3D (Interactive helicopter flying through using VRML)

VIII. LINK ANALYSIS (CHEN, 2002)

*Authors vote via links
*Pages with higher in link have higher quality
*Page rank. It is to limit distribution of random walk
* Hubs: pages that point to many good pages
*Authorities: pages pointed to by many good pages.
*Operation over a vincity graph, refined by IBM clever Group

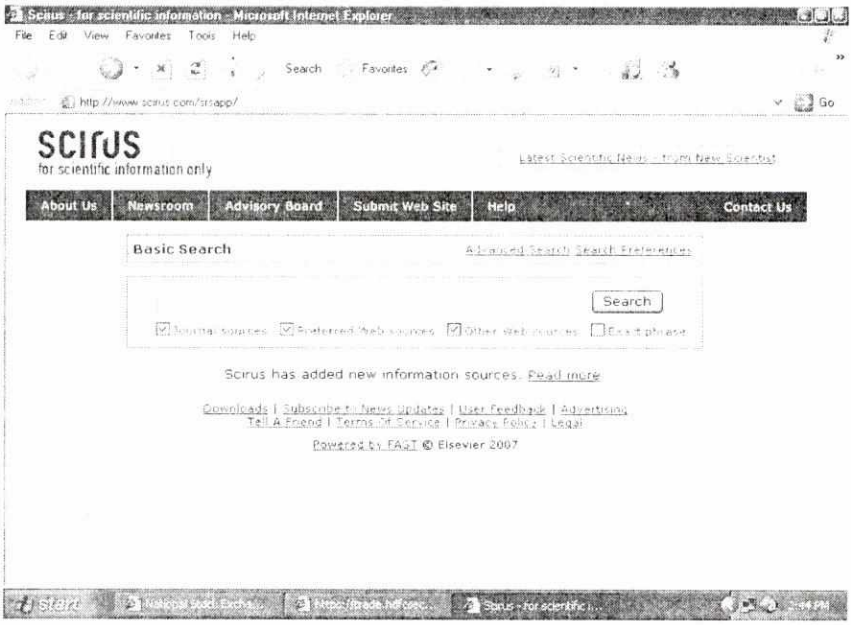
IX. TEXT CATEGORIZATION

Much work has been done in this direction which is based on classificatory ideas that have been documented by Fibro Sebastiani **(Fabrizo, n.d)** . It redefines categories as symbolic labels where stop words (topic neutral words) stamming (i.e. conflating different inflectional forms of the same word in a single class) has been demonstrated. **(Chen, Houston, Sewell & Schatz, 1998; Mooney & Roy, 2000)**

X. SMART DATA AGGREGATION

Data has been aggregated in a smart way by many search engines by designing and accessing knowledge portals navigated by clients through user friendly interfaces developed by specialized search engines taking clues from multifaceted research in digital library projects. The case is well demonstrated by Scirus search engine as shown in figure 2. However to overcome the overload of information it provides many choices date wise and document wise. Further research work is in full progress to evlolve more smart data aggregation by developing personalized annotations to digital libraries. **(Neuhold, Niederee, Claude &Stewart, 2004).**

Fig. 2. Scirus search engine showing facilities for navigation in a smart manner



CONCLUSION

In today's environment of increased expectations, information consumers demand tools and capabilities that help them access relevant information and enable them to manipulate the information that they retrieve. As the deep web continues to grow deeper, the importance of Directed Query Engine tools increases proportionately and the Distributed Explorer application is well positioned as the first generation architecture for the PSII. The vision behind the PSII is to create an integrated network for the physical sciences where content, technology and service converge to make resources readily accessible, openly available, useful, and usable. There is no value in information that is inaccessible, but great value in relevant information accessible with just a few clicks and commands.

REFERENCES

Chen, H., Houston, A L , Sewell, R R & Schatz, B R. (1998). Internet Browsing and searching user evaluation of category map and concept space techniques, *Journal of American society for Information* Science, 49, pp. 582-603.

Chen, Hsinchun. (2002). *Trailblzing a path towards knowledge and transformation of e- library, e –Government and e-commerce.* Arizona: knowledge Computing Corporation.

Fabrizo, Sebastiani. (n.d). Automatic document classification. Retrieved June 18, 2004 from http://faure.iei.pi.cnr.it/~fabrizo/publications/ACcs02.pdf/

Gopinathan, MA. 1985). Postulation approach to analytic synthetic classification. *Lib science with slan to Documentation,* 22, pp.204-229.

Mooney, R and Roy, L. (2000). Content based Book Recommending using Learning for text categorization. In *Proceedings of Fifth ACM conference on Digital libraries.* (pp 195-204). New York: ACM Press.

Neuhold, Erich. Niederee, Claude and Stewart, Avare. (2004). Context –driven Access to personalized digital multimedia libraries. In ICDL-04 (pp 451-460.).New Delhi:TERI.

Razdan,A. J. Femiani & Rowe, J. (2003). 3D Methods to aid Handwriting analysis & OCR. In Proceedings of symposuim on Document image understanding technology, (April-9-11, 2003).

Reddy, RaJ (1999). *Global digital Libraries: Building the infrastructure.*Report International Technology Research Instiute (ITRI) .Retrieved July 08, 2005, from http://wtec.org/logola/digilibs/

Shabahat Husain. (1995) *Classification: facets and approach.* New Delhi: TataMcGraw- Hill.