RGPVNOTES.IN

Subject Name: **Modern Information Retrieva**

Subject Code: **CS-7004**

Semester: **7th**

Subject Notes

CS 7004 [Modern Information Retrieval]

Topics to be covered: *Introduction to online Systems, Digital Library searches and web Personalization.*

Introduction to online IR Systems:

The underlying basics of online systems have arguably changed the least over time. Lancaster's overall systems approach, viewing information retrieval as a complex system that can be broken into many separate system components for better understanding, is the approach long favored by researchers who realize that each subsystem must be understood to understand or improve the whole.

While the underlying technology of inverted file structure has improved dramatically to provide efficient retrieval of massive full text databases, the importance was established in early online systems. Although often criticized and now faced with many alternatives, Boolean logic remains the standard for information retrieval systems.

The most common criticism of Boolean logic systems throughout the 1980s and 1990s was that end users had trouble understanding Boolean logic and thus query formulation is too difficult.

One thing that had to be changed to make online systems friendlier or easier to use was to improve the interface. Today's Web search engines have extremely simple interfaces that hide the inner workings of a complex system, and commercial information retrieval systems have improved over the decades, although interfaces for online information retrieval systems are still not considered user-friendly.

Online IR Systems:

The use of the computer for bibliographic information retrieval was first demonstrated in the 1950s, and initiated by the National Library of Medicine in 1964 using batch processing. Also in the 1960s, federally funded projects were carried out to develop prototype online systems which were then implemented in government research laboratories.

The last production service, Lochheed's DIALOG system, was implemented for NASA and subsequently made available to other government locations before becoming a commercial activity in the early 1970's.

Today DIALOG operates worldwide with databases offered via the Internet to Libraries and other organizations as well as individuals.

With a few exceptions, database vendors do not produce information but rather make it available to searchers via a common search interface. Database vendors license databases from the Producers, process the databases to introduce as much standardization as is feasible(e. g. standard field names), mount the database through the creation of inverted indexes, create database descriptions and aids to searchers in a standard format, and conduct training sessions for clients. These organizations offer a value-added service by providing a common gateway to multiple databases. A database vendor may offer cross-database searches; for example, DIALOG allows the searcher to search simultaneously a predetermined or searcher-selected grouping of databases to create a merged set of references, then process the set to remove duplicates.

IR in Online Retrieval Systems:

Since the inception of these online retrieval services, their retrieval functionality has been primarily on the Boolean model for retrieval, in contrast to research in the IR field which has focused on improving retrieval performance through non-Boolean models, such as the vendor space model. A number of factors guided the choice of the Boolean model as the basis for these services. Research in indexing and retrieval at the time, particularly the cranfield studies, a series of experiments comparing natural and controlled vocabulary indexing, suggested that a natural language" retrieval provided a level of retrieval performance comparable to manual indexing. Boolean logic was already being used in some libraries for manual retrieval systems, such as edge-notched cards and optical coincidence cards, and seemed to offer a natural mechanism for implementing retrieved based on combinations of words in documents.

Despite developments in IR research which suggested that alternative models might provide improved retrieval performance, Boolean retrieval has remained the commonest access method offered by database vendors, although in recent years some systems have added a form of natural language input with ranked output processing as an alternative access method.

Online Public Access Catalogs (OPACs):

An OPAC (Online Public Access Catalog) is an online bibliography of a library collection that is available to the public. OPACs developed as stand-alone online catalogs, often from VT100 terminals to a mainframe library catalog. With the arrival of the Internet, most libraries have made their OPAC accessible from a server to users all over the world.

User searches of an OPAC make use of the Z39.50 protocol. This protocol can also be used to link disparate OPCS into a single "union" OPAC.

Although a handful of experimental systems existed as early as the 1960s, the first large-scale online catalogs were developed at Ohio State University in 1975 and the Dallas Public Library in 1978.

The newest generation of library catalog systems are distinguished from earlier OPACs by their use of more sophisticated search technologies, including relevancy ranking and faceted search, as well as features aimed at greater user interaction and participation with the system, including tagging and reviews. These new features rely heavily on existing metadata which is often poor or inconsistent, particularly for older records.

These newer systems are almost always independent of the library's integrated library system (ILS), instead providing drivers that allow for the synchronization of data between the two systems. While older online catalog systems were almost exclusively built by ILS vendors, libraries are increasingly turning to next generation catalog systems built by enterprise search companies and open source projects, often led by libraries themselves.[5][6] The costs associated with these new systems, however, have slowed their adoption, particularly at smaller institutions.

An example of a next generation OPAC system is included in the Libramatic software package.

Three generations of OPAC:

First generation: known-item finding tools through search by author, title, control number.

Phrase searching OPACs', as they are generally called, were in a way the machine readable forms of conventional catalogues providing such access points as class mark, author, title , subject as phrase and simple left to right phrase matching. Such systems had certain obvious drawbacks, for the probability of exact matching between search phrases with indexing terms was rather small: Much of the computer capabilities were wasted as the system worked like a card catalogue. It was not user-friendly as user/system interaction was quite limited.

Second generation: increased search technology with access by subject headings, keywords, boolean queries problems included failed searches, navigational confusion enhancements represented large investments for a library.

Most of the existing OPACs are still at this stage. Influenced by the commercial bibliographic database, second generation OPACs have adopted many of their features likes 'online help messages', `alphabetical index displays' for searching search terms and using `Boolean logic' for their combination and effective retrieval.

Despite the improvements, the second generation OPACs have made the first generation, Hildreth regards them as 'deficient tools' for effective subject searching, for the following reasons:

❖They offer little or no help in translation of entry query terms into the vocabulary used in the catalogue;

❖They provide no help to the user in making alternate search statements and techniques, when the initial approach fail;

❖They do not in all cases lead to a successful free text search(e.g. of the title words); to the corresponding subject headings or class numbers assigned to a broader range of related material;

❖The retrieval records are generally devoid of such information as table of contents , abstracts and book reviews, that might help user to judge the usefulness of the documents;

Third generation: focus on open systems architectures, improved GUI, support for Z39.50 and Dublin Core, hypertext links, java programming, ranked results sets problems include slow pace of development failure to match trends in the Web.

The above listed deficiencies were investigated and some of the remedies that emerged were incorporated into third generation OPACs to enhance their subject searching capability. These systems are enriched by the inclusion of additional controlled and uncontrolled access points. Queries are accepted as a 'natural language' statement eliminating the need for the user to know quarry formulation and search techniques. Some of the systems use partial match techniques instead of Boolean operators. The retrieved sets are sometimes ranked according to the query relevance. These catalogues ensure vastly improved search system interaction at every level of the search process.

Digital Libraries:

Digital Libraries (DLs) are advanced and complex information (retrieval) systems, which offer many valued services besides searching and browsing, such as document preservation and recommendation reference services selective information dissemination, among others. All services are provided over various types of multimedia data (e.g., audio, video) in a distributed fashion.

Definition of Digital Libraries: DLs are organized and focused collection of digital objects, including text, images, video, and audio, along with methods of access and retrieval, and for selection, creation, organization, maintenance, and sharing of the collection.

Traditional libraries are among the first institutions to use IR systems, to create catalogs of records for the material from the library. The catalogs can be search by users in the library or over the Web (online

public access catalogs). These catalogs use database technology; the records are structured according to standards such as MARC (title, a few subject headings, and a classification number).

The basic 'Ss' and fundamental concepts of a 'minimal' DL.

Streams: sequences of arbitrary items (e.g., bits, characters, pixels, images) representing the content of a DL.

Structures: can be viewed as labeled directed graphs, which impose organization on the DL content.

Spaces (e.g., vector, probabilistic): used as support for services and for presentation purposes; can be seen as sets with operations that obey certain mathematical constraints.

Scenarios: stories that describe the behavior of services and consist of sequences of events or actions that modify states of a computation in order to accomplish a functional requirement.

Societies: sets of entities and activities and the relationships among them.

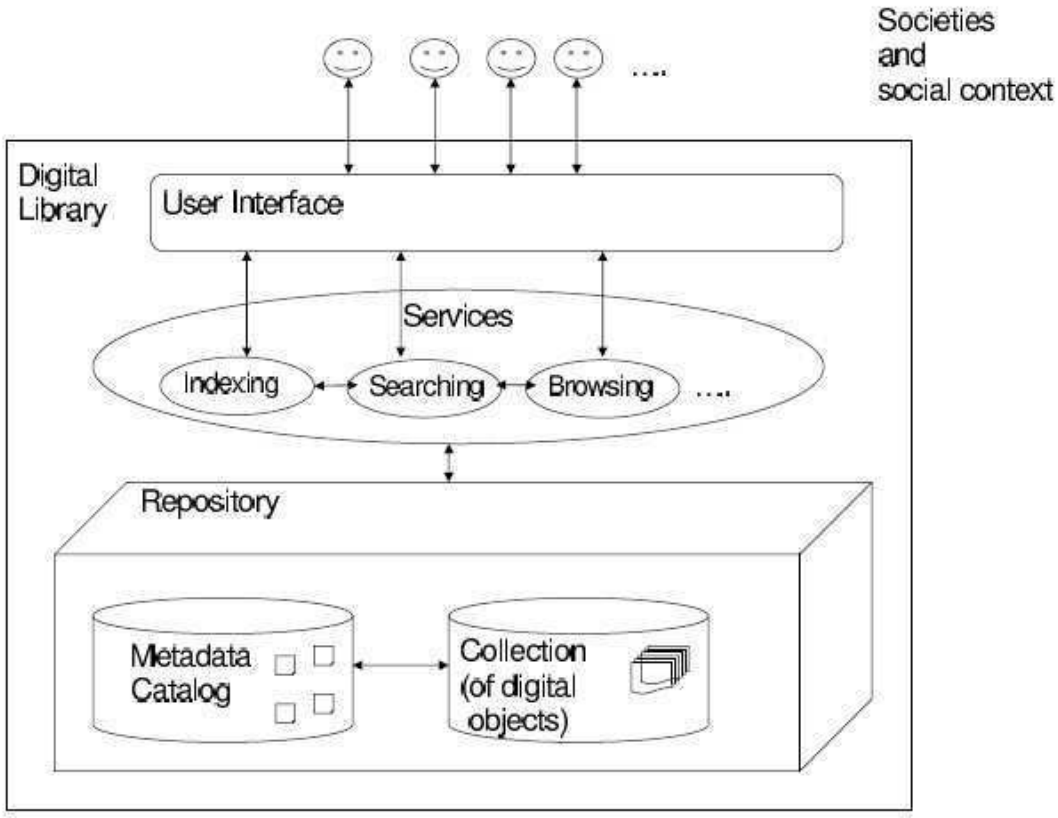

Figure 5.1: General Reference Architecture

- A DL is comprised of a collection of digital objects (e.g., digital documents, images, etc) and a catalog of metadata records that serve either to describe, to organize, or to specify how the objects in the collection can be used and by whom.

- Ideally every object should have a corresponding metadata record in the catalog, and this record should have a specific structure defined by a schema.
- Collections and catalogs are usually stored together in a repository that provides access and management capabilities to collections and catalogs.
- Services to create digital objects or metadata records, to preserve content, to add value to it, and to satisfy information needs are built on top of the repositories and are used by actors in a social context.
- Services can cooperate in terms of reuse or extension of capabilities to create more advanced services from simple ones.
- Usually a digital library provides simple searching and browsing services and indexing services to support the former two as a minimal set of services.
- The user interface serves as a ”glue" to organize and display all the provided services.

Modern libraries are being transformed to digital libraries as a result of the growth in electronic publishing, which makes information available in a digital form. Through the Web, a single interface provides access to local resources, as well as remote access to databases in the sciences, humanities, and business, including full-text journals, newspapers, and directories.

Special collections, in multimedia not only in text format, become available through the same gateway. For more details about the technology of digital libraries see, for example, (Lesk, 1997). Many libraries, particularly academic and large public libraries, have undertaken digital library project to achieve interoperability and ease of use and access. Two such projects are the Los Angeles Public Library’s Virtual Electronic Library project (http://www.lapl.org), and University of Pennsylvania’s Digital Library (http://www.library.upenn.edu).

A digital library could have no connection to an actual library, for example the ACM Digital Library (http://portal.acm.org/dl.cfm) that contains journal and conference publications in Computer Science. Digital libraries are more than complex IR systems. They are social systems centered around various communities of users. They also have component for building, cataloging, maintaining, and preserving repositories. There are many international or national digital library projects. One such project is the Digital Libraries Initiative (DLI) (phase one 1994-1998, phase two in progress) supported by the National Science Foundation (NSF), the Department of Defense Advanced Research Projects Agency (DARPA) and the National Aeronautics and Space Administration (NASA).

The DLI phase one contained large research projects at six universities: University of Illinois Urbana-Champaign, Carnegie-Mellon University, Stanford University, University of California at Berkeley, University of California at Santa Barbara and University of Michigan. These projects are developing the next generation of tools for information discovery, management, retrieval and analysis.

Web Personalization:

Global information retrieval and anywhere, anytime information access has stimulated a need to design and model the personalized information search in a flexible and agile way that can use the specific personalization techniques, algorithms, and available technology infrastructure to satisfy high-level functional requirements for personalization.

Personalized Information Retrieval and Access: Concepts, Methods and Practices surveys the main concepts, methods, and practices of personalized information retrieval and access in today's data intensive, dynamic, and distributed environment, and provides students, researchers, and practitioners with authoritative coverage of recent technological advances that are shaping the future of globally distributed information retrieval and anywhere, anytime information access.

Except when indexes are updated, a simple search engine delivers the same results for a query, But in reality, search performance may be improved if answers to the following questions can be obtained

- Who is searching?
- What role are they playing?
- What are they interested in?
- Why are they searching?
- Where are they located?

Personalized information retrieval represents only one aspect of a broad field of Personalization research Teevan et al quantified the potential for gain from personalizing search For queries supplied by the experimenters, they found a diversity of imputed intents Even when the imputed intents were the same, subjects disagreed substantially in the ratings they gave to results.

Pitkow et al studied the value of using a client-side personalization system termed as Outride.

Outride builds up a model of the user-based on their searching, browsing, demographic and application use profile.

Search results are re-ranked with reference to a vector-space representation of the user's profile.

The authors observed dramatic reductions in both:

(a) The time taken to complete a search task

(b) The number of user actions such as mouse clicks or keyboard entries

(c) Search tools can be customized according to characteristics of groups, individuals, or tasks

(d) There is also a particular potential for contextualizing search by employees within an organization

RGPVNOTES.IN

We hope you find these notes useful.

You can get previous year question papers at https://qp.rgpvnotes.in .

If you have any queries or you want to submit your study notes please write us at rgpvnotes.in@gmail.com

LIKE & FOLLOW US ON FACEBOOK
facebook.com/rgpvnotes.in