Department of Library and Information Science

Bharathidasan University
Tiruchirappalli-620024

Name of the Programme: M.Lib.I.Sc

Course - 2.2:Information Processing and Retrieval Systems

Course Code:P21MLS7

Unit-I: Information Retrieval System – concepts – Tools and Techniques; Models

Course Teacher: Dr.C.RANGANATHAN
Professor
e-mail: cranganathan72@gmail.com

Information Retrieval System

Information

- Information is nothing but idea communicated by others or obtained through personal study and investigation.
- Knowledge, intelligence facts or data which can be used, transferred or communicated.
- Information is represented as a body of knowledge which may take two factors.

Cont.,

- It may be recorded in publication
- Information may be passed directly or orally to be use.
- Retrieval
  - What is available and where is it applicable

System

- A set of components integrated to achieve a goal, system comprising of the components namely information, retrieval tools and the users.

Information Retrieval System

- Any system that is designated to facilitate this literature searching activity may legitimately be called an IRS – Lancaster.
- Searching and retrieval of Information from storage according to specification by subject-Moores.
- It is the process of location and selecting data relevant to a given requirements – J.H. Shera

Objectives

- Quantitative / Qualitative
- Text
- Drawing, graphs, maps etc.
- Computer Programme
- Description of objects
- Names location
- Bibliographical reforms

Function of IRS

- It provides either information recorded in documents or information about documents.
- It provides both current and retrospective information to readers.
- It saves the time of the reader by providing necessary information in time.
- It ensures each reader his or her information.
- It avails each piece of information its reader.

Cont.,

- IRS helps the searcher of information to retrieve the document on information easily and with in the stipulated period of time.
- It provides multiple access points conveniently and economically by using different retrieval tools.
- It helps to control the information explosion occurred in various field of knowledge.

Information Need

- Information has become a resource
- Demands of users of information services.
- Two categories
  - 1. The need to locate and obtain a copy of a particular document for which the title is known (Know item need).
  - 2. The need to locate documents dealing with a particulars subject or capable of answering a particular question (know subject need)

Know item need

- Approaches about on information that, a specific document exists and that certain clue about the author, title, editor is known by the users.

Know Subject Need

- Self explanatory
- Ability of information centre to supply known items when demanded by the users community is its "document delivery capability".
- The ability of the centre to retrieve document or a particular subject or provide answer to a specific question is its 'information retrieval capability'.

Information Retrieval Models

- In order to facilitate decision-making and problem solving, **intelligent knowledge based information retrieval models** comprise three basic aspects:

  a) knowledge base

  b) An inference engine - is capable of deriving appropriate information from the combination of rules for deriving a synthesized knowledge. This process of deriving is based on inferential logic using quantitative and non-quantitative techniques.

  c) A user interface

Information Retrieval Models

- Boolean Retrieval Model

  Standard Boolean Retrieval Model

  Weighted Boolean Retrieval Model

- Fuzzy Logic Model

  The conceptual fuzzy logic was introduced by Professor Lotfi A Zadeah.

- Set Theoretic Model
- Vector Space Model
  - Often weighted systems are discussed as vectorised information systems. This association comes from the SMART system at Cornell University, created by Dr. Gerald Salton [1979].

Cont..,

- A vector is a one-dimensional set of values, where the order/position of each value in the set is fixed and represents a particular domain. Each vector represents a document and each position in a vector represents a different unique word (processing token) in the database.
- The VSM is contrary to the Boolean Retrieval Model in which retrieval is based on the hundred percent (exact) match. The VSM allows retrieval of the most similar to the query without the exact match.
- Probabilistic Retrieval Model
- Linguistic Model
- Mathematical Model
- Psychological Model
- Economic Model
- Hypertext Linkage Model

Expert Systems for Information Processing and Retrieval

- **COMIT**: an early AI language developed at **MIT**, designed for information retrieval.
- **REFSEARCH**: developed by Joseph **Meredith and team**, and attempted representation of entire universe of reference works. This program is developed as an instructional tool for teaching librarians the basic principles of reference work.
- **RESEDA**: developed by **Zarri**, is a system based on restructuring the knowledge encoded in printed texts.
- **CANSEARCH**: Politt [1986, 1987] has developed Cansearch, an expert system for access to cancer therapy literature in the Medline database.
- **IR-NLI** (Information Retrieval – Natural Language Interface): IR-NLI translates the user's initial query into a formal problem statement.
- **Neural network computing** involves building intelligent systems to mimic human brain functions. ANNs attempt to achieve knowledge processing based on the parallel processing method of human brains, pattern recognition based on experience, and fast retrieval of massive amounts of data.

Cont…,

- **Information Storage and Retrieval (ISAR) system** deals with three basic aspects:
  - Information representation
  - Information storage and organisation
  - Information access.
- There are different **kinds of ISAR systems** such as:
  - Database Management System (DBMS)
  - Text Retrieval System
  - Management Information System (MIS)
  - Decision Support System
  - Knowledge Based System.
- **ISAR system evaluation -** Claverdon, B.C. Vickery, Alina Vickery

Cont..,

- The **structure of MARC records** is an implementation of national and international standards, e.g., Information Interchange Format (**ANSI Z39.2**) and Format for Information Exchange (**ISO 2709**).
- **MARC21 - Field 856 -** contain information that identifies the electronic location of an item (electronic resource), including enough information to retrieve the item.
- The **15 core elements** of **Dublin core (1995)** are – Title, Creator, Subject, Description, Publisher, Contributor, Date, Type, Format, Identifier, Source, Language, Relation, Coverage, Rights. The latest additions are the Audience and Rights Holder.
- **DCMI** is currently hosted by the **OCLC** Online Computer Library Center, Inc., a not-for-profit international library consortium.
- The **Text Encoding Initiative** (TEI) was founded in 1987 to develop guidelines for encoding machine-readable texts of interest to the humanities and social sciences.

Cont..,

- **DTD** : Document Type Definition – defines the structural rules of a type of document. These rules include a complete list of allowable elements and attributes, special character entities, rules for external files (such as images), as well as the hierarchical structure of all elements. Examples of documents type definitions include TEI, EAD, etc.
- **PURL** : Persistent Uniform Resource Locator is an approach to the URL permanence problem proposed by **OCLC**. A PURL remains stable, while the document's background URL will change as it is managed (e.g., moved) over time. A PURL is created by a Web administrator who is registered as a PURL 'owner' and who maintains a mapping of the PURL to a current and functioning URL.

Cont..,

- **Computer** is an electronic device and it only understands voltage ON and OFF. That is why a notation is developed to represent voltage **ON and OFF as '1' and '0'** respectively. This particular system of notation is known as '**Binary system**' and coding of characters using such system is known as '**character encoding**'. In ASCII, when 'A' labelled key is pressed, computer understands it as '01000001' (or decimal 65) and prints 'A' on the screen.
- **ASCII** was developed by a working committee known as **X3.4** constituted by IBM, AT&T and its subsidiary Teletype, in 1963 and revised in 1967. Basically this code was 8-bit character coding system, but it uses only 7-bits, that means only 128 characters are accommodated, instead of 256 (0-255). It should be noted that the numbers '0-9' are given 48 to 57; Capital letters 'A-Z' are given 65 to 90 and small letters 'a-z' are given 97 to 122.

Cont..,

- **ISCII** (Indian Standard Code for Information Interchange) is extended ASCII. It uses last 128 characters position for characters representation in Indic scripts in the range 0-255 provided by ASCII i.e. in ISCII, ASCII character are placed in the lower half (0-127) of the 8-bit code table while Indian script characters are in the upper half (160-255). ISCII caters to the following 10 Indian scripts - Devanagari, Gujarati, Punjabi, Bengali, Assamese, Oriya, Telugu, Tamil, Malayalam, Kannada. First version was released in 1983 and adopted by the Bureau of Indian Standards (IS 13194:1991) in 1991 after revisions in 1986 and 1988.
- **Unicode** is a 16 bit character coding code (allows for 65,536 characters), developed by UNICODE consortium which is a group of software organisations like IBM, Microsoft, Xerox, Oracle, etc. The first version of UNICODE came in 1991. Unicode is fully compatible with ASCII, i.e., the first 128 characters are ASCII characters. UNICODE has defined different language zones for their corresponding scripts. For example, Devanagari characters start at 2304 and end at 2431.

Cont…,

- The **ancient scripts** known are Bramhi around 7th BC and Kharosti during 5th BC. Of which Bramhi became the mother script for many Indian scripts. Meanwhile many scripts came into being like Grantha and later on Nagari, i.e., Devanagari, during 5th and 7th century respectively.
- **Graphics and Intelligence based Scripting Technology** (GIST), developed by CDAC

(India), supports major Indic languages. GIST technology uses ASCII code for character representation using the extended ASCII table (8-bit). GIST product uses underlying standards like ISCII, ISFOC (for font representation on screen) and INSCRIPT (common keyboard layout for Indian scripts). The software called iPlugin gives facility to develop the website in Indic languages.

UTF - Universal Character Set Transformation Format

UCS - Universal Character Set

- **Glyphs** are exposition of characters in graphical form, i.e., how a character appears on display, for example, bold, italics, or whatever.

Cont..,

- **Standards** Developed by **ISO/TC 46/SC9** for Bibliographic References Including Electronic Documents
- ISO 690- 2:1997, Information and documentation – Bibliographic references – Part 2: Electronic documents or parts thereof. First edition.
- ISO 999:1996, Information and documentation – Guidelines for the content, organisation and presentation of indexes. Second edition.
- ISO 3297:1998, Information and documentation – International Standard Serial Number (ISSN). Third edition.
- ISO 3901:2001, Information and documentation – International Standard Recording Code (ISRC). Second edition.
- ISO 5963:1985, Documentation – Methods for examining documents, determining their subjects, and selecting indexing terms, represent contents of documents systematically and consistently. First edition.
- ISO 10324:1997, Information and Documentation – holdings statements – summary level.

Cont..,

- **Standards** Related to Design and Hosting of Electronic Documents with Multimedia Components. These are under consideration and formulation stage by **ISO**.
- ISO 9241: Ergonomic principles for visual display terminals.
- ISO/IEC 10741: What happens to the cursor control when users interact with text editors.
- ISO/IEC 11581: Usage and appropriateness of icons in the user interface.
- ISO/IEC 13251: Collection of graphical symbols for office equipment.
- ISO 13407: Designing user interfaces with humans in mind.
- ISO/IEC 14754: Defines the basic gesture commands.
- ISO 14915: Recommendations for multimedia controls and navigation.
- ISO/IEC 18019: A standard for the design and preparation of software user documentation.
- ISO/IEC 18035: Icon symbols and functions for multimedia applications.
- ISO/IEC 18036: Icon symbols and functions for World Wide Web browsers.

Cont..,

- The **World Wide Web Consortium** (W3C) established on December 14, 1994.
- Its function is to oversee the formal organisation of HTML, as well as the various web-related protocols and languages, including CSS, XML and XHTML.
- **Applet -** A program inserted into a web page.


Best Wishes