TEERTHANKER MAHAVEER UNIVERSITY

TMU

MORADABAD

तीर्थंकर महावीर विश्वविद्यालय

सम्यक् दर्शन • सम्यक् ज्ञान • सम्यक् चारित्र

Accredited with NAAC A Grade

12-B Status from UGC

Information Storage and Retrieval
DBLIS201

Unit-1
Introduction to Information Retrieval

1. Introduction

With the technological advancement of science and as well as computer science in modern era, the data and information generation in every discipline of the universe of knowledge have seen a staggering growth over the last few decades. Storing, managing, querying and retrieval of huge amount of data and information needed a sophisticated procedure and suitable technology. Only generation of information is not necessarily the goal of humankind but also to cater the information need of the users. So, it’s to be understood that representation, organization and retrieval of relevant informationareimportant issues whichisactuallydealt by Information Retrieval.Information need of userhasaverycomplexnature.Manymodelshavebeendevelopedtounderstandtheinformation

need of human being, still, undoubtedly that remains a problem area and raises many open questions. The information need of human is changing due to application and integration of advanced technologies. Prior to the computerization and digital era, all the records (mostly documents and artefacts) were maintained in libraries or at personal collection level and retrieval used to be done by cataloguing schemes and other local practices.

2. NeedforIR

TheneedofIRcameintopictureduetosomefactors.Theyare-

i. Size and number of documents increased where no traditional cataloguing system can give technical support.

ii. Different disciplines (Earth Observation, Biotechnology, Genetics etc.) started producing different types of data with computer support and in multiple number of file formats which need to be indexed, stored, organized or retrieved. These data are mostly semi-structured (Video, audio) or unstructured (WebPages, E-resources).

iii. Libraries had a little or limited scopes in terms documents processing, handling different e-resources or sharing heterogeneous data and information over the internet.

iv. On the web, different organization started publishing and sharing information which shouldbe pre-processed, filtered and modelled to give a general structure in the web environment. Whereas, documents need to be indexed, scanned and coining information on bibliographic elements.Subsequently,thisbigdifferencecreatedatechnicalparadigmshiftandnecessitated to invent new theory and concepts to handle e-resources.

v. Librarian’s approach towards indexing is based on pre-coordination system and both success and efficiency of the indexing used to heavily depend on classification system (e.g. Colon Classification). Ex: - Chain indexing system developed by Dr. S.R. Ranganathan. But maintainingpre-coordinateindexingsystemcostsanenormoushumanlabour andalsoit lacks computation with mathematical or statistical approach. Unless the users know the properindex term needs to be given at search time, retrieval of relevant documents may be difficult. Post-coordination approach was necessary which is actually implemented in different search engines. This chapter will give an introduction to the subject.

3. Differentformsofmediaanddocuments

Mediaofinformation-

- Text
- Image
- Graphics
- Audio(Sound,Speech,Music)
- Video
- Animation

Documents

Document is a piece of written, printed or electronic matter the provides information or evidence or that serves as an official records. Documents may be of different types. They are-

- **MonomediaDocuments**:Text,Documents,OfficialRecordsetc.
- **MultimediaDocuments**:Documentswithdifferentmedia
- **HypertextDocuments**:Documentswithlinks(alsocalledasnon-lineardocument)
- **HypermediaDocument**:Multimedia+Hypertext
- **UserGeneratedDocuments**:Blogs,CommentsandTweets

4. WhatisInformationRetrieval?

Though understanding information need is a complex task but one easy way to express them is to transform the need into query form in natural language which is can be processed to map relevant documents and retrieve from storage space. So, in a word, Information Retrieval includes representation, storage, organization, accessing information which actually meets up the user need. Information retrieval started with mainly unstructured data like texts which actually doesn't haveclear, semantically overt and easy-for-computer structure [1]. On the contrary, database mainly deals with structured format of data where data are stored and managed with schema and proper definition of domains. Typically, a database applies relational algebra to establish the relations among the entities. Depending on the query in database, unlike database, IR system takes a different approach. In a nutshell, the primary objective of IR is indexing documents, make an indexed collection of them and giving a searching interface in order to retrieve them with certain level of relevance. Though, in last 20 years, there has been a huge research inputs from different organizations and universities and the objectives and activities of IR have been widened a lot. Now IR includes document clustering and categorization, classification of documents, system architecture, information and data visualization, alliedservices,rankingof documents,semanticlinking,filteringandothers.Searchengineshavebeen developed based on the concepts, principles and techniques developed by IR. Based on the different types of services,IRcanbecategorisedas websearch,personalisedIR,enterprise/institutionalservice based IR, domain specific IR etc. By nature, IR can be categorised as Web based IR system, digital libraries, multimedia IR system and distributed IR systems.

Advantages-

Over the time, the web expressed itself a potential platform of universal repository of human knowledge capital, channel of effective communication and sharing information. As web has an enormous amount of information, we need a systematic and procedural computational environment which can manage and retrieve this data/information with ease. Several communication protocols, software and hardware have been developed to make it possible. Now, the importance of IR is felt when there was a necessity to locate or to get those shared information without restrictions. The advantages of IR is-

- IR system is designed in such a way, it can accept queries in natural language and execute matching operation with its indexed term at back-end and locate the expected document from its term-document matrix.

- After executing the queries, search engine represents the results with ranks as a specific ranking algorithm (e.g. Page Rank) runs on the fetched result. Preferably, the most relevant documents get top ranks than non-relevant ones.

- As most of the IR systems (Search Engines) index the documents on incremental basis, web-based crawlers crawl the web pages in the hyperspace within certain time interval and get the updated information and further index the crawled information. Thus we get the latest information from the search spaces.

- IRsystemhasopeneduphugebusinessopportunitiesthroughweb environment.

5. BriefHistoryofInformationRetrieval

Approach to manage and organize large collection of information actually came from librarianship. It can be unambiguously claimed that cataloguing is the primordial soup for thebirth of Information Retrieval. Earlier days, mostly different books, documents, sacred manuscripts, scriptures, epics, spiritual documents were kept and indexed using cataloguing schemes. Eliot and Rose claimed in 3rd century B.C. Greek poet, Callimachus, first created own cataloguing schemes for managing his personal collections. In ancient periods, some big libraries were built. For example, library at Alexandria (280 B.C.) had more than 700,000 documents. Nalanda University had one huge library for document storage. But, the existence of any mechanism to organize, classify or retrieve them is still unknown.

In 1891, Rudolph filed a patent to US patent office for a machine composed catalogue cardsjoined together, which could be wound past a viewingwindow enablingrapid manual scanningof the catalogues. Soper in 1918 filed another patent for a device where catalogue cards with holed, related to categories, were aligned in front of each other to determine if there were entries in a collection with a particular combination of categories. If light could be seen through the arrangement of cards, a match was found.

The necessity of designing some mechanical devices that can be used for searching a cataloguefor a particular entry was felt in due years. Emanuel Goldberg was the first person who worked to solve that problemin the 1920s and '30s and indigenously. By nature, it's an optical device which basically searches for a pattern of dots or letters within the catalogues on a roll of microfilm. Goldberg patented many of his inventions in photography. Figure 1 shows the diagram of the patent filed in USPTO in 1928. "Hereit can be seen that catalogue entries were stored on a roll of film(figure1).Aquery(2)wasalsoonfilmshowinganegativeimageof thepartof thecatalogue being searched for; in this case the 1st and 6thentries on the roll. A light source (7) was shone through the catalogue roll and query film, focused onto a photocell (6). If an exact match was found, all light was blocked to the cell causing a relay to move a counter forward (12) and for an image of the match to be shown via a half silvered mirror (3), reflecting the match onto a screenor photographic plate (4 &5)"[1].

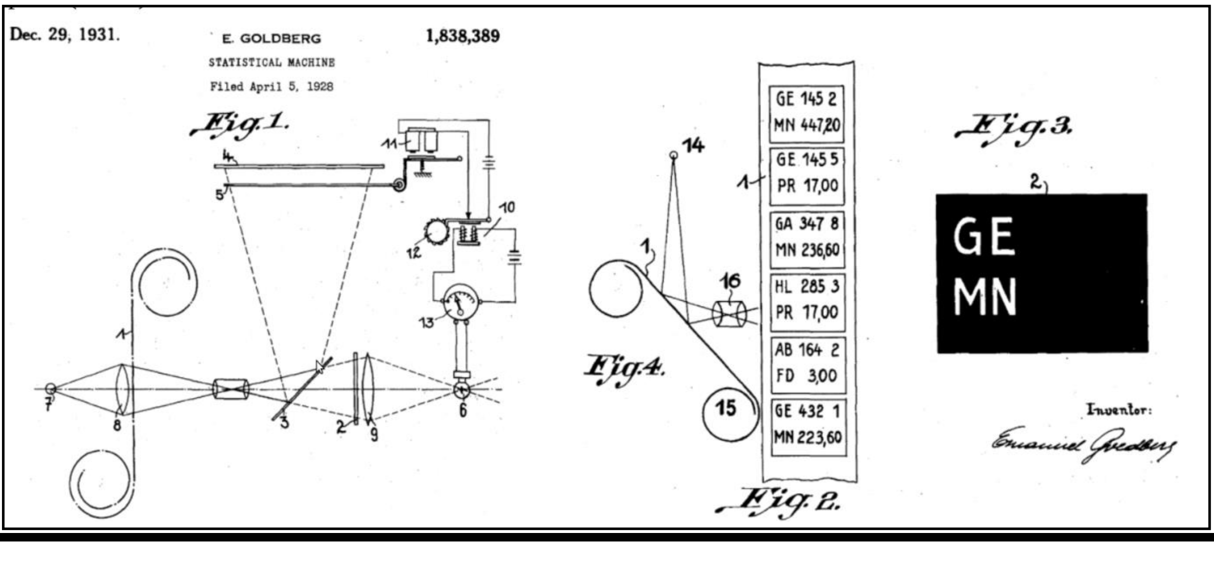

Fig.1:GoldmanOpticalMachine

After this big invention, in 1935, Davis and Draeger also made several experiments in similar line on microfilm based searching. As per Mooers, their work influenced Vannevar Bush and developed famous Memex System in 1945.

Radolph Shaw implemented Rapid Selector in US department of Agriculture (USDA) library .This machine was developed under the supervision of engineers in MIT and they worked on the earlier version of Rapid Selector on consent from Vannever Bush and delivered to USDA in 1949. “It was reported to search through a 2,000 foot reel of film. Each half of the film’s frames had a different purpose: one half for ‘frames of material’; the other for ‘index entries’. It is stated that 72,000 frames were stored on the film, which in total were indexed by 430,000 entries. Shaw reported that the selector was able to search at the rate of 78,000 entries per minute [1].”

In 1950, Luhn also made a selector using punch card, light and photo cells and this system could search over 600 cards per minute. Another important feature of this system is it could search the pattern of consecutive characters within a long string. Calvin Mooers in a conference in 1950 first coined the term “Information Retrieval” [2].

6. EarlyUseofComputers

In 1948, Holmstrom showed that Universal Automatic Computer (UNIVAC) could capable of searching for text references attached to subject code which used to be stored on magnetic tapes and could process at 120 words per minute. This is the first known fact where computer was used tosearch for contents. During 1950s, many projects were undertaken related to IR in different organizations (General Electric etc.).

ImportantMilestones

- Co-ordinateandUnitermIndexingbyMortimerTaube(1951)
- Cranfield 1(1957) study by ASLIB under the supervision of C.W. Cleverdon using uniterm and other classification systems (Universal Decimal Classification, Alphabetical Subject Catalogue, Faceted Classification Scheme). The main objective of the experiment centred around the investigation of

  i. TheCostofIndexing

  ii. Thecostofpreparingthephysical index

  iii. Thecost of searching

- Cranfield—WRUtest(ParalleltoCranfield-1test)
- Cranfield-2 test (1963-1966) was executed byC.W. Cleverdon,Mills and Keen and evaluated on the basis of two measures – Recall and Precision. Though, recall and precision measures were also used in Cranfield-WRU test.
- Gerard Salton (Early 1960's) proposed vector-space model for Information Retrieval and the perfomance of retrieval and ranking of the result was measured by cosine coefficient of document and query vector.
- In1962,AllanKentpublished“InformationAnalysisandRetrieval”.
- Weinberg report (1963) “Science, Government and Information” identified the problem of information transfer process and managing growing number of information along with its crisis. The report also put forward the urgency to address and formulate advanced techniques to retrieve information and manage and store them with convenience. This report made recommendations separately for Technical Community and Government Agencies.
- In 1968, the project report was published on the Intrex database design in MIT. This system could read the machine readable flexible, analytically-structured, catalogue-record format. Effort was also given “to the creation from each document of a set of complete index term phrases and to the problems of matching these unconstrained terms with similarly unconstrained subject request phrases”[3].

- SMART (System for the Mechanical Analysis and Retrieval of Text) Information Retrieval system developed by the leadership of Gerard Salton in Cornell University in 1960s. This system incorporated many important concepts like vector space model, relevance feedback, and Rocchio Classification. In 1968, Salton published his famous book titled as "Automatic Information Organization and Retrieval"

- J.W. Sammon (1969) gave the idea of visualisation interface integrated to an IR system in his famous paper "A nonlinear mapping for data structure analysis [4]".

- During 1966-67, F.W. Lancaster evaluated the MEDLARS (Medical Literature Analysis and Retrieval System) Demand Search Service. MEDLARS eventually gave birth to AIM-TWX and he also evaluated that during 1970-71. MEDLARS and AIM-TWX were the previous versions of MEDLINE/PubMed.

- Firstonlinesystems--NLM'sAIM-TWX,MEDLINE;Lockheed'sDialog;SDC'sORBIT.

- In 1975, three publications from Salton actually gave tremendous impetus to research in Information Retrieval community. They are -

i. ATheoryof Indexing[5]

ii. Atheoryoftermimportanceinautomatictextanalysis[6]

iii. Avectorspacemodelforautomaticindexing[7]

- ACM SIGIR Conference started in 1978 which subsequently emerged as the apex conference in this field.

- Belkin, Oddy, and Brooks gave the concept of ASK (Anomalous State of Knowledge) for information retrieval in 1982.

- One important invention happened during 1982-88 was formulation OKAPI model. It was developed at Poytechnic of Central London. Okapi is a set-oriented ranked output design for probabilistic type retrieval of textual material using inverted index [8].

- In1989,TimBerners-LeeproposedWorldWideWebinCERNLaboratory.

- TREC conference started as part of TIPSTER text program in 1992 and it was sponsored by US Defense and National Institute of Standards and Technology (NIST).

- PageRank algorithm was developed at Stanford University by Larry Page and Sergey Brin in 1996.

- Latent Dirichlet allocation (LDA), a generative/topic model in NLP was developed by David Blei, Andrew NG, and Michael Jordan in 2003. LDA is similar to probabilistic Latent Semantic Analysis (pLSA) and Latent Semantic Indexing (LSI). LSI gained huge popularityin WWW and was hugely used in Search Engine Optimization (SEO).

-

- In1997,Google Inc.wasbornwhichhasnowrulingdominantlyinsearchingengine domain.

7. Summary

The present situation of web and the environment of search engine did not evolve within moments ratherit's theproductof decades-longresearch.This chapter brieflydelineatedimportanceand history of Information Retrieval.

8. References

1. Retrievedfromhttp://ciir-publications.cs.umass.edu/getpdf.php?id=1066

2. C. N. Mooers, ‘The theory of digital handling of non-numerical information and its implications to machine economics’, in Association for Computing Machinery Conference, Rutger University, 1950.

3. Retrievedfromhttp://dspace.mit.edu/bitstream/handle/1721.1/1249/R-0360-14277844.pdf

4. Sammon, John W. "A nonlinear mapping for data structure analysis."IEEE Transactions on computers 18.5 (1969): 401-409.

5. Salton,Gerard.Atheoryofindexing.Vol.18.SIAM, 1975.

6. Salton, Gerard, Chung-Shu Yang, and CLEMENT T. Yu. "A theory of term importance in automatic text analysis."Journal of the American societyfor Information Science26.1 (1975): 33-44.

7. Salton, Gerard, Anita Wong, and Chung-Shu Yang. "A vector space model for automatic indexing."Communications of the ACM 18.11 (1975): 613-620.

8. Mitev, Nathalie N., Gillian M. Venner, and Stephen Walker. Designing an online public access catalogue: Okapi, a catalogue on a local area network. The British Library, 1985.

Uint-2

Basic Concepts and Components of Information Retrieval Systems

1. Introduction

We conceptualize the knowledge system into which an IR system is implanted to consist of three of component parts: a) people in their role as information-processors, b) documents in their role as carriers of information, and c) topics as representations. We are connected with the life cycle of each of these three objects and with the dynamic interactions amongthem. Thusthe objective of an informationretrieval system is to enable users to find relevant information from an organized collection of documents. In fact, most information retrieval systems are, truly speaking, document retrieval systems, since they are designed to retrieve information about the existence (or non-existence) of documents relevant to a user query. Lancaster comments that an information retrieval system does not inform (change the knowledge of) the user on the subject of their enquiry; it merely informs them of the existence (or non-existence) and whereabouts of documents relating to their request. However, this notion of information retrieval has changed since the availability of full text documents in bibliographic databases. Modern information retrieval systems can either retrieve bibliographic items, or the exact text that matches a user’s search criteria from a stored database of full texts of documents. Although information retrieval systems originally meant text retrieval systems, since they were dealing with textual documents, many modern information retrieval systems deal with multimedia information comprisingtext, audio, images and video. While manyfeatures of conventional text retrieval systems are equally applicable to multimedia information retrieval, the specific nature of audio, image and video information has called for the development of many new tools and techniques for information retrieval. Modern information retrieval deals with storage, organization and access to text, as well as multimedia information resources.

2. FeaturesofIRSystems

An information retrieval system is developed in order to help users to discovery relevant information from a storehouse containing collection of documents. The idea of information retrieval assumes that there exist several documents or records comprising data that have been arranged in a suitable order for easy retrieval. The storehouse contains many bibliographic information, which is quite different from other kinds of information or data. Let us consider some examples such as if we maintain a database of information aboutan institution or a supermarket, all we have are the different types of records and related facts, such as, for a college, names of students, faculties, staffs, their positions, qualifications and so on; in the case of a supermarket, names of different commodities, market prices, quantity and so forth. For such scenarios the retrieval system is designed to search for and retrieve specific facts or data, such as the qualification of a particular faculty, or the market price of a certain type of rice. Conventional database management systems, such as Access, Oracle, MySQL, etc, deal with structured data, where the arrangement or structuring of data takes place on the basis of the specific attributes of the data elements. For example, in a database of recipe, the various data elements could be the attributes of specific recipe records, such as recipe instruction, recipe yield, type of recipe, ingredients required, etc. In contrast to this, a database of items sold in a supermarket could be the name of the item with its barcode, manufacturer, supplier, price and so forth. So, the first database in this example will be structured according to the specific attributes of institution, whereas for the other database will be organized in accordance to the attributes of specific commodities. The main objective of these databases is to enable the user to search for specific records that be matched with one or more specific conditions or search criteria, for example, details of a certain recipe containing a particular ingredient; details of a specific product within a specific range of market price; a list of all the faculties that are involved with a specificcourse; or the products of a particular type grown at certain statesin the country, for example basmati rice available in north eastern stats in India.

Unlike a conventional database management system, an information retrieval systemdeals with unstructured data also. The main purpose of designing an information retrieval system is to meet the user requirements. It enables in document retrieval in-order to answer to the users' queries. The retrieved information can be in represented in different forms. The database can store abstracts of some bibliographic resources or full texts of documents,suchasjournalarticles,conferenceproceedings,newspaperarticles,textbooks,encyclopedias, legal documents, and statistical records,etc along with audios, graphics, images and videos information. No matter what the database may contain, be it bibliographic resources, full-text documents or multimedia information – the system assumes that there exists a target group of users for whom the system is designed and fulfill their requirements. Users may have certain queries or information needs, and they search for required information, the information retrieval system should be able to fetch the necessary bibliographic references of those documents bearing the required information; some systems also retrieve the actual text, image, table or chart relevant to the information needs of the user.

Letusconsideraverysimpleexampletounderstandthebasicfunctioningofaninformationretrievalsystem. Let usconsider asimplescenariowhereauser wantstodiscoveryinformationabout aterm,say‘nature’,ina book. One approach would be to start with the very first word in the first sentence present in the book, and continue to search for the term ‘nature’ until we find it or we come to the end of the book. However, in real life, this is not the scenario. Instead, to save our time we preferably use an index – the ‘back-of-the-book index’ – to look for an ideal match for the search term, and if we find a match then we take note of the corresponding references – the page number(s) where the term occurs – and we move to the specific page(s)

to find the information and the given context. In their simplest form, most information retrieval systemswork in this way.

Although historically information retrieval systems were established to help end users find relevant information from bibliographic and textual databases, in this 21st centuryinformation retrieval systems is used in almost each and every facet of our daily lives, for example, to retrieve a song on YouTube or e-mail received or sent on a specific date; to find sms sent to or by a particular person; to find a person's entity onthe web; to search for an e-book in an online library catalogue or in a digital library; to search for a book available for purchase in Amazon.com and so on.

3. ScopeofIR System

a. Unstructured Information: This information either does not have a pre-defined data model or is not organized in a pre-defined order. Unstructured information is typically text-heavy, but may contain datasets such as dates, numbers, and facts as well. This results in irregularities and ambiguities that make it difficult to understand using traditional computer programs as compared to data stored in fielded form in databases or annotated (semantically tagged) in documents (Wikipedia).Examplesof "unstructured data" may include books, journals, documents, metadata, health records, audio, video, analog data, images, files, and unstructured text such as the body of an e-mail message, Web page, or word-processor document. While the primary content being conveyed does not possess a defined structure, it generally comes packaged in objects (e.g. in files or folders or documents, ...) that themselves have some metadata and are thus a combination of structured and unstructured data, but normally it is referred to as "unstructured data".[7] For example, if we consider an HTML web page it is tagged, but HTML mark-up typically serves the purpose of presentation. It is not beingable to capture the significance or function of tagged elements in order to assist automatedprocessing of the information content of the page. XHTML tagging does allow machine processing of elements, although it typically does not capture or convey the semantic meaning of tagged terms. There are several techniques such as data mining and text analytics and noisy-text analytics, information visualization which give different methods to search for patterns in, or otherwise interpret from the available unstructured information. The most populartechnique for providing structure to several unstructured resources usually involve manual tagging with metadata or part-of- speech tagging for further text mining-based structuring. Unstructured Information Management Architecture (UIMA) provides a common model for processing this information to extract meaning and create structured data about the information [5].

b. Structured Information: It is information that is already structured in fields,such as “name”, “age”, “gender”, “hobby”, “address”, “profession”, “salary”. This is the typical example of what we find ina record of a relational database table. When information is organized in a structured form, it is usually relatively easy to search it, since one can directly query the database : give me the list of names whose profession is student in the table PERSON, where age is greater than 25 and name starts with the letter B. Structured data first depends on creating a data model – a model of the types of business data that will be recorded and how they will be stored, processed and accessed. This includes defining what fields of data will be stored and how that data will be stored: data type(string, integer, etc) and any restrictions on the data input (number of characters; restricted to certain terms such as Mr., Ms. or Dr.; M or F).

Structured data can be handled easily as they can be easily entered, stored, queried and analyzed. Due to the increase in cost and performance limitations of storage, memory and processing, relational databases and spreadsheets using structured data were the only way to effectively manage data. It is a common phenomena where anything that couldn't fit into a tightly organized structure requires to be stored on paper in a filing cabinet. Structured data is often managed using Structured Query Language (SQL) – a programming language created for entering, managing and querying data in relational database management systems. Originally developed by IBM in the early 1970s and later developed commercially by Relational Software, Inc. (now Oracle Corporation).

It is to be noted that information retrieval systems and database systems merely find what is already there:forexample, fromstudentsdatabaseshis/hermarks, fromthemarksitcan pointto theposition ofthestudent in class. An expert system on the other hand goes beyond just finding facts- it creates new information by inference: it identifies a student and gauges their merit in different subjects and their future prospect.

4. TypesofIR System

IR has concentrated more on finding the documents consisting of written text; much IR research focuses more specifically on text retrieval – the computerized retrieval of machine-readable text without human indexing. But it has spread across other interesting areas. Such as:

Speech Retrieval: Speech is an information-rich element of multimedia. Now there exist several techniques where information can be extracted from a speech signal in a number of different ways. Thus there are several well-established speech signal analysis research fields. These fields include speech recognition, speaker identification, voice detection, sentiment analysis and fingerprinting. The information that can be extractedfromtools and methods developed in thesefields can greatlyenhance multimedia systems and help mankind in various aspects.

Cross language information retrieval: It is an application area of information retrieval, which deals with fetching information written in a particular language different from the language of the user's query. E.g., Using Hindi queries to retrieve English documents. It is one of the challenging fields and a lot of research is going on in this area.

Question-answering IR system: It is a computer science discipline within the domains of informationretrieval and natural language processing (NLP), which is involved with building systems that automatically answer questions posed by humans in a natural language. A QA implementation, usually a computer program, mayconstruct itsanswersbyqueryingastructureddatabaseof knowledgeor information,usuallya knowledge base. More commonly, QA systems can pull answers from an unstructured collection of natural language documents. (Wikipedia)

Image Retrieval:It is part of sub-field of information retrieval. It helps the retrieval system for browsing, searchingand retrievingimages froma large database.The database maycontainonlydigitalimages, images along with text or may contain other types of resources like graphics, videos, audios along with the image, etc.Mostpopularandcommontechniquesofimageretrievalutilizesomemethodofaddingmetadatasuchas use of captioning, keywords, or descriptions to the images so that retrieval can be performed over the annotation words. The manual process of image annotation is not only time-consuming but is also alaborious and expensive affair; to address this; there has been a large amount of research done on automatic image annotation and image detection. Moreover, with theincrease in usage of social networks and a shift in paradigm from web to data web warrants new technology framework have inspired the evolution of several web-based image annotation tools.

Music Retrieval: Music information Retrieval (MIR) is the interdisciplinary field of retrieving useful information from music. MIR, although small yet it is a growing field of research with many real-world applications. Several researchers working in MIR may come from different background which includes computer science, instrumentation, musicology, psychology, academic music study, signal processing, machine learning or some combination of these.

Inadditiontotheabovementionedretrieval systems,IRalsodealswithanytypeof entityor object: workof art, software, courses offered at a university, people, products of any kind, etc. Text, speech or images, printed or digital, carry information, hence information retrieval.

5 FunctioningofIRSystem

An information system essentially makes ensure that users should be satisfied with the service. The system will be able to accomplish tasks, solve problems, and make decisions, based on the user needs. In short an information retrieval system should 1) find out the requirement of a target group of users, 2) a collection of relevantdocumentsandotherinformationresourcesshouldbemadeandindexedappropriately,and3)match documentswithuserneedsin-ordertofetchrelevantdocuments.Todeterminetheuserneeds,itinvolvesin studying information needs of users in general as a basic for designing responsive system (such as determining what study materials required for library and information science students typically need to do assignments in content management), and actively soliciting the needs of specific users, expressed as query descriptions, so that the system can provide the information. To have a successful retrieval system, it should figure out what information the users require to solve a problem. Query matching involves in mapping a query description with relevant documents in the collection; this is the task of the IR system.

All operations pertaining to information retrieval surround around usefulness and relevance of documents. The use of a document is dependent upon on three major things, topical connectedness, applicability, and originality. A resource is considered to be topically significant for a particular context, question, or task if it consists of information that either instantlyprovides answer tothe queryor can be used, in combination with other information, to infer an answer or perform the task. The appropriateness of the answer completely dependsupontheuser for agivencontext.It isoriginalif it providesaninput totheuser'sknowledge.Let us consider a simple situationwhere, a basketball player is important for a teamif his abilities and playingstyle fit the team strategy, applicable if he is compatible with the coach, and possess unique talent if the team is missing a player in his position.

Utility can be measured in monetary terms: "To what extent the document is useful for the user?" "What is the role of the player for a team?" "What is the recall and precision of the search engine"? From the literatures point of view, the term "relevance" is used for different purpose; it can indicate utility or topical relevance or pertinence. Many IR systems focus on finding topically relevant documents, leaving further selection to the user.

Relevance is a matter of degree; some documents are highly relevant and indispensable for the user as it servesthepurposeoftheusers'need; othersmaynot contributemuchtotheusers'requirements.For example if a user seeks information for 'orange' which is a fruit, all the documents about the fruit orange are relevant. Other documents may have the word 'orange' but might not indicate about the fruit (see ranked retrieval inthe section on Matching).From relevance assessments; measures of retrieval performance can be computed such as

recall = (relevant items correctly retrieved) / (all relevant items in the collection).

discrimination=(irrelevantitemscorrectlyrejected)/(allirrelevantitemsinthecollection) precision =

(relevant items retrieved)/ (all items retrieved)

Evaluationstudiesnormallyuserecallandprecisionoracombinationofboth;butthereexistsalotof argument whether these can be considered as the best measures for information retrieval systems.

6. BasiccomponentsinvolvedinIRprocess

An IR system performs retrieval operation by indexing documents and designing queries, thereby leading to representation of documents and representation of queries,respectively; the systemthen matchestheindexed documents with that of user query and displays the matched documents found and the user selects the relevant items. These operations are tightly intertwined and are directly dependent on each other. The search process often goes through severaliterations: several casesfeature similaritymeasurement isusedin order to distinguish the relevant documents from irrelevant ones and thereby it is used to improve the query or the indexing (relevance feedback).

Indexing:CreatingDocumentRepresentations

Indexing (from the library science point of view can also be referred as cataloging, metadata assignment, or metadata extraction) is the manual or automated process creating indexes for record collections. Having indexes allows researchers to more quickly find records for specific individuals; without them, researchers might have to look through hundreds or thousands of records to locate an individual record. We focus hereon subjectindexing– act ofdescribingorclassifyingadocument byindex terms or othersymbols in order to indicate what the document is about,tosummarize itscontent or to increaseits find ability. In other words, it isaboutidentifyinganddescribingthesubjectofdocuments.Indexesareconstructed,separately,onthree

distinct levels: terms in a document such as a book; objects in a collection such as a library; and documents (such as books and articles) within a field of knowledge (Wikipedia). Indexing can also be document-oriented – the indexer captures what the subject of the document, or request-oriented – the indexer assesses the document's relevance to other features of interest to users; for example, indexing the recipes in a cookbook in accordance to the course-type or meal or primary ingredients, etc making the resource interestingfor theusers.Abstractingisrelatedtoindexing– act of providingasummaryof thefull document giving the main content of the document or sometimes it may also include important results (informative abstract, summary). A lot of researchers have their interest on designing algorithms for building automatic summarization.

Automatic indexing begins with feature selection and extraction, this demands in extracting all the words from a text, this is followed by elimination of stop-words (words which are filtered out prior to, or after, processing of natural language data (text).There is not one definite list of stop words which all tools use and such a filter is not always used. Some tools specifically avoid removing them to support phrase search), stemming (the process for reducing inflected words to their stem, base or root form—generally a written word form. The stem need not be identical to the morphological root of the word; it is usually sufficient that related words map to the same stem, even if this stem is not in itself a valid root) , counting (using only the most frequent words), and mapping to concepts using a thesaurus or ontology (Wikipedia). In case ofimages, extractable features include color distribution, texture or shapes detection . For music, extractable features comprises of frequency of occurrence of notes or chords, harmonies, melody, main pitch, beats per minute or rhythm in the piece.

Features are generally processed further for retrieval. The system makes use of a classifier that links the raw orrefined features with that of a descriptor from a pre-established index language. A classifier can be built manually by making each descriptor act asa query description and building a query formulation for it. Moreover a classifier can be built automatically by making use of training sets, for example, the list of documents forbiotechnology, for machine learning of what features predict what descriptors. There exist several techniques that enable prediction of different words and word combinations by using the same descriptor, thereby making it easier forusers to find all relevant documents on a given context. The process of assigning documents to (mutually exclusive) classes of aclassification is also known as text categorization. Analyzingthe documents having similar features and clustering them in one group lead to identification of unique classes in which the documents belong. These are some initial steps of document classification.

QueryFormulation:CreatingQueryRepresentations

Information Retrieval means making use of the available information in-order to anticipate the extent to which a given document issignificant or useful for aparticular users'information need as outlined in a free-form query description, also calledtopic description or query statement. A user’s query can be transformed, manually orautomatically, into a formal query representation (also called query formulation) when combined with features, it helps to predict the usefulness of a document with respect to the query. The information need of the users can be identified by analyzing the query in terms of the system’s conceptual schema, ready to be matched withdocument collected in the database. A query may be in search of text words or phrases that the system should acknowledge and search (free-text search) or any other entityfeature,suchasdescriptorsassignedfromacontrolledvocabulary,anauthor’sorganization,orthetitleofthe journal where a document was published.A query can simply give features in an unstructured list (for example, a “bag of words”) or combine features using Boolean operators (structured query). Examples: The Boolean query specifies three conditions, AND, OR, NOT. If a query contains AND operator it indicates narrow search and retrieve records containing all of the words it separates. Similarly if the query includesOR operator it broadens the search and retrieves records containing any of the words it separates. Thesymbol '|' can be used instead of 'or' (e.g., 'mouse | mice | rat' is equivalent to 'mouse or mice or rat'). Lastly NOT operator indicates narrow search and retrieve records that do not contain the term following it. If there exists some relevant documents, the system can usethem as a training set to build a classifier with two classes: relevant and not relevant. These relevant and non-relevant documents will lead to the measurement of recall and precision. The requirement for the information need and formulating the query often acts as a cup and plate as they move together, directly dependent upon each other.An IR system can show a subject hierarchy forbrowsingandfinding gooddescriptors,oritcanasktheuseraseriesofquestionsandfrom the answers construct a query. For buying an online food item, the system might ask the following three questions:

- Whatkindoffooddoyouprefer(vegetarian,non-vegetarian,...)?
- Areyouallergictoanyparticularingredient(prawn,carrot,cuminseeds,..)?
- Whatkindofcuisineyouprefer(Italian,Indian, American,..)?

The system should help the users by suggesting synonyms and narrower and broader terms from itsthesaurus. This will help the users to visualize all the features to consider, without which it would not have been feasible. Throughout the search process, users further clarify their information needs as they read titles and abstracts.

Matchingthequeryrepresentationwithentityrepresentations

The document relevance is predicted by identifying the relevant features of the query with that of the document. In case of an exact matchthe system is able to mark the documents that satisfy all the conditions of a Boolean query (it predictsrelevance as 1 or 0). In-order to improve recall, the system can make use of elaborating the synonyms (if the query asksfor dessert, it finds sweets as well) and hierarchic expansion or inclusive searching (it finds dairy product as well). Since relevance or adequacy is a matter of degree, many information retrieval systems (including mostWeb search engines) rank the retrieved results by a score of expected relevance (ranked retrieval).Consider the query “Study of concept analysis in information retrieval”. In this case each term's contribution is a product of three weights: The weight of the queryterm(thesignificance of thetermto the user), thetermfrequency(tf) (the number ofoccurrences of the term in the document, synonyms count also), and the infrequency of the term in context of the document orinverse document frequency (idf) on a logarithmic scale is measured. If document frequency = .01 (1 % or 1/100 of all documents, the term is to be included), then idf = 100 or 10^2 and log(idf) = 2.

Selection

The user searches for the most relevant result and selects the appropriate items. Results can be organized in rank order (the search process can be stopped once the users' need is fulfilled); in case of groupings the documents based on subject, automatic classification scheme or clusteringtechniques (similar items can be examined side by side) can be applied. The display of titles along with the abstract with key termshighlighted is considered to be the most useful (as title alone is tooshort, the full text too long). For certain scenarios users may require assistance while making the connection between anitem found and the task at hand.

RelevanceFeedbackandInteractiveRetrieval

Once the user has evaluated the significance of a few items found, the query can be made better. The system can thereby provide assistance for the users in enriching the query by displaying a list of features (assigned descriptors; text wordsandphrases,and so on) foundin manyrelevant items andanother listfromirrelevant items. In some cases the system can automatically improve the query by identifying those unique features which can distinguish between relevant from irrelevant items and thus are good predictors of relevance.

7. PurposeandFunctionofIRSystem

Purpose

An information retrieval system serves the purpose to retrieve the resources or information required by the target audience. It is also important that the right information should reach to the right people at right time. Thus, the main aim of an information retrieval system is to collect and organizeinformation in one or more fields in order to help the users to access the retrieved resources. The use of information retrieval systemscan be explained considering a simple scenario:

- Awriterputsforthhis/herideaonadocumentusingsomeconceptsforagiven context.
- Somewhere around the globe there might be some target audience or a person who is in need of that unique idea but is unable to find it; in other words, some people is ignorant of the ideas put forward by theauthor in their work.
- Here Information retrieval systems bridge the gap by matching the writer’s ideas expressed in the document with that of the users’ requirements or demands for that idea.

Thus, an information retrieval system functions as a bridge between the world of creators or generators of information and the users who are ignorant of that information. Hence some researchers state thatinformation retrieval is an information-communication system.

Function

An information retrieval system deals with different sources of information on one hand and on the other hand it has to cater to several users' requirements. It must:

- availablecontentsaretobeanalyzedintheinformationsourcesaswellastheusers' queries,and then
- thentheuserqueriesarematchedwiththeavailabledocumentin-ordertoretrievetherelevant resources.

Thedifferentfunctionsofinformationretrievalsystemsareasfollows:

- Toidentifytheinformation(sources) relevant totheareasof interest of thetarget users' community; this is a challenging job especially in the web environment where virtually everybody in the world can be the potential user ofa web-based information retrieval system.
- Toanalyzethecontentsof thesources(documents); thisisbecomingincreasinglychallengingasthe size, volume and variety of information sources(documents) is increasing rapidly; web information retrieval is carried outautomatically using specially designed programs called spiders.
- To represent the contents of analyzed sources in a way that matches users'queries; this is done by automaticallycreatingoneor more index files, and isbecomingan increasinglycomplex taskdueto the volume and variety ofcontent and increasing user demands.
- To analyze users' queries and represent them in a form that will be suitable for matching the database; this is done in a number of ways, through the design ofsophisticated search interfaces including those that can provide some help tousers for selection of appropriate search terms by using dictionary and thesauri, automatic spell checkers, a predefined set of search statements and so forth.
- To match the search statement with the stored database; a number of complexinformation retrieval models have been developed over the years that are usedto determine the similarity of the queryand stored documents.
- To retrieve relevant information; a variety of tools and techniques are used todetermine the relevance of retrieved items and their ranking.
- Tomakecontinuouschangesinallaspectsofthesystem, keepingin mindtherapiddevelopmentsin information and communication technologies (ICTs)relating to changing patterns of society, users and their information needs andexpectations. (Chowdhury)

8. Summary

In short, IR involves in finding some desired information which is stored in a storehouse commonly called database. A typical IR system should meet the following functional and nonfunctional obligations. It must enablethe user tocreate,insert, modifyand delete,documents inthedatabase. It shouldbe aplatformfor the users to search for documents by entering queries, and examining the retrieved documents. An IR systemwill typically need to support large databases, some in the megabyte to gigabyte range, and retrieve relevant documents in response to queries interactively--often within 1 to 10 seconds. This field has come out with several path-breaking results as several research labs are working to develop many modern techniques for better precision.

9. References

1. Sparck Jones, K. and Willett, P., Overall Introduction. In Sparck Jones, K. and Willett, P. (eds) Readings in Information Retrieval , San Francisco, Morgan Kaufmann Pub. Inc., 1997, 1–7.

2. Parsaye, K., Chignell, M., Khosafian, S. and Wong, H., Intelligent Databases: object-oriented, deductive hypermedia technologies , New York, John Wiley, 1989.

3. Lancaster,F.W.,InformationRetrievalSystems,NewYork,JohnWiley,1968.

4. Belkin, N. J., Anomalous States of Knowledge as a Basis for Information Retrieval, CanadianJournal of Information Science , 5 , 1980, 133–43.

5. Meadow, C. T., Boyce, B. R., Kraft, D. H. and Barry, C., Text Information Retrieval Systems , 3rd edn, London, Academic Press, 2007.

6. Lancaster, F. W., Information Retrieval Systems: characteristics, testing, and evaluation , 2nd edn, New York, John Wiley, 1979.

7. Kent,A.,InformationAnalysisandRetrieval,3rdedn,NewYork,BeckerandHeys,1971.

8. Vickery,B.C.,TechniquesofInformationRetrieval,London,Butterworth,1970.

9. Vickery, B. and Vickery, A.,Information Science Theory and Practice , London, Bowker-Saur, 1987.

10. Rowley, J.,The Basics of Information Systems , 2nd edn, London, Library Association Publishing, 1996.

11. Liston, D. M. and Schoene, M. L., A Systems Approach to the Design of Information Systems. In King, D. W. (ed.)Key Papers in the Design and Evaluation of Information Systems, New York, Knowledge Industry, 1978, 327–34.

12. Voorhees,E.andHarman,D.(eds),TextRetrievalConference,Cambridge,MA,MITPress,2005.

13. Sparck Jones, K., What’s the Value of TREC: is there a gap to jump or a chasm to bridge?SIGIR Forum ,40 (1), 2006, 10–20.

14. Hearst,M.,SearchUserInterfaces,Cambridge,CambridgeUniversityPress,2009

15. Chowdhury,G.G.,IntroductiontoModernInformationRetrieval,3rded.London,2010.

Unit-3

Users of Information Retrieval Systems

1. Introduction

The user is an important component of any Information Retrieval System (IRS). The ultimate aim of an IRS is to connect the user, quickly and efficiently to the proper information. User is the last link or the recipient of information, also known as ‘end-user’. There are other terms used to represent the concept of user such as patron, client, member, customer, etc. In the context of IR the term 'user' is employed to represent the seekers of information. The person who is actively seeking access to information. The person who, when successful in search and retrieval, obtains and uses the information is described as user. Sometimes user is referred as searcher also. For an IRS to be effective and efficient it is very necessary to understand the following:

a. Who are the users, their needs, and what is the nature of their needs?
b. How they seek the required information?
c. What is the use pattern they exhibit in using the information?

In this module we will discuss about the various aspects of users, their categories and nature. The concept of information need and types of information needs are discussed and more specific information needs in different areas of activities are also explained. The information seeking behaviour of users in order to satisfy their information needs. The various methods generally known as user studies carried out to find the pattern of overall interaction of user with the IRS is also discussed.

2. Users and Their Nature

The person who is actively seeking access to information and who, when successful, obtains and uses the information is described as user. Users can be categorised on the basis of different characteristics, such as the extent of use of IRS for satisfying their information need or type of activity they are involved in.

Broadly, the important group of users can be distinguished according to the kind of activity in which they are engaged:

a. Researchers in basic and applied sciences.
b. Practitioners and technicians engaged in developmental and operational activities in the various fields of technology and industry: agriculture, medicine, industrial production, communication etc.
c. Managers, planners and other decision makers who are engaged in developmental activities in both private and public sector.

However, the user groups can be identified according to other characteristics such as:

a. By nature of work:
Engineers, scientists, policy makers, researchers, planners, managers, persons in different professions, etc.
b. Psychological criteria:
Users with superiority complex, with inferiority complex, selfish, abnormal, normal, etc.

c. By nature of activity:
Study, research, specialization, level of education and responsibility, expert, novice etc.

2.1 User Functions

In any IRS, there are several functions performed by the users. The functions include the sequence of activities performed by users to access the information, such as searching, browsing, selecting and evaluating the information objects of their interest. It also includes activities relatedto obtain and use the information objects once the access seeking is successful.

3. Types of Information Needs

In order to have a user oriented system it is imperative to focus the attention to 'user' and his information needs. The accurate assessment of information needs of users forms the primary basis for all information activities.

Before looking at the types of information needs it becomes necessary to understand the concept of 'information need'. Information need comprises of two terms information and need. It may be defined as need for information; information need is a factual situation in which there exists an inseparable interconnection between 'information' and 'need'. It is also to be understood that theinformation need exists objectively, that is they are oriented towards reality, practice and task. Tohave a true perspective of 'information need', the dictionary meaning of the term 'need' and other closely related terms such as; requirement, want and demand should be analyzed.

Maurice Line, (1974) attempted to define these terms in the perspective of information and tried to solve the difficulty of separating the concept of need, want, requirement demand and use, as under:

a. **Need:** What an individual ought to have, for his work, his research, his edification, his recreation, etc. In the case of a research, a needed item of information is one that would further his research. There may be an implied value judgment in the way the term is used. A need may or may not be identified as a want. A need is a potential demand.

b. **Want:** What an individual would like to have, whether or not the want is actually translated in to a demand. Individual may need an item they do not want, or want an item they do not need (or even ought not to have). A want like a need is a potential demand.

c. **Requirement:** It can mean what is needed, what is demanded, and can therefore be usefully employed to cover all three categories.

d. **Demand:** What an individual asks for, more precisely a request for an item of information believed to be wanted (when satisfied, the demand may prove not to be want after all). Individuals may demand information; they do not need and certainly need or want information they do not demand. Demand is partly dependent on expectation, which in turn depends partly on existing provision of library or information services. A demand is a potential use.

e. **Use:** What an individual actually uses. A use may be satisfied demand, or it may be the result of browsing or accident. A use usually represents a need of some kind. Use can be partial indicator of demand, demand of wants, and want of needs.

On the similar attempts, Taylor (1968) has explored the information need from the perspective of psychology of human behaviour, as follows:

a. **Visceral need:** An actual but unexpressed need for information.
b. **Conscious need:** An ill defined area of decision
c. **Formal need:** An area of doubt which may be expressed in concrete terms.
d. **Compromised need:** A need translated into what the resources and files can deliver.

Similarly, on the psychological grounds N J Belkin (1968) proposed the concept of Anomalous State of Knowledge (ASK) hypothesizing that the information need arises from the recognized anomaly in the users' state of knowledge concerning some topic or situation and that, in general, the user is unable to specify precisely what is needed to resolve that anomaly.

Information need is a condition in which certain information contributes to the achievement of a genuine or legitimate information purpose. Information need is a relationship that exists between information and its intended purpose or use.

Information needs can be divided into the following categories:

a. **Social or pragmatic information Needs:** Under this category we can put information that is required to cope with the day to day life, such as weather details of a location, Bus and train timings, etc.
b. **Recreation information needs:** This category involves information satisfying the recreational and cultural interests of the users, such as the upcoming books, TV show timings etc.
c. **Professional information needs:** The information required to operate competently and efficiently within the professional environment of the user. This involves information regarding new trends and practices being followed.
d. **Educational information needs:** The information required to satisfy academic requirement at an institution or to learn new skills.

Information need can also be classified into, kinetic and potential information need. The kinetic needs are directed towards satisfying a special problem in hand diagnosed and are of immediate concern, while potential needs remain hidden under the layers of attitude, impulse and values.

4. Information Needs in Different Areas of Activity

In this section we will discuss the information needs in some areas of activity. The areas covered are, Industrial information needs, planning information needs, information need in business, decision-making, research and development information need, Information need for business.

Before going in detail about the various information needs in different areas of activities it is worth to understand how to ascertain the information needs of the clientele.

Ms. Pauline Atherton (1977) listed some methods of ascertaining the information needs:

a. Study the organizational chart of the institution.
b. Study of its functions, activities chart of the organization.
c. Study of its annual reports, project reports and other publications.
d. Survey of users' requirements using questionnaire.
e. Interviewing users:
    i) Interview of superiors of user (persons higher in the hierarchy)
    ii) Interviewing user
    iii) Interview of subordinates of user (person controlled, taught, guided etc.)
f. Study of papers, books, etc. published by the user.
g. Attending seminars, colloquia, etc. in which the users participate.
h. Observing user at his work place.
i. Personal informal contacts with users.
j. Meeting users in small, preferably homogeneous groups periodically.
k. Feedback from information services rendered.
l. Providing for suggestions from users, about their subject interest, author interest, institutional interest, etc.
m. Attending technical meetings within the institution at which projects and problems may be discussed.
n. Scanning correspondence and reports prepared and received by the user.
o. Study of documents used by user.
p. Study of reference queries received from the users.
q. Participation in work orientation programmes.
r. While orienting and guiding users in using the libraries resources, tools and techniques.
s. Study of classification schemes and handbooks.
t. Liaison.

Industrial Information Needs

Success of an industry depends on its ability to receive the vital information in time. It is well established now that the more updated an industry, the more successful it becomes. Information need of industries falls into following broad categories:

a. Technological information
b. Company oriented information
c. Economic information
d. Policy information

However, any new industry in the process of its establishment may need the information concerning: scope and prospects for the industry, location, land, machinery and equipment, raw material, utilities, transportation, staff and labour, finances, regulations and procedures, market strategy.

Planning Information Needs

Planning is process of determining the course of action. Proper planning helps in achieving the goals by following a well set path of interrelated events. As the planning is future oriented, to do it well an accurate assessment of past and present situation of the relevant environment is a must.

The information needs in planning activity can be understood by understanding the steps that are followed while planning. Planning activity involves five interactive steps and the planner mustbe supplied with the needed information at each step.

e. Planning establishes goals and objectives. This requires large amount of information related to present and past events and situations.
f. Planning also identifies the events and activities that must be performed to achieve the goals. This step also requires considerable amount of information relating to each event and activities.
g. The next step is to describe the resources and/or talents necessary to perform the identified activity. Information related to available resources such as individuals who will implement and control the activities is paramount at this step.
h. Defining the duration of each identified activity: This requires lot of prior experience and other information about the sub-activities.
i. Final step is to determine the sequence in which the identified activities must be performed for best results.

Decision-making and Information Needs

Decision-making is a process of selecting the most desirable or the optimum alternative toresolve a problem or to attain a goal. Decision-making ranges from taking routine decisions to the complex ones. There is a direct relationship between decision-making and information; the more the decision maker is informed the better decisions can be made. Information is an essential ingredient of decision-making. Decision-making is pragmatic, rational, information using process. Hence it is very much necessary to provide accurate and timely information to the decision makers so that their information needs is fulfilled and informed decisions are made.

Different levels of decision making requires different types of information as under:

a. Strategic decision-making requires strategic information. Strategic decisions are characterized by a great deal of uncertainty and are future oriented. It includes activities like establishing policies, policy making, organizing and attaining an overall effectiveness for the organization.

b. Tactical decision-making requires tactical information. This pertains to short term activities and allocation of resources for the attainment of the objectives. At the tactical level of decision-making, standards are fixed and the results of decisions are deterministic.

c. As decision-making involves broadly intelligence, design of course of action, and choice of appropriate course of action, the information needs at each stage can be satisfied by information from internal sources, prior experiences and the information about the environment where the decisions are going to implemented.

Research and Development and Information Needs

Research is the most important activity for any society or industry for its development. Research attempts to find solutions to the problems being faced by a society or an organization. It is very rigorous process and involves processing and use of information to generate new knowledge. The generated information at different steps is contained in various documentary forms such as periodicals, reports, thesis, conference proceedings, review monographs, etc.

Research and development activities involves two activities:

a. Basic and fundamental research
b. Applied research and technical development

The information needs of researchers involved in the activity of research and development should be satisfied. The R&D professionals make use of wide range of information of direct relevance to the topic of research. Researchers require information for following purposes:

a. To aid in perception or definition of problem
b. To formulate a scientific or technical solution
c. To place work in proper context
d. To relate work ongoing research in progress
e. To select design/strategy for data collection
f. To select a data gathering technique
g. To design or select equipment or apparatus for conducting a study
h. To enable full interpretation of the collected data
i. To integrate findings into current state of knowledge

Seeing the above uses of information, it can be said the R&D scientists have to be supplied with the adequate information of right order at right time. The R&D activity is paramount for the socio-economic development of a nation.

Information Needs in Business

Any business operates in an environment that consists of economic, legal, political, social and technological factors. Each factor creates need for different types of information needs. The information needs vary from very general type of information to more detailed information relating to different aspects of the business.

Some of the information needs can be listed as:

a. Capital procurement and mobilization
b. Technical know-how

c. Knowledge of existing policies, practices and regulations.
d. Market conditions and requirements
e. Foreign trade
f. Management information

The information needs mentioned above if adequately satisfied with the right information the business excels.

5. Information Seeking Behaviour of Users

In the above sections we discussed about the information need, its types and the different information needs that exist in various human activities. The moment the information need is realized it becomes important to satisfy this need. As already discussed NJ Belkin ASK (1968) model proposes that, information need is an anomaly that arises in the users' state of knowledge. It is but natural for the human mind to fill this anomaly. There comes the concept of information seeking behaviour(ISB).

T D Wilson (1981) stated ISB as, the attempt of the user in obtaining the needed information results from the recognition of some need, perceived by the user. In order to satisfy the information needs, the user actively undergoes the information seeking process. In other words ISB can be defined as strategies and actions undertaken to locate discrete knowledge elements to satisfy the information need. The behaviour may take several forms -- the user may make demands upon formal systems such as information systems or upon other systems which may perform information functions. The user also seeks information from other people through information exchange, which involves an element of reciprocity, recognized by sociologists as fundamental aspect of human interaction. During the process failure may be experienced with the system as well from other sources when seeking information. Dissatisfaction after the use of information may lead to generation of new information need.

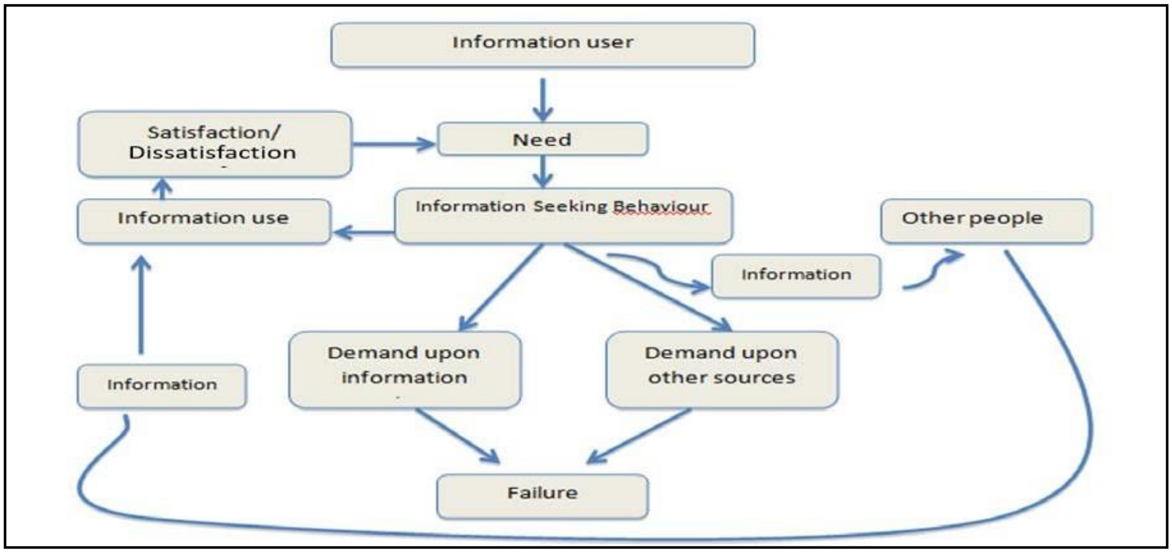

Fig.1: Information seeking behaviour of users

According to Girija Kumar (1990), information seeking behaviour involves following processes:

a. Identifying objectives
b. Defining needs
c. Accessing information systems
d. Establishing sources of information
e. Information acquisition
f. Use of information
g. Satisfaction/ dissatisfaction

Information seeking behaviour is concerned with the integrative utilization of three basic resources:

a. people,
b. information, and
c. system.

It is said, the behaviour that yields the highest information satisfaction is the best information seeking behaviour.

6. User Studies

In this section we will discuss the various methods generally known as user studies carried out tofind the pattern of overall interaction of user with the information retrieval system. Studying the behaviour and information needs of users in a systematic manner can help in increasing the efficiency of the system. Since last two decades user studies has gained lot of attention in library and information science. The user study aims to understand the processes of information transfer,which will further lead to improvement in the similar systems.

The term user study is mainly concerned with studying information processing activities of users. It essentially implies the study of the use of the demand or need of information. User study mainly focuses on 'users' to measure their information, needs, use behaviour and use pattern. Also for a meaningful service user study is must.

Menzel (1966) has categorized the user studies into three categories as under:

a. Behaviour Studies
b. Use studies
c. Information flow studies

Studies which are carried out to find out the pattern of overall interaction of the user community with the communication system, without reference to any specific information receiving event are called information behaviour studies.

Studies which are conducted to find out the use of any communication medium such as primary periodical, secondary periodical, other sources are called use studies. The studies which are conducted to find the pattern of flow of information in the communication system are the information flow studies. The factors of user study include:

a. identification of user group
b. assessment of use information needs
c. identification of user approaches and attitudes in finding, locating, and obtaining the information
d. orienting the user in finding, locating, and obtaining the information
e. matching the user and his information in such a way that the maximum benefits could be derived from the system.

Reasons for conducting user studies:

The following reasons could be pointed for conducting a user study:

a. identifying the actual systems and weaknesses of library resources and services,
b. identifying the levels and kinds of user needs,
c. identifying faculty and student priorities for library and information resources and services.
d. identifying the limitations or problems which seem to discourage the use of the system.
e. identifying the level of involvement of user in the system.
f. improving the organization and planning of the overall system.

The user study may also adopt interdisciplinary approach to the study of the user. There may be psychological or sociological approach. The aim of the user study is to develop dynamic interface between the system and the user. Hence the scope of user study is quite wide and ever increasing in its dimension.

7. Summary

In this module we tried to understand the answers to the following questions:

a. Who are the users and how can we categorise them?
b. What are their needs, and of what is the nature of their needs?
c. What are the specific information needs in some areas of human activities?
d. How the users seek the required information and by depicting what kind of behaviour?
e. What is the use pattern they exhibit in using the information?

Finally we understood the various aspects of users, their categories and nature. The concept of information need and types of information needs and more specific information needs in different areas of activities are also explained. The information seeking behaviour of users in order to satisfy their information needs is also discussed. Also the various methods generally known as user studies carried out to find the pattern of overall interaction of user with the IRS is also discussed.

8. References

1. Atherton, Pauline A. Handbook for information systems and services. Paris: UNESCO, 1977.

2. Belkin, N.J. Anomalous state of knowledge as basis of information retrieval. Canadian Journal of Information Science. 29, 1968.

3. Debons, Anthony et al. Information Science: an integrated view. Boston: G.K. Hall, 1988

4. Garvey, William D. The dynamic scientific information user. In communication: the essence of science. by William D Garvey. New York: Pergamon Press, 1979

5. Girja Kumar. Defining the concept of information needs. In Binwal, J.C. et al. Social Science information: problems and prospects. New Delhi :Vikas Publishing House, 1990.

6. Gorry, G.A. and Morton, M.S. Scott. A framework for management information systems. Sloan Management Record. 1977.

7. Kumar, P.S.G. Library and users. Delhi: BRPC, 2005

8. Line, M.B. Draft definitions: information and library needs, want, demand and uses. ASLIB Proceedings. 26, 1974.

9. Menzel, H. Information needs and uses in science and technology. ARIST, 1, 1966.

10. Olanigan, S.A. Information needs of the consultants to business enterprises. International Library Review. 19, 1987.

11. Paisley, W.J. Information needs and users. ARIST, 3, 1968.

12. Prasad, H.N. Information needs and users. Varanasi: Indian Bibliographic Centre, 1992.

13. Taylor, R.S. Question negotiation and information seeking in libraries. College and Research Libraries. 29, 1968

14. Wilson, T.D. On user studies and information needs. Journals of Documentation. 37, 1981.

Unit-4
Evolutions in Information Retrieval

1. Introduction

The growth of Information and Communication Technology (ICT) has influenced the wayinformation is being searched and retrieved. It has brought revolution in IR. There are several advancements that have taken place in this area over the period of time.

In this module, we discuss some of the IR techniques and technologies that evolved in the recent past. We discuss some of the significant IR standards and protocols such as Z39.50, SRW/SRU, and CQL. We also report the state-of-the-art research in IR field, for instance, the initiative of the global digital library, application of intelligent systems like an expert system in library cataloguing, classification and abstracting, the application and the issues of intelligent hypertext and hypermedia systems, and the research on human-computer interaction.

2. Information Retrieval Standards and Protocols

A standard means an agreement by what way to perform a task or carry out some activity to obtain a predictable result [6]. All standards available by the National Information Standards Organization (NISO), U.S.A are established by an agreement, which is based on the expertise ofdesigners, application developers and vendors, and product users. All the standards available by NISO are approved by the American National Standards Institute (ANSI). There are various standards and protocols exist today for IR systems. In the following sections some of the very popular search and retrieval standards and protocols, such as, Z39.50 [1], CQL [4], SRW [3], andSRU [2] are discussed.

2.1 Z39.50

Z39.50 is used both at the national and international level as a standard protocol that defines computer-to-computer information retrieval technique. It is a non-proprietary and vendor- independent. Z39.50 was originally approved by the National Information Standards Organization (NISO) in 1988. In 1998, International Organization for Standardization (ISO) adopted Z39.50 and issued ISO 23950 Information and documentation - Information retrieval (Z39.50) [6]. Using Z39.50 a user through his/her system can search and retrieve information from other Z39.50 compliant computer systems without having the prior idea about the syntax ofsearch that is used by the other systems. The primary goal of Z39.50 is to reduce the complexity and difficulties involved in searching and retrieving electronic information [6]. Z39.50 makes the life of the end-users simple to search and use the wealth of information available on the Internet. In Z39.50 enabled system environment, when a user of one system search for an information in another system, he does not need to know how the other system works.

Z39.50 operates in a client/server architecture. The protocol acts as a common language that all Z39.50-enabled systems can understand. It is like the Esperanto language, which bridges several “languages and dialects” that various information systems “speak” [6]. The communication and interoperation for Z39.50 take place both at the client and the server, hence, they must be able to speak the same Z39.50 language. TCP/IP [57] Internet communications protocol is used as a standard by almost all Z39.50 implementations to connect the systems and compliant software ofZ39.50 to translate between them for searching and retrieval of information.

Since all the associated technical activities occur behind the scenes, the users only see their familiar search and display interface. Z39.50 standardizes the messages used by clients and servers for communication, regardless of what software, platform or systems are used for achieving the interoperability [6]. A Z39.50 enabled client system can communicate with diverse servers. Similarly, a Z39.50 enabled server is searchable by client systems developed by differentvendors.

2.1.1 How does Z39.50 Works?

As stated above, a user on a client system can search through Z39.50 enabled interface without knowing how a server system works. Z39.50 governs the entire process of how a client translates the query into a standard format to send to a server [6]. After receiving the query, the server applies the Z39.50 rules to translate the query into a format that the local database understands, performs the search and sends the result to the client system. After receiving the result, the interface software at the client processes the results returned through Z39.50 with the goal of displaying them as close as possible to the way records are displayed in the user's native system [6].

SRW

SRW stands for **Search/Retrieve Web Service** protocol [3]. Its aim is to minimize the cross- language problems. The goal is to allow access to several networked resources and support interoperability among distributed databases, using a common utilization framework [3]. It is developed by collective implementers with more than 20 years of experience of the Z39.50 Information Retrieval protocol with nascent developments in the technological arena of the web.

SRW provides both Simple Object Access Protocol (SOAP) [36] and URL-based access mechanisms to a wide range of possible clients from Microsoft's .Net initiative to simple Extensible Stylesheet Language Transformations (XSLT) [35] and JavaScript transformations. This influences the Contextual Query Language (CQL) (discussed below) that provides an expressive but intuitive ways to search formulation [3]. SRW directs the usage of open and industry-supported standards like eXtensible Markup Language (XML) and XML Schema, and where desired, SOAP and XPath [3].

SRW provides semantics in search of databases having metadata and objects, both text and non- text [3]. As SRW has been developed on Z39.50 semantics, this makes easier for existing Z39.50 systems decreasing the barriers to new information providers, to enable their contents available via a standard search and retrieve mechanism. SRW defines web service incorporating several Z39.50 features, most notably, the Search, Present and Sort Services [3].

SRU

SRU stands for **Search/Retrieve via URL**. It is a standard XML-based protocol for search by utilizing CQL (http://www.loc.gov/cql/), a standard syntax for query representation [3]. The prime difference between SRU and SRW is that the former uses HTTP as the transport mechanism and the latter is based on SOAP protocol and uses XML streams for both the query and the results. This depicts that the query is communicated as a URL and the XML is receivedas if it were a web page. POST method an alternate way for using the HTPP transport technique is not acceptable in SRU protocol. The advantage is that a wide variety of transport mechanisms can be used in this case for instance e-mail.

CQL

CQL stands for Contextual Query Language (formerly known as, Common Query Language, http://www.loc.gov/cql/)). It is designed for use with SRW which is a search protocol successor to Z39.50 (as discussed in the previous section). CQL is an abstract and extensible query language for maximum interoperability amongst the connected systems. The goal is to reduce thedifficulty to learn and use while retaining the capability to allow complex searches. Primarily CQL is used in the bibliographic domain, but it is not restricted to this context alone. CQL provides standards and tested mechanism to specify a query to select records from a database that may be used either internally or remotely [3]. The bibliographic database of Library of Congress has an SRW/CQL interface available for all 28 million records. CQL can also be used in OpenOffice to identify, locate and integrate a huge amount of data within the application [3].

3. Global Digital Library

With the rise in the worldwide use of the open systems such as the Internet and World Wide Web(WWW) has enabled us to experience several real world entities in cyberspace like "virtual libraries" [42]. It is the role of librarians to take active participation and work together for reducing the issues and problems related to the information framework, which has a much globalpresence.

Global Digital Library (GDL) [42] is a prototype which aims to connect several national libraries and some major libraries, museums, archives, and information organizations with each other. Undoubtedly, there exists a need for cooperation globally in the field of building "digital" knowledge and sharing them in this digital information age.

Paradigm Shift towards a Global Learned Society

The growth of information and communication technologies, there is an increasing need to have access to information globally in order to have a much finer and better picture of the society in which we are living [42]. We are more curious about knowing our culture, our surroundings, our history, our economy, our growth in science and technology, etc. Nowadays, information has become a key to productivity and with the progress in technology, economic progress, and societal change, libraries have new challenges to face. This change wants our libraries to not only provide or satisfy the information needs of users who visit library but also to provide access to services and information resources to the users not present physically in the library i.e. user at home, at work, in school, or in any place where they need them [12].

Challenges

In the last few years, there is a rapid growth in the use of Internet for uploading digital contents on the World Wide Web (WWW) needed for non-commercial and commercial purposes. Nowadays it is a matter of a few seconds for us to write, talk, confer with, or send textual, audio and visual content to desired person in any part of the world [42]. The pattern for information seeking and use of a library and other information services and their delivery has changed dramatically. With the physical library, digital library has also become a reality. Now it is essential to share information physically as well as over the cyberspace to satisfy the need for both types of users.

Obstacles to Universal Access

There exist a lot of hurdles and issues related to the information infrastructure. Some of these difficulties are [7][40][41][42]:

- Several legal issues may arise related to intellectual property, copyright, confidentiality and privacy, security, personal, business equity, etc.;
- Difference in culture may influence the way of information communication;
- The presence of generational gaps;
- The sheer complexity of information architecture both at the global and national level;
- To have an effective and adequate inventory of available resources comprising the knowledge of information;
- The ability to locate, identify and retrieve relevant and quality information;
- Due to the huge amount of information, the complexity arises related to "undesirable" "indecent" information.

Irrespective of these difficulties and unsolved issues, it is expected that the relevant technologies will soon be available which will enable us to link all the global information to form "The GlobalLibrary" and delivery of multimedia information [12]. Note that in the recent time we have seen various initiatives of digital libraries (mostly in terms of institutional repositories), but they are mostly dispersed in nature. What is still missing, or what should be done is the global initiative by the major libraries and information organizations and works together towards finding the solutions for the above difficulties and issues. Hence, to build a true “The Global Library”, an effective, substantive and higher level of cooperation between the library and information leaders is needed both at the global and national level [42].

4. Intelligent Information Retrieval

Intelligent information retrieval, as defined by Sparck Jones [43], is a computer system having the capability to infer knowledge with the help of its previous knowledge for establishing a link between the requirement of its user and a set of candidate document. This is a system which can perform intelligent retrieval. The realization of researchers to use knowledge in the information retrieval system has led them to think about the artificial intelligent system which also has the similar purpose, and one among these classes is an expert system.

Expert System (ES)

As Peter [58] stated in artificial intelligence, an expert system is “a computer system which emulates the decision-making ability of human experts.” The expert systems are designed to solve complex problems by reasoning over knowledge stored in a knowledge base. The knowledge in the knowledge base is primarily represented as IF-THEN rules rather than conventional procedural code [59]. The first expert systems were invented in the 1970s and then proliferated in the 1980s [44] [45] [46]. The expert systems were the first among the true realization of Artificial Intelligence software.

An expert system has two primary components: inference engine and knowledge base [47] [44]. A knowledge base consisted of rules and facts, i.e., the knowledge about the real world objects. Inference engine applies these rules to the known facts for deducing new facts or knowledge. Inference engines can debug and can also provide the explanations for the deduced knowledge. The knowledge bases are designed in a similar fashion like the object oriented programming and stores the knowledge in a structured form. The knowledge in a knowledge base is structured in a form of classes, subclasses, and instance.

As expert systems evolved, several new techniques were adopted into various types of inference, engines. Some of the most important of these, as mentioned by Mettrey are [48]:

- Truth Maintenance: These systems record the dependencies in a knowledge base, so that when facts are changed the dependent knowledge also get altered accordingly. For example, if the system learns that Aristotle is no longer known to be a man, it will repeal the assertion that Aristotle is mortal.

- Hypothetical Reasoning: In hypothetical reasoning, the knowledge base can be sub-divided up into several possible views, or worlds. This enables the inference engine to explore several possibilities in parallel. In case of the previous example, the system may want to find the consequences of both the assertions, what will be true if Aristotle is a Man and what will be true if he is not?

- Fuzzy Logic: This was one of the first extension of the simply using rules for representing knowledge. The main idea behind fuzzy logic is to associate the probability with each rule. Hence, not to assert that Aristotle is mortal, instead to assert that Aristotle may be mortal with some probability value.

- Ontology Classification: With the introduction of object classes in the knowledge bases, a new kind of reasoning is enabled. With this addition of object classes, it is possible now to reason about the structure of the objects, instead of just simply reasoning the value of the objects. In the case of the above example, Man can be represented as an object class and R1 can be redefined as a rule which defines the class of all men. These types of specific inference engines are known as classifiers. As Mettrey stated [48], classifiers are very effective for unstructured volatile domains and a key technology for the Internet and the emerging Semantic Web [60].

Expert System for Library Professionals

Expert systems can be used to emulate the jobs of library professionals in a library. They can be applied where the intelligent activities are involved and generally carried by the library experts. Note that primarily the expert systems have a more specific goal to achieve than an information system. Expert systems are to make decisions, and not just to produce reports [16, p.91]. In this section, we discuss some of the significant expert systems specifically built for libraries.

Expert System in Cataloguing

A system called AUTOCAT [20] was produced in Germany. The system was designed to generate bibliographic records of physical sciences periodicals available in machine-readable form. Another significant work was done by Weibel, Oskins and Vizine-Goetz [21] at OCLC. They built a prototype based on rules known as "the OCLC Automated Title Page Cataloguing Project." The tool was designed to prepare descriptive cataloguing from the title pages. Another important project, namely, Qualcat (Quality Control in Cataloguing) was undertaken at the University of Bradford. The goals of the project were to develop expert systems to select the best records, to link the databases and centralized authority control, to build a fully automated control package for day to day running, and to investigate interface problems for cataloguing [61].

Expert System in Classification

Classification is a difficult task to accomplish using an expert system. This is even true for a human expert. The main problem is although the schedules are available to determine subjects and class numbers, the relationships between the objects (here, the documents) and classes are often not explicit. There are some expert systems that have been developed on item, patent and book classification, for instance, by Sharif [22]; Cosgrove and Weimann [49]; Valkonen and Nykanen [50]; Gopinath and Prasad [23].

In 1986, Paul Burton conducted an exploratory research at the University of Strathclyde, United Kingdom. The aims were to assess the merits of different ways of knowledge representation and suitability of expert systems in classification [39, p.64]. As an outcome of the experiment, a prototype expert system was designed. The system was able to provide the Dewey class number based on the information provided by the users. In another research, OCLC developed an expert system, called Cataloguer's Assistant. The system was tested in Carnegie-Mellon University to reclassify the mathematics and computer science collection [39, p.64].

Expert System in Document Delivery

There are very few references available related to expert systems in document delivery [39]. Brown [51] explained the use of expert system technologies in equipment division of Raytheon Company's for coordinating requests of specifications and standards documents with purchases made via the acquisitions unit. Abate [52] reported about an ES in the library of a law firm which was developed for delivery of document in decision making using the ES shell and VP-Expert.

Expert System in Abstracting

The researches on abstracting have been primarily focused on abstracting the scientific articles from journals and conference proceedings. The first reported work on automatic abstracting was in 1958 by H.P Luhn. There are few other works also have taken place on automatic abstracting, for instance, DeJong [53], Lebowitz [54], and Husk [55]. DeJong developed the system known as FRUMP which analyses articles from newspapers using frame-based techniques. The articles were first scanned and then data were automatically fed into the different slots within frames. Scripts were then executed to generate summaries of the information held in the relevant frames. Besides generating abstract for scientific articles, the research on abstracting also extended to other kind of materials. For instance, Rau, Jacobs and Zernik (1989) [18] developed a system known as SCISOR that generate reports on corporate acquisitions and mergers.

Besides the above areas, expert systems have been developed in many other areas of LIS. For instance, expert system in acquisition, in collection development and in indexing. Expert systems have been also used as an intelligent intermediaries for database selection, for query formulation,and so forth. For further details on expert systems in various areas of LIS, students are suggested to go through the review article [39]. It worth to mention here that a very few expert systems are reportedly operating in practice in the various areas of LIS. Some systems have progressed commercially but have later failed and been withdrawn from the market (e.g., Tome Searcher and Tome Selector).

5. Hypertext and Hypermedia Systems

Hypertext

Hypertext is a text which is displayed on a computer screen or other digital device with references (hyperlinks) to other text that a reader can access, or where text can be followed progressively at multiple levels of detail [62]. WordNet 2.1 (https://wordnet.princeton.edu/) defines hypertext as a “machine-readable text that is not sequential but is organized so that related items of information are connected.” The hypertext words connected through hyperlinks can be clicked by a mouse or by touching the screens. Apart from linking between the texts, the linking can be created between pictures, between tables or any other presentational content formsusing the hyperlinks. Web Pages in World Wide Web are mostly written in Hypertext Markup Language (HTML). This provides an efficient, flexible connection and sharing of information on the Internet. Hypertext documents are of two types either static (i.e. prepared and stored in advance) or dynamic (continually changing according to the response of user’s input) [62]. Statichypertext is used for the cross-references of a collection of data in documents, books on CDs/DVDs, and the web pages, etc. The most significant and popular implementation of hypertext is the World Wide Web. In 1963, Ted Nelson coined the word “hypertext.”

Hypermedia

Hypermedia, a logical extension of hypertext, is a non-linear medium of information space which includes plain text, audio, video, graphics and hyperlinks link [63]. Hypermedia contrasts with its broader term multimedia (a content consists of varieties of content, such as, text, images, animation, and video), which can be deployed to describe non-interactive linear presentations along with hypermedia. The term was coined by Ted Nelson in 1965. The World Wide Web is a classic example of hypermedia, whereas a non-interactive cinema presentation is an example of standard multimedia. It is non-interactive due to the lack of hyperlinks.

Information Retrieval Based on Hypertext and Hypermedia

The research on online search is focused on designing systems that would assist the professional intermediaries to retrieve a smaller set of result from a relatively larger set of records, e.g., scholarly journal abstracts, library catalogue records, etc. [64, p.105]. The focus is primarily on designing systems that would help or replace the professional intermediaries. Professional online searchers perform the search in a systematic manner. They clarify the search queries with the users before they provide the result. To search, the professionals in advance plans the search, consult the thesauri, and combine the terms using the logical operators (AND, OR, NOT) and also by adjusting proximity limits (the set of words within which query terms must co-occur) and scoping limits (the set of documents over which search takes place) [64, p.105].

Today's electronic retrieval systems were mainly designed to replace the professional intermediaries or to emulate their performance. The systems focused, on indexing and cross- referencing for the organization and retrieval of the resources, instead on meaning, readability and understanding of information. Hence, the systems designed for end users must follow this philosophy and also constitute the suitable information seeking strategies [64].

Hypertext systems have the difference with respect to the present online retrieval systems. The online retrieval systems encourage the personalized, informal and content-oriented information- seeking strategies. In Hypertext system users can provide information during the retrieval process with the help of getting the context, and during browsing by saving, connecting, or transferring images or text [64, p.106]. Further, the current research trend is to support end users by providing flexible and powerful interfaces between the people (the end user) and computer. The idea is that the smart interface would be able to balance the end user browsing patterns usingefficient analytical techniques similar like those used by the professional intermediaries.

In case of Hypermedia systems, they exhibit various types of relationships among the information elements. Typical examples of similarity relationships include similarity in meaning, similarity in logical sequence and temporal sequence, containment, etc. [65] [66]. Hypermedia allows these relationships to be used as links which as a result enables content navigation within the information space. Based on this, we can also build the taxonomies of links, which we can further discuss and analyze how best they are utilized. One possible example of a taxonomy could be based on mechanics of links (i.e., single source with single-destination, multiple-source with single-destination), the directionality of connections (i.e., unidirectional, bidirectional), and the mechanism for anchoring (i.e., generic links, dynamic links). Another alternative taxonomy example could be based on types of information relationships represented, in particular related tothe organization of information space (e.g., structural links), related to the content of informationspace (e.g., associative and referential links) [65][66].

6. User Interface

In present era as the digital repositories are growing in terms of volume as well the diversified information content, the need for effective information retrieval systems are becoming increasingly high. In this section, we focus on an important component “interface” of an information retrieval system. We discuss the issues related to using and designing effective interfaces.

IR as a Problem Solving Process

Henninger, S. and Belkin [67] have divided the field of information retrieval into two categories: system-based and user-based. According to them, the system-based IR is concerned with the efficient search techniques for matching with the query and document representations. While the user-based IR is concerned with the cognitive state of the searcher and the context in which a problem is to be solved.

Generally speaking, an user of an IR system feel lonely or seek help whenever they perceive that they lack some knowledge to perform a task or to solve information search problem. This happenin particular when an user is searching for something which s/he knows little or nothing. In this context, it is expected that an IR not only just retrieve the information, but also help the users to describe and formulate the query. The idea is the system not only provides good query language, but also need to support an interactive dialogue model [67]. An iterative interface can interact with the users like the way professional intermediaries does and can understand and render help in solving their search problems.

Issue of Vocabulary

It is a very common and well-known problem of IR. It is quite often observed that even though the information is available in a repository, the users do not find it. It is due to the missing link between user’s search term and the IR defined terms, which are mostly assigned by the professionals. Users use multiple terms to refer to the same thing. So, unless there is match between the user search term and the IR defined terms, the result is going to be null. Also the situation becomes further worst when there is a mismatch between the ways a user characterize an object and the way a professional sees it. In most of the repositories, professionals describe the resources based upon the properties of the objects, which are mostly the inherent properties. Unless the users are also able to perceive and characterize in the same way, the query is likely to fail. It is noted in [67] that users look for information that is used for something and therefore forthem how an object is used has greater relevance than its inherent properties.

Interfaces for Retrieval Systems

Current IR systems have addressed these some of inherent properties of information seeking strategies and indexing in various ways. Browsing has been placed to facilitate the iterative and ill-defined information seeking strategies [67]. A mixed approach including the support for information seeking strategies, such as, browsing and direct query facility, information visualization, feedback mechanism, etc. can be employed in designing an effective and efficient IR interface.

Interface Design Strategies

The design strategies of information retrieval systems must not only address problems related to look-and-feel, but it should also address the issues of ill-defined information seeking strategy including others. Besides designing an attractive and presentable interface (i.e., good look-and- feel), an interactive interface design is also crucial. Dialogue models based on relevant feedback and reformulation of query addressing the ill-defined nature of information seeking would enableusers to learn from the repository and refine their information need. The IR systems need to havesupport for number of interaction patterns, such as making query and browsing for satisfying various kinds of search techniques users may need to use [67].

7. Summary

In this chapter we have discussed some of the IR techniques and technologies that evolved in the recent past. We have discussed some of the significant IR standards and protocols. We have also reported the state-of-the-art research in IR field, for instance, the initiative of global digital library, application of intelligent systems like expert system in library cataloguing, classification and abstracting, the application and issues of intelligent hypertext and hypermedia systems, and the research on human computer interaction.

8. References

1. The Z39.50 Information Retrieval Standard. Retrieved from http://www.dlib.org/dlib/april97/04lynch.html. Accessed on Sep. 20, 2014.

2. SRU: Search/Retrieve via URL. Retrieved from http://www.loc.gov/standards/sru/companionSpecs/srw.html. Accessed on Sep. 20, 2014.

3. Apache OpenOffice: The Free and Open Productivity Suite. Retrieved from http://www.openoffice.org/bibliographic/srw.html. Accessed on Sep. 20, 2014.

4. The Contextual Query Language. Retrieved from http://www.loc.gov/standards/sru/cql. Accessed on Sep. 20, 2014.

5. Buntine. W., Taylor, M. P., & Lagunas, F. (2006). Standards for Open Source Information Retrieval, In Proceedings of the Open Source Information Retrieval Workshop (OSIR).

6. Z39.50 A Primer on the Protocol. Retrieved from www.niso.org/publications/press/Z3950_primer.pdf. Accessed on Sep. 20, 2014.
7. Alexandria Declaration of Principles (prepared by Robert M. Hayes and Ching-chih Chen) (1995). Microcomputers for Information Management, 12 (1-2), 3-8. Also In Planning Global Information Infrastructure, ed. by Ching-chih Chen. Norwood, pp. 1-6.

8. Bobinha, José Luis B. and Delgado, José Carlos M. (1996, November). Internet and the New Library. In Proceedings of NIT '96: The 9th International Conference on New Information Technology, November 11-14, 1996, Pretoria, South Africa. Ed. by Ching-chih Chen. Newton, MA: MicroUse Information. pp. 1-12.

9. Chen, Ching-chih. (1986, December). Libraries in the information age: Where are the microcomputer and laser optical disc technologies taking us? Microcomputers for Information Management, 3 (4), 253-266.

10. Chen, Ching-chih. (1990, December). The challenge to library and information professionals in the visual information age. Microcomputers for Information Management, 7 (4), 255-272.

11. Chen, Ching-chih. (1993, November). Thriving in the digital communications environment: PROJECT EMPEROR-I's experience -- From print to multimedia, analog to digital. In Proceedings of NIT '93: The Sixth International Conference on New Information Technology (PP. 59-68). West Newton: MicroUse Information. pp. 59-68.

12. Chen, Ching-chih. (1994, September). Information superhighway and the digital global library: Realities and challenges. Microcomputers for Information Management, 11 (3), 143-156.

13. Lynch & Garcia-Molina. (1996). Interoperability, scaling, and the digital libraries research agenda. Microcomputers for Information Management, 13 (2), 85-132.

14. Negrosponte, Nicholas. (1996, May). Caught browsing again. Wired, p. 200.

15. Ziegler, Bart. (1994, May 18). Building the highway: New obstacles, new solutions. The Wall Street Journal, pp. B1, B3.

16. Schoech, D. (1999). Human services technology: understanding, designing and implementing compuer and internet applications in the social services (2nd ed.). Haworth, New York.

17. Morris, Anne. (ed.) 1992. The Application of Expert Systems in Libraries and Information Centres. London : Bowker-Saur.

18. Rau, L. F.; P.S. Jacobs and U. Zernik. 1989. SCISSOR : information extraction and text summarization using linguistic knowledge acquisition. Information Processing and Management, Vol.25 no.4: 419 - 428.

19. Richardson, John. 1989. Towards an expert system for reference service: a research agenda for the 1990. College and Re-search Libraries, Vol.50 no.2: 230 - 248.

20. Endres-Niggemeyer, B. and G. Knorz. 1987. AUTOCAT: knowledge-based descriptive cataloguing of articles published in scien-tific journals. Knowledge Based Systems. Munich, October 20 – 21, 1987.

21. Weibel, Stuart; M. Oskins and Diane Vizine-Goetz. 1989. Automatic title page cataloguing: a feasibility study. Information Processing and Management, Vol.25 no.2: 187-203.

22. Sharif, Carolyn A. Y. 1988. Developing an expert system for classification of books using micro-based expert systems shells. British Library Research Paper 32.

23. Gopinath, M. A. and A.R.D. Prasad. 1994. A knowledge representation model for Analytico-Synthetic classification. In: Know-ledge Organization and Quality Manage-ment. Proceedings of the 3rd ISKO Con-ference; 20-24 June 1994; Copenhagen, Denmark. pp. 320 – 327

24. T.H. Nelson, Literary Machines, Swarthmore, Penn., 1981. Available from Nelson.

25. Conklin, “Hypertext: An Introduction and Survey,” Computer, Sept. 1987, pp.

26. D. Goodman, “The Two Faces of Hyper- Card,” Macworld, Oct. 1987, pp. 122-129.

27. B. Shneiderman, Designing the User Interface: Strategies for Effective Human Computer Interaction, Addison-Wesley, Reading, Mass., 1987.

28. Belkin, N.J., Oddy, R.N, Brooks, H.M. ASK for Information Retrieval: Parts I&2, Journal of Documentation, 38(2,3), 1982, pp. 61-71; 145-164.

29. Card, S.K., Robertson, G.G. Mackinlay, J.D. The Information Visualizer, an Information Workspace, Human Factors in Computing Systems: CHI Ô91 Proceedings, ACM, 1991, pp.181- 194.

30. Furnas, G.W., Landauer, T.K., Gomez, L.M., Dumais, S.T. The Vocabulary Problem in Human- System Communication, Communications of the ACM, 30(11), 1987, pp. 964-971.

31. Kwasnik, B.H. How a Personal Document's Intended Use or Purpose Affects Its Classification in an Office, Proceedings SIGIR '89, ACM, 1989, pp. 207-210.

32. Thompson, R.H., Croft, W.B. Support for Browsing in an Intelligent Text Retrieval System, International Journal of Man-Machine Studies, 30, 1989, pp. 639-668.

33. Winograd, T, Flores, F. Understanding Computers and Cognition: A New Foundation for Design. Ablex, 1986.

34. Chowdhury, G. C. (1999). Introduction to Modern Information Retrieval. Library Association Publishing.

35. XSLT. Version 1.0. http://www.w3.org/TR/xslt

36. Simple Object Access Protocol (SOAP) 1.1, Don Box, David Ehnebuske, Gopal Kakivaya, Andrew Layman, Noah Mendelsohn, Henrik Nielsen, Satish Thatte, Dave Winer, Editors. DevelopMentor, IBM, Microsoft, Lotus Development Corp., UserLand Software, Inc., 30 July 2003. This version is http://www.w3.org/TR/2000/NOTE-SOAP-20000508/.

37. MindManager. http://www.mindjet.com/

38. IdeaFisher. http://www.innovationmanagement.se/imtool-articles/ideafisher-pro-the-perfect-catalyst-for-your-mind/

39. Silva, Sharon Manel De. (1997). A review of expert systems in library and information science. Malaysian Jour. of Lib. & Inf. Sc., vol. 2, no. 2, pp. 57-92.

40. Chen, Ching-chih, ed. (1995a). Planning Global Information Infrastructure. Norwood, NJ: Ablex. 518 p.

41. Chen, Ching-chih. (1995b). Digital global cultural and heritage information network: Personal viewpoint. In Planning Global Information Infrastructure, (pp. 181-185). Ching-chih Chen (Ed.). Norwood, NJ: Ablex. pp. 167-180.

42. Chen, Ching-chih. "Global digital library: Technology is ready, how about content?." Proceedings of NIT'96, Pretoria, South Africa (1996): 41-56.

43. Sparck Jones, K., 'Intelligent retrieval'. In: Jones, K. P. (ed.), Intelligent information retrieval: Informatics 7, London, Aslib, 1983, 136-42.

44. Zahendi, F., Intelligent systems for business expert systems with neural networks, Boston, MA, Wadsworth Publishing, 1993.

45. Ford, N., Experts systems and artificial intelligence: an information manager's guide, London Association Publishing, 1991.

46. Morris, A., 'Expert systems for library and information services: a review', Information processing and management, 2 (6), 1991, 713-24.

47. Brooks, H. M., ' Expert systems and intelligent information retrieval', Information processing and management, 23 (4), 1987, 367-82.

48. Mettrey, William (1987). An Assessment of Tools for Building Large Knowledge-BasedSystems. AI Magazine 8 (4).

49. S.J. Cosgrove, J.M. Weimann, (1992) "Expert system technology applied to item classification", Library Hi Tech, Vol. 10 Iss: 1/2, pp.33 - 40.

50. Valkonen, P. and Nykanen, O. (1991), “An expert system for patent classification”, World Patent Information, Vol. 13 No. 3, pp. 143 -8.

51. Brown, Lynne C. Branche. 1993b. Expert sys-tems and document delivery in an automa-ted acquisitions environment. Library Ac-quisitions: Practice & Theory, Vol.17, no.3: 353 - 357.

52. Abate, A. K. 1995. Document delivery expert. Journal of Interlibrary Loan, Document Delivery and Information Supply, Vol.6, no.1: 17 - 37.

53. DeJong, Gerald. 1982. An overview of the FRUMP system. In: Lehnert, W. G. and Ringle, M. H. (eds.). Strategies for Natural Language Processing. London : Lawrence Erlbaum, pp. 149 - 172.

54. Lebowitz, M. 1986. An experiment in intel-ligent information systems – RESEAR-CHER. In: Davies, R., (ed.). Intelligent Information Systems: Progress and Pros-pects.

Chichester, England : Ellis Hor-wood. pp. 127 - 149.

55. Husk, G. D. 1988. Techniques for automatic abstraction of technical documents using reference resolution and self-inducing phrases. Master’s dissertation. University of Lancaster.

56. Shneiderman, Ben. Designing the User Interface: Strategies for Effective Human-computer Interaction. Addison-Wesley, Reading, MA, 1997.

57. Internet protocol suite. https://en.wikipedia.org/wiki/Internet_protocol_suite

58. Jackson, Peter (1998), Introduction To Expert Systems (3 ed.), Addison Wesley, p. 2, ISBN 978-0-201-87686-4.

59. Expert system. https://en.wikipedia.org/wiki/Expert_system

60. Berners-Lee, T., Hendler, J. and Lassila, O. (2001). The Semantic Web: a new form of Web content that is meaningful to computers will unleash a revolution of new possibilities. Scientific American. http://www.scientificamerican.com/article.cfm?id=the-semantic-web

61. Qualcat: Automation of Quality Control in Cataloguing. British Library Research and Development, 1994, pp. 112.

62. Hypertext. https://en.wikipedia.org/wiki/Hypertext

63. Hypermedia. https://en.wikipedia.org/wiki/Hypermedia

64. Marchionini, G. and Shneiderman, Ben. (1993). Finding facts vs. browsing knowledge in hypertext systems. In "Sparks of innovation in Human-Computer interaction," edited by Ben Shneiderman (ed.). Ablex Pub., New Jersey, pp. 103-122.

65. Lowe, David and Hall, Wendy (1999). Hypermedia & the Web: An Engineering Approach. Willy.

66. What is hypermedia? http://users.ecs.soton.ac.uk/lac/LoweNHall/extracts/Hypermedia.html

67. Henninger, S. and Belkin, N. J. (n.d.). Interface Issues and Interaction Strategies for Information Retrieval Systems. http://www.sigchi.org/chi96/proceedings/tutorial/Henninger/njb_txt.htm

Unit-5

Web Based Information Retrieval

1. WorldWide Web

The World Wide Web (WWW) is a huge system based on client-server architecture withmillions of distributed servers worldwide. Here, every server maintains a collection ofdocuments and each document is stored as a file (although documents can also be generated on request). A server accepts requests for providing a document and then sends it to the client. Requests for storing new documents can also be made to servers.

TheWWWstartedasaprojectatCERN,theEuropeanParticlePhysics LaboratoryinGeneva,to provide access to shared documents using a simple hypertext system for its large and geographically dispersed group of researchers. A document in this context can be anything that can be displayed on a user's computer terminal, such as personal notes, figures, drawings, reports, blueprints, and so on. By connecting documents to each other, it becomes easy to integrate them from various places into a new document without the need for centralized changes. The only thing required was to establish a document providing connections to other relevant documents.
The easiest technique to refer to a document is by a reference known as Uniform Resource Locator (URL).This specifies the location of a document.

Here a client or a user interacts with Web servers with the help of a special application called browser. A browser has the responsibilityof properly displaying a document also, accepts inputs from a user usually by allowing the user select a reference to another document, that it subsequently fetches andthen displays.

DocumentModel

IntheWebenvironment allinformationisrepresentedintheformofdocuments.Therearemany ways to express a document in a web environment. Some documents are as simple as an ASCII text file, while others are done by a collection of scripts which gets automatically executed if the document is downloaded into a web browser.

A document can contain references to other documents in the form of hyperlinks. i.e. when a documentisopenedinawebbrowser,hyperlinkstootherdocumentsarevisibletotheusers.The user can follow a link by clicking on it.

HyperText Mark-up Language or simply HTML is the basic building block of the Web documents. This is a mark-up language that provides keywords to provide structure into various sections. For example, each HTML document is divided into a heading section and a main body. HTML also provides tags or keywords to distinguish headers, tables, lists and forms. It also enables to insert images or animations in a document. Besides these structural elements, HTML provides various keywords to provide instructions to the browser how to render the document. Below is a simple HTML code provided:

```
<HTML><!--StartofHTMLdocument
<BODY><!--Startofthe mainbody
<H1>HelloWorld</H1><!--Basic texttobe displayed
<P><!--Startofnew paragraph
<SCRIPT type = "text/javascript"><!-- Identify scripting language
document.writeln("<H1>HelloWorld</H1>");//Writealineoftext
</SCRIPT><!--Endofscriptingsection
</P><!--Endofparagraphsection
</BODY><!--Endofmainbody
</HTML><!--EndofHTMLsection
```

Naming

The WWW uses a naming scheme to identify documents, known as Uniform ResourceIdentifiers or simply URIs (Berners-Lee et al., 1998). URIs consists of a Uniform Resource Locator (URL) which identifies a document by including information on location of the document and the Uniform Resource Name (URN) which acts as true identifier by providing reference to a document.

2. TypesofInformationand Resources

An Information Source is a source of Information, i.e. anything that may provide information about something or provide knowledge about it. Different types of problems require different information sources. They may be categorised into Primary Sources, Secondary Sources and Tertiary Sources.

Primarysources

Primary sources are original information sources. They are from the time period involved and have not been passed through interpretation or evaluation. They are usually the first formal appearance of information in physical, print or electronic format. They present original thinking, share new information or report a discovery.

Examplesinclude:

- Artefacts(e.g.coins,allfromthetimeunder study);
- Audiorecordings(e.g.radioprograms)
- Internetcommunicationsonemail, listservs;
- Interviews(e.g.,oralhistories,telephone, e-mail);
- Journalarticlespublishedinpeer-reviewed publications;
- Newspaperarticleswrittenatthe time;
- Originaldocuments(i.e. birthcertificate,will,marriagelicense,trial transcript);
- Proceedingsofmeetings,conferencesand symposia;
- Records of organizations, government agencies (e.g., annual report, treaty, constitution, government document);
- Surveyresearch(e.g.,marketsurveys,publicopinionpolls).

Secondarysources

Secondary sources of information are those which are either compiled from or refer to primary sources of information. They are interpretations and evaluations of primary sources. Secondary sources are not evidence, but rather commentary on and discussion of evidence.

Examplesinclude:

- Bibliographies;
- Biographicalworks;
- Dictionaries,Encyclopaedias(alsoconsideredtertiary);
- Textbooks
- Monographs
- Website(alsoconsideredprimary).

Tertiarysources

Tertiary sources consist of information which is a distillation and collection of primary and secondary sources which include:

- Bibliographiesof Bibliographies;
- Directories;
- GuidestoLiterature

3. UsersInteractionand Search

Searching for the required information on Web can be a frustrating and disappointing experience due the availabilityof huge amount of information on the Web. These web contents are different from the traditional resources that are available in libraries and in online databases because Web contents are heterogeneous, networked and available in multimedia types. Such as text,hypertext, images, audio, video, etc. Here information creation is dynamic and beyond the physicalboundaries.On theWebuser'sprofileis heterogeneouslike theresources,andthemajor portion of which are novices with wide variety of subject backgrounds and different computer and web literacy. So, the user interaction with the different interface of search engines largely depends on the skills of users to frame the search queryin such a manner so as to bridge the gap between the way a programmer has indexed a particular content under what keyword and the actual requirement of the user. In the traditional Library environment a users approaches a reference librarian for the information need and it is the librarian who analyses the user's query and then provides with the relevant information. So, Librarians act as a bridge between the user and the information. But in Web environment users, themselves have to perform the task of the librarian, search engine will just provide all the documents matching user's query keywords.Then user has to perform the task of finding the relevant information from the collection.

4. DifferencebetweenclassicIRandWBIR

Information retrieval is a very important task done by every person to meet their daily life requirements. Their need and purpose may vary depending upon the users requirements. In the present web environment, we are surrounded by various types of digital resources and searching them have become a routine activity. But satisfying the information need of the users on the web is not a simple task. It is a difficult and time consuming task. In this scenario finding a relevant information is like searching a needle from haystack. Present web-based search tools have been inspired by classical information retrieval systems but due to the nature of web environment and the diversity in users' behaviours brings several challenges in searching and retrieval patterns. In the classic IR systems both the resources and the users were more or less predictable and homogeneous. The digital contents from online as well as offline databases, Online Public Access Catalogues (OPACs) mainly contain data stored in a structured manner. Due to thisnature of stored data, the search and retrieval process was much easier and more predictable. Accordingly, the user group were mainly comprised of people from academics, researchers, subject experts or librarians. They were well aware of the search keywords to use for finding a particular document. But with the emergence of Web there is a flooding of digital information in a very unstructured, uneven and heterogeneous manner so to cope up with this situation Web based IR systems evolved. These systems provide accessibility to web based digital contents. They use programs to maintain and update a list of web documents added to web at a regular interval or according to the need. Then these web IRs mainly look for the searched keywords in the maintained list or the index file to retrieve the original document present on the web. WebIRs user's interface is much more user friendly than the traditional IR systems by keeping inmind the issue of increase in inexperienced user group of web IR systems. So, now Web IR interfacedesignershavetolookmoreintotheinformationseekingbehaviourofthistypeofusers than in the classical IR systems where users were mainly experts of a particular domain under consideration.

5. SearchEngines

Search engines are computer programs that search for particular keywords entered by users and returns a list of documents in which they were found, it is especially a service that searches contents on the web. But they not only search for keywords rather some search for other things also and these are not “engines“ in the classical sense like mouse is not a “mouse” in digital world.

Typesof SearchEngines

Searchenginescan bemainlycategorised intofourtypes:

- Crawler-based search engines are useful if we have specific search keyword in our mind butif our search topic is a general one then these type of search engines may provide several irrelevant documents to a search request, e.g. AltaVista.
- Human-powered directories are good if our search is a general topic, then this type of search engines powered with human crafted directories will guide us and help toconverge oursearch and fetch refined responses, e.g. DMOZ.

- Hybrid search engines use a combination of both crawler-based results and directory results, e.g. Google.
- Meta-search engines are good for saving time by gathering results from different search engines at a single interface. It is excellent if we wish to know whether something is available about a particular topic or not on the web, e.g. Dogpile.

FeaturesofSearch Engine

The features like basic text search facilities, like Boolean search, proximity search, phrasesearch, truncation, field-specific search and limiting search are provided by almost all the search engines.

Booleansearchuses3differenttechniquessuchas:

- combinationofkeywords withAND,OR,NOT
- byprefixing‘+’or‘-’operators with keywords
- byselectingoptions like‘all ofthewords’as in Google.

Proximitysearchwhichlooksfortwoormorewordswhichoccurwithinacertainnumberof words from each other, e.g. Google supports ‘AROUND’.

- FieldSearchusing‘intitle’or‘allintitle’beforethesearchtermsinGoogle.
- Phrasesearch tosearchforaphraseinadocumentusingdoublequotes.
- Limitingsearchforadocument bydateorfiletypein Google.

The search interfaces in the modern day search engines enables users to use above features withoutmucheffort.Manyadvancedsearchinterfaces alsoprovidesenoughhelpinformationfor users to perform the search on the search interface itself.

AdvancedSearchEnginesand Applications

Present day search engines are like encyclopaedias operating on the internet, allowing users to search and retrieve relevant digital contents. But from users perspective only requirement is to search for a desired content using appropriate search engine. Because different search enginesare meant for different purpose and requires different skill set to use it. Advanced search engines will satisfy the most of the users queries by providing advanced search options, thus efficiently providing solutions to users queries.

Someofadvancedsearchengineswithsuchadvancedapplicationsareasfollows:

- **For General Search**: If users requirement is written information, the general search engines like Google is efficient one. Google with its advanced search options enableusers to perform more specific search queries.

- **Reverse Image Search**: If a users requirement to search for images then a advanced search engines like TinEye is a efficient one as this can read the content and thus making it searchable while a general search engines can look for only file names or user defined tags.

- **Similar Image Search**: The advanced search engines like GazoPa can look for similar features in the image like texture, colour or structures but cannot recognize exact copies of a given content.

- **Invisible Search**: The CompletePlanet advanced search engines have the application of searchingthe desired content from the data storedin databases which are almost invisible to the general search engines. Because general search engines mainlyindex the resources from the websites byfollowing the hyperlinks one after another. This type of hidden web is known as Deep Web.

- **SemanticSearch**:Semanticsearchismeantforsearchingtermsinameaningfulmanner i.e. terms with exact meaning, context and definition. The search engines like Yummly based on such type of semantic search algorithms are efficient in obtaining relevant results.

6. WebDirectories

Web directory also known as link directory is a type of directory available on the WWW. Its purpose is to provide links to other web sites and further classify those links. It is not like a search engine and do not display list of websites based on user searched keywords; rather this maintains alist ofweb sites accordingto class and subclass. This classification ofwebsites is not based on individual pages but on the complete website. Websites for their inclusion in a web directory in a particular category is either submitted by the site owner followed by reviewprocess for the approval or done by the web directory maintainer. Some of the Web directories are as follows:

- YahooDirectories istheoneofthebest andoldest directoryon theWeb
- The Open Directory is a human-edited directory. Also known as DMOZ (Directory Mozilla).

7. Ontology

Theconceptoriginated morethantwothousandyearsagofromphilosophyandmorespecifically from Aristotle’s theory of categories. The original purpose was to provide a categorization of all existing things in the world. Ontologies have been lately adopted in several other fields, such as Library and Information Science (LIS), Artificial Intelligence (AI), and more recently in Computer Science (CS), as the main means for describing how classes of objects are correlated, or for categorizing the document resources. Many definitions of ontologies have been provided. According to Gruber, ontology is defined as, “an explicit specification of a conceptualization”. Later on Studer et al extended the definition and defined ontology as "a formal, explicit specification of a shared conceptualisation". Studer’s definition includes the idea of shared in the notion of conceptualization and formal relations among the concepts. The explicit, formal representation of a shared conceptualization involves a perspective of a specific reality, and isconstitutedintheconceptualstructureofaknowledgebase.Theultimateobjectiveofontology is to share the knowledge it represents. An ontology defines the terms and their formal relations within a given knowledge area. The main features of ontology are:

i. Ontologyprovideashared understandingofdomains;
ii. Ontologyisusefultorepresentandtofacilitatethesharingofdomainknowledge between human and automatic agents;
iii. Ontologyis usefulfortheorganization andnavigation ofwebsites;
iv. OntologyisusefulforimprovingtheaccuracyofWebsearches.Websearchescan exploit the generalization and/ or specialization of information.

Semanticsearch

Enormous amount of information is produced everydayto caterthe needs ofhuman being. There are so many users who have different information needs and most of them are dependent on the web to fulfil their needs. Web is a big semi structured database which provides a vast amount of information. WWW constitutes of near to 12 billion web pages (Gulli, A., and Signorini, A.,2005). Through the rapid growth of web, it has become an easy way to access information. But for the rapid increase of information users are facing new challenges. So locating the precise information also raises a big challenging task. Moreover, most of the search engines primarily give priority to index huge amount of information-contained files or web-pages, so that, once searched it can give maximum number of hits. But, evaluating this enormous links by several parameters is strictly a statistical procedure. Semantic sense and relevance measures on the search space are still grey areas of research and raises many open questions. This issue can be solved using semantic search techniques.

Semantic search is a combined efforts of semantic web and search engine technologies which is designed to solve complex queries, automatic clustering and managing the large number of web documents. From Web point of view, each Semantic Web data is addressed by URLs and retrieved as form of Semantic Web documents. So, the Semantic Web is an extensive collection of static or dynamic semantic web documents (SWD) distributed over the Web-space. Meta- information to each ontology also executes the searching and matching operation very optimally and retrieves ontologies stored in an ontology registry, providing a compact representation for efficient search and reuse of related ontologies. Some of the semantic search engines are such as Swoogle and Watson.

8. Summary

Web based retrieval[WBIR] was introduced making a distinction between classic IR and WBIR.Features ofweb browsers and search engines were described.Hyper Text Markup Language (HTML) and use of various tags were discussed. Advanced search engines and their applications were discussed.Advanced concepts of semantic search in semantic web and search engine technologywere also covered in this module.

9. Reference

Peiling Wang et al. (1998) An exploratory study of user searching of the World Wide Web: a holistic approach. Proceedings of the 61st Annual Meeting of the American Society for Information Science, October 25–29, Pittsburgh, PA, pp. 389–399.

Mansourian, Y. (2004). "Similarities and differences between Web search procedure and searching in the pre-web information retrieval systems". Webology, 1(1), Article 3. Available at: http://www.webology.org/2004/v1n1/a3.html.

Croft, W. B., Metzler, D., & Strohman, T. (2010). Search engines: Information retrieval in practice (p. 88). Reading: Addison-Wesley.

A Comparison of Search Engines For Finding Resources by By Yuanlei Zhang, April 28, 2004, http://www.yuanlei.com/studies/articles/is567-searchengine/page2.htm

http://www.guidingtech.com/16116/google-search-little-known-around-operator/Accessed on Jun. 20, 2015.

http://www.webology.org/2010/v7n1/a76.html.AccessedonJun.20, 2015.

http://bcs.org/upload/pdf/ewic_tl06_s2paper3.pdf. Accessed on Jun. 22, 2015.

Aristotle'sCategories,2007.http://plato.stanford.edu/entries/aristotle-categories/

Dini, Luca (2004). NLP technologies and the semantic web: risks, opportunities and challenges. IntelligenzaArtificiale 1(1), pp. 67-71.

Gruber,T.R.(1993).Atranslationapproachtoportableontologyspecifications.Knowledge Acquisition, 5(2), pp.199–220].

Studer,R.,Benjamins,V.R.andFensel,D.(1998).Knowledgeengineering:principlesand methods.

http://www.das.ufsc.br/~gb/pg-ia/KnowledgeEngineering-PrinciplesAndMethods.pdf

Dutta, B., Chatterjee, U. and Madalli, Devika P. (2013). From Application Ontology to Core Ontology. In the Proceedings of International Conference on Knowledge Modelling and Knowledge Management (ICKM 2013), Bangalore, India. ISBN: 978-93-5137-765-8.

SemanticWebMadeEasy.http://www.w3.org/RDF/Metalog/docs/sw-easy

Antoniou,Grigoris and Harmelen, Frankvan.Asemanticwebprimer. London:MITPress, 2004.

"Webdirectory".Dictionary.AccessedonJul.30,2015.

WendyBoswell. "Whatis aWeb Directory".About.com. AccessedonJul.30, 2015.

http://websearch.about.com/od/enginesanddirectories/a/subdirectory.htm. Accessed on Jul. 30, 2015.

Basu, A. and Paul, M. (2015). A Case study on Semantic Web Search Engines. In: Proceedings of Libraries in Next Era (LiNE), India. ISBN- 978-81-930849-0-8.

Gulli, A., & Signorini, A. (2005, May). The indexable web is more than 11.5 billion pages. In Special interest tracks and posters of the 14th international conference on World Wide Web (pp. 902-903). ACM.W. Roush, “Search beyond Google,” Technology Review, 2004.