Introduction to Information Retrieval

I. Objectives

The Objective of this module is to:

- Introduce the basic concepts and history of Information Retrieval (IR).
- Familiarize the reader the essential functions of IR.
- Introduce the different models which are used in IR.
- Familiarize the students with various applications of information retrieval system in various fields.
- Introduce the concepts of pre-coordinate and post-coordinate indexing systems that are used to index the document collections.

II. Learning Outcomes

After reading this module the expected outcomes are:

- The student understands the basic concept and needs of IR
- The student will gain the basic knowledge of IR applications such as document clustering and categorization, classification of documents, system architecture, information and data visualization, and ranking of documents
- The student will gain the knowledge of how IR is useful in the development of search engines
- The student will gain the knowledge of different models of information retrieval system such as vector space model, relevance feedback model, and etc.
- The student will be able to answer the questions on development of information retrieval systems and their inventors.

III. Structure

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 information are important issues which is actually dealt by Information Retrieval. Information need of user has a very complex nature. Many models have been developed to understand the information

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. Need for IR

The need of IR came into picture due to some factors. They are -

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 should be 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, this big difference created a technical paradigm shift and necessitated 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 maintaining pre-coordinate indexing system costs an enormous human labour and also it lacks computation with mathematical or statistical approach. Unless the users know the proper index 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. Different forms of media and documents

3.1 Media of information -

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

3.2 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-

- **Monomedia Documents**: Text, Documents, Official Records etc.

- **Multimedia Documents**: Documents with different media
- **Hypertext Documents**: Documents with links (also called as non-linear document)
- **Hypermedia Document**: Multimedia + Hypertext
- **User Generated Documents**: Blogs, Comments and Tweets

4. What is Information Retrieval?

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 have clear, 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, allied services, ranking of documents, semantic linking, filtering and others. Search engines have been developed based on the concepts, principles and techniques developed by IR. Based on the different types of services, IR can be categorised as web search, personalised IR, enterprise/institutional service 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.
- IR system has opened up huge business opportunities through web environment.

5. Brief History of Information Retrieval

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 the birth 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 cards joined together, which could be wound past a viewing window enabling rapid manual scanning of 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 catalogue for a particular entry was felt in due years. Emanuel Goldberg was the first person who worked to solve that problem in 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. "Here it can be seen that catalogue entries were stored on a roll of film (figure 1). A query (2) was also on film showing a negative image of the part of the catalogue 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 screen or photographic plate (4 &5)"[1].

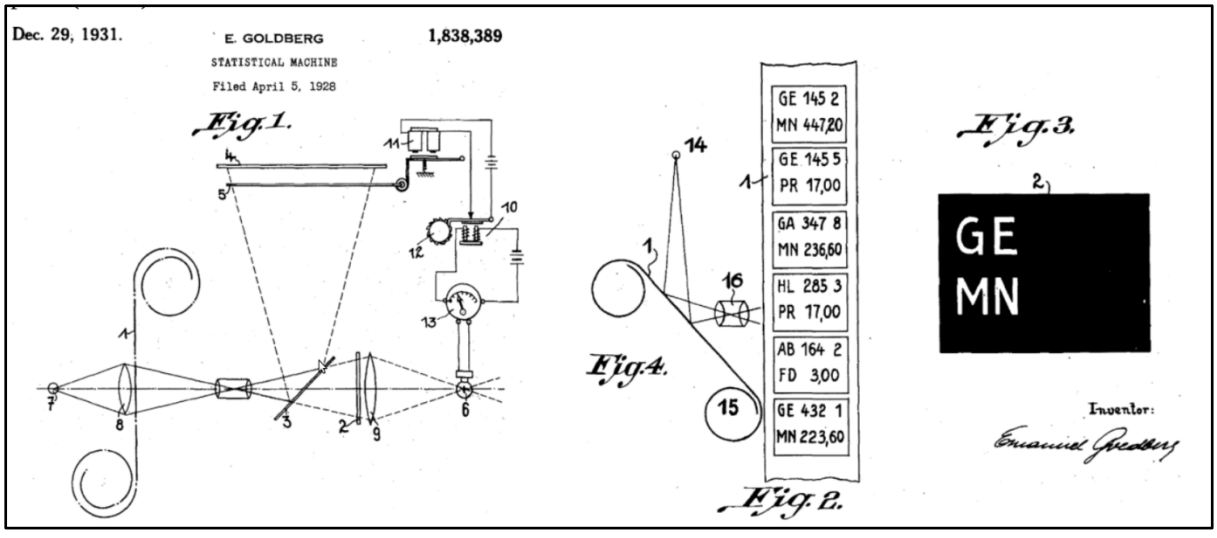

Fig.1: Goldman Optical Machine

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. Early Use of Computers

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 to search for contents. During 1950s, many projects were undertaken related to IR in different organizations (General Electric etc.).

Important Milestones

- Co-ordinate and Uniterm Indexing by Mortimer Taube(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. The Cost of Indexing

ii. The cost of preparing the physical index

iii. The cost of searching

- Cranfield—WRU test(Parallel to Cranfield-1 test)
- Cranfield-2 test (1963-1966) was executed by C.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.
- In 1962, Allan Kent published “Information Analysis and Retrieval”.
- 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.

- First online systems--NLM's AIM-TWX, MEDLINE; Lockheed's Dialog; SDC's ORBIT.

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

i. A Theory of Indexing [5]

ii. A theory of term importance in automatic text analysis [6]

iii. A vector space model for automatic indexing [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].

- In 1989, Tim Berners-Lee proposed World Wide Web in CERN Laboratory.

- 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 popularity in WWW and was hugely used in Search Engine Optimization (SEO).

- In 1997, Google Inc. was born which has now ruling dominantly in searching engine domain.

7. Summary

The present situation of web and the environment of search engine did not evolve within moments rather it's the product of decades-long research. This chapter briefly delineated importance and history of Information Retrieval.

8. References

1. Retrieved from http://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. Retrieved from http://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. A theory of indexing. 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 society for Information Science 26.1 (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.