Volume 5, Issue 12, December 2015 ISSN: 2277 128X

International Journal of Advanced Research in Computer Science and Software Engineering

Research Paper

Available online at: www.ijarcsse.com

An Overview of Information Retrieval Techniques

Prof. Dipak R. Pardhi
Assistant Professor and Head of
Department of Computer Engineering
Godavari College of Engineering Jalgaon (MH) India

Lalitkumar B. Borase
M.E. [Second Year]
Department of Computer Engineering
Godavari College of Engineering Jalgaon (MH) India

***Abstract* — This paper describes a brief study on the techniques of Information Retrieval. Here we have discussed how Information Retrieval system developed starting from the ancient time and what advancement are being done till now. Further we have focused on the role played by search engines in information retrieval and their types. We have suggested that information retrieval model must contain some features to know user's intent behind the search so that search engines must return results that may prove informative to the users.**

***Keywords*—Information Retrieval; Models; Web search; Semantic search; Indexing**

I. INTRODUCTION

An Information Retrieval system provides the information that is relevant to a user's query. An IR system performs search action on a huge structured and unstructured data like web pages, documents, images, videos, etc.

Definition: Information retrieval is the action of obtaining information resources significant to an information need from a collection of IR (Information Retrieval). Search can be based on metadata or on full-text (or other content-based) indexing [9].OR

The definition of IR (Information retrieval) according to (Manning et al., 2009) is - Information retrieval (IR) is finding material (usually documents) of an formless character (usually text) that satisfies an information need from inside large collections (usually stored on workstations).The IR was started a long before the origin of computers and internet. The first approach to manage and collect large information originated with the creation of electromechanical searching devices which were used in Libraries. In librarianship where books or papers were indexed and kept, in 1940s the first computer-based searching systems were built. Now-days IR system use automatic indexing of words in text where as in earlier days these electromechanical and computational devices used manually generated catalogs for searching. Emanuel Goldberg was the first person to build such an electromechanical device in 1920s and 1930s [1].

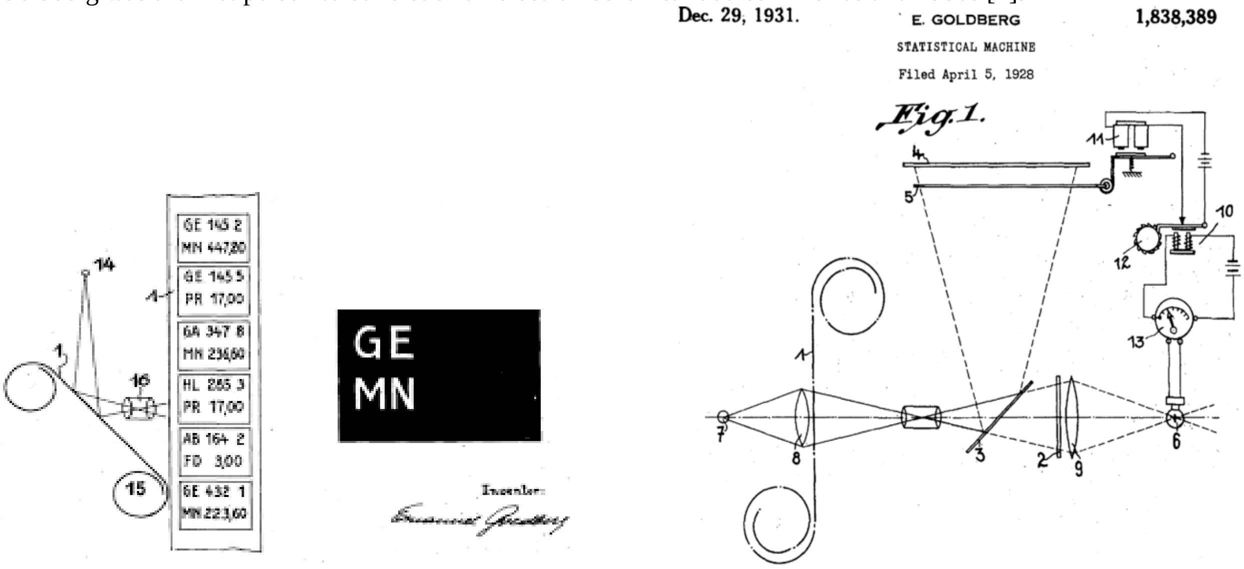

Figure1. Electromechanical Searching Device [1]

The need of IR occurs when the collection of data was not being handled by these old traditional cataloging techniques. Which can be understand by Moore's law which state that due to continual processor speed increase, there has been a constant replication in digital storage capacity every two years. Search engines play the major role in information retrieval in today's IR systems. A search engine searches for and identifies items in a database that correspond to keywords or characters specified by user. Now a lot of stress is being laid on semantic search i.e. to know the intent of user behind query being searched by the user. The two important components of IR are: how to index documents and how to retrieve them. Indexing and ranking are two building factors of IR on which its working depends.

© 2015, IJARCSSE All Rights Reserved

II. INDEXING

It is a common method for keeping track of data so that it can be accessed quickly. Like an index in a book, it is a list in which every entry contains the name of the item and its place. However, computer-based indexes may point to a physical place on a disk or to a logical location those points elsewhere to the actual place.

- *Indexer* is a program which indexes the sites, peaceably with received principles. It finds the new web sites choose the important information and transfer further.
- *Index Interface* records and load data from index. It also control recorded contents and search in index. It is implemented as database.
- *User Interface* takes the question from user and delivers it to the core of the system and receives the answer.

Techniques of Indexing:

- *Forward Index:* It is a table of founded fragments of text (tokens) together with their positions in text.
- *Inverted Index:* It depends on sorting table with tokens and groups them with positions where they are in web and pages.
- *Inverted File Index:* In this only the address of the documents are recorded in database.
- *B.Tree Index:* These are the alternate to the index sequential files and they are the tree structure with sorted data.

III. RANKING

The IR (Information retrieval) systems are electromechanical and computer system used the same style called Boolean retrieval. In an IR when the user gives a query, the most relevant document to the query is retrieved by consulting the index. Now the ranking of these relevant documents are done according to their importance, degree of relevance, etc. ranking systems depends on judgment of websites how site is similar to query.

Techniques of ranking:

- *Vector Space Model:* It is also known as term vector model. It is an algebraic model which represents documents and queries as vectors of identifiers in the word space. It used in relevancy location of indexed data, information filtering and information retrieval [3].
- *Page Rank Techniques:* It is a link analysis technique in which importance of a document is estimated by analyzing the link structure of a hyperlinked set of documents. Page Rank determine a rough estimate of how important the website is by counting the number and quality of links to a page of that website. The underlying assumption is that more essential websites are probable to receive more links from another websites. It is named after Larry Page (co-founder of Google).
- *Page Rank Methods/Techniques* in mathematical term- We assume page A has pages T1….Tn which point to it (i.e., are notations). The parameter d is a damping factor which can be place between 0 and 1. We usually set d to 0.85. Also C (A) is defined as the number of relations going out of page A (a normalizing factor). The Page Rank of a page A is a value in the range 0 to 1 and is given by:

$$PR(A) = \frac{1-d}{N} + d\left(\frac{PR(T_1)}{C(T_1)} + \cdots + \frac{PR(T_n)}{C(Tn)}\right)$$

A search engine is a program that searches for identifies item in a database that correspond to keywords or characters specified by user. Performance of a search engine is measured on the basis of how fast and correct result it gives to the user's query. User's intent (i.e. what exactly user is looking for) is a major factor which plays an important role in refining the result of search engine [7].

IV. SEARCH ENGINES

Search engine components:

- *Crawler:* It browses the document collection and fetches documents. It also traverses the web by recursively following the links from seed.
- *Indexer:* This system builds an index of the documents.
- *User interface:* It is the visible content to use. It takes the query from the user and delivers it to core of the system and receive the answer.
- *Relevance Feedback:* User may give relevance feedback to the search engine.

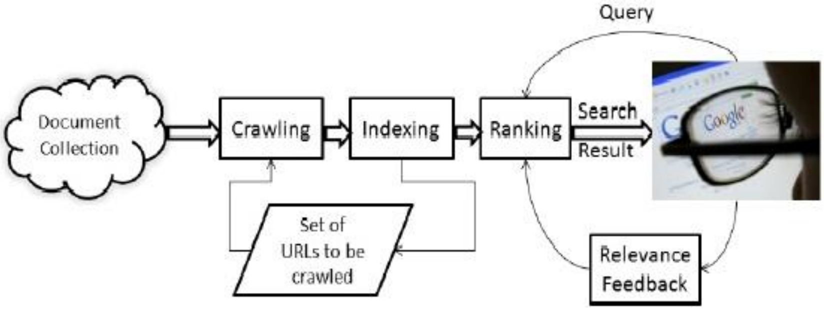

Figure2. Structure of a search engine

Types of search engines:

Mainly there are two types of search engines Desktop search engines and Web search engines.

A. Desktop Search Engines

These are the desktop search tools present within the user's personal Computer files and are designed to find information on the user's system, text documents, sound files, images, video, including web browser history and e-mail archives. There are many desktop search engines designed by various companies [2]:

- *Microsoft Windows 7 Search*: It is the default search functionality used in Microsoft Windows 7 and was considered best desktop search engine in 2009.
- *Google Desktop Search:* Google as the world leading search engine company offers a lot of products dealing with information search and visualization e.g. Google Earth. Chrome and also Google Desktop Search is a well-established in desktop search engine market.
- *Hulbee Desktop:* It is the newest product in the market and it is a Swiss made desktop search engine which offers a data cloud for serendipity search in addition to the result set.
- *X Friend personal Desktop search:* It is a desktop search engine from Convotis AG.
- *Archivarius 3000:* Archivarius is a product from Likasoft.
- *Find and Run Robot (FARR):* It is a desktop search engine that is in various points totally different from all above mentioned search engines. It has only a keyboard-controlled interface, it has no separate index but it can be used as a program launcher and as interface for internet queries.

B. Web Search Engines

A Web Search engine is a software system that is planned to search for information on the World Wide Web. It can be further classified on the basis of information retrieval methods and searching methods. There are different web search engines designed by various companies [5]:

- *Google Web Search*: It is most-used search engine on World Wide Web and owned by Google Inc. It handles more than 3 billion searches a day.
- *Yahoo Search:* It is owned by Yahoo Inc. and it was as of Jan 2014 the second largest search directory on the web by query volume. It features keyword and topic searching.
- *Exite:* It is owned by IAC Search and Media. It claims to be world most complete and flexible search engine. With a complete text index of more than 50 million web pages.
- *Bing previously known as MSN Search:* It a search engine from Microsoft.

V. TYPES OF SEARCH

A. Semantic Search

Semantic Search technique is a process used to produce searching for individuals conducting research on the Internet that is, when there is no every document you're seeking but when you're trying to locate a number of documents that will help you locate the information you want. To improve search accuracy by understanding searcher intent and the relative significance of terms as they appear in the searchable data space. Here we use phrase with query. Rather than use ranking algorithms (as several trendy Internet search engines do), to predict relevancy, Semantic Search uses the science of significance in language "semantics" to generate highly significant search outcomes [5].

Some of Semantic Search Engines:

- *Hakia:* The search queries were mapped to the results and ranked using an algorithm that scores them on sentence analysis and how strongly they match the concept associated to the query. But unfortunately it was closed down in April 2014.
- *Kosmix:* It focuses on informational search makes it suitable for topics when you want information on it rather than appear for a particular answer or URL.
- *Sensebot:* The technology powering this engine creates a outline of the top outcome that are returned for a user query, often opposing the need to drill down into the URLs to get the information that one is seeking.
- *Swoogle:* Swoogle is strictly for the semantic web. The engine indexes documents developed on the concepts and values for semantics.
- *Powerset:* The Microsoft-acquired search engine Powerset. All search outcomes on Powerset come from wikipedia, building it the eventual way to search Wikipedia, using semantics.
- *DuckDuckGo:* It is a search engine that
  Emphasis defensive searchers' privacy and avoiding the "filter bubble" of personalized search result.
- *Truevert:* It has labeled itself as a green search engine. All results are filtered and organized from one specific perception – with the topic of environmental alertness in mind. Searching for any term will be put in the context of environmental concerns.

B. Keyword Based Search

A keyword search is a type of search where only keywords are used we don't use phrase with query. Internet search engines like Google, Yahoo and Bing have popularized keyword based search i.e. these are the best examples based on keyword based search. Here keywords submitted by users to the search engine are returned in ranked list of documents as an output [5].

C. Performance Parameters

Performance parameters of various information retrieval techniques:

- *Precision:* Ratio of the number of relevant documents relevant to the number of documents that have been retrieved [4].

$$Precision = \frac{Number\ of\ relevant\ documents}{Number\ of\ retrived\ documents}$$

- *Recall:* Ratio of the number of relevant documents actually available in the repository to the number of relevant documents that have been retrieved [4].

$$Recall = \frac{Number\ of\ relevant\ documents}{Number\ of\ relevant\ \ retrived\ documents}$$

VI. CONCLUSIONS

In this paper, we explored the history of information retrieval. We reviewed the present information retrieval techniques being used and the advancement done from the time of electromechanical searching devices to modern age computer based search engines. Earlier a person seeking for some information would probably go to a local library and, using a card catalog, locate books or documents that hopefully answered that need with very small number of questions. This was because the scope of information available to them was limited and all this procedure to retrieve information was very time consuming and difficult. Now because of web search and the system that are accessible today are so easy to use, the current state of IR system is that the user can now access hundreds of terabytes of data related to web pages, images, academic papers, scanned books, programs, video clips, news, music, films, and television.

This review has shown that the route to create useful IR (Information Retrieval) systems required much innovation and thought over a long period of time. Even though search engines have made IR systems advanced. No single search engine is universally best for all searches even for the same user. Therefore if only one search engine will exist hugging whole web, which will be the center of whole existing internet this problem can be resolved. The main concern at present in the field of searching is to know the intent of user behind the query. Therefore a lot of work is to be done to improve the semantic search.

There are still many opportunities to improve search engines although they seem a simple tool compared to such visions of the future.

.

REFERENCES

[1] Mark Sanderson, W. Bruce Croft, "The History of Information Retrieval Research," Proceedings of the IEEE, Vol.100, May 13th, 2012.

[2] Brend Markscheffel, Daniela Buttner, Daniel Fischer, "Desktop Search Engines- a State of the Art Comparison," 6th International conference on internet technology and secured transactions IEEE, December 2011.

[3] Shivangi Raman, Vijay Kumar Chaurasia, Vekantesan S, "Performance Comparison of various Information Retrieval Models Used in search Engines," International Conference on Communication, Information & Computing Technology IEEE, 19-20 Oct, 2012.

[4] Mei Kobayashi, Koichi Takeda, "Information Retrieval on the Web" in ACM Computing Surveys, Vol.32, No.2, June 2000, pp. 1279-1292.

[5] Duygu Tumer, Mohammad Ahmed Shah, Yiltan Bitirim, "An Empirical Evaluation on Semantic Search Performance of Keyword-Based and Semantic Search Engines: Google, Yahoo, Msn and Hakia," Fourth International Conference on Internet Monitoring and Protection IEEE, 2009.

[6] G.Atsaros, D.Spinellis, P.Louridas, "Site Specific versus General Purpose Web Search Engines: a Comparative Evaluation," Panhellenic Conference on Informatics IEEE, 2008.

[7] Marcin Owczarek, Bartosz Sakowicz, Andrej Napieralski, "The Comparison of Modern Search Engines," TCSET IEEE, February, 2004.

[8] Amanda Spink, Bernard J.Jasen, Chris Blakely, Sherry Koshman, "Overlap Among Major Web Search Engines," Third International Conference on Information Technology: New Generations (ITNG) IEEE, 2006.

[9] Google (2014) Information Retrieval– Wikipedia [Online]. Available: http://en.wikipedia.org/wiki/Information_retrieval