ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 2, February 2014

The Information Retrieval, a Future Barrier

G.Ganesh Sriram[1], T.Pravallika[2], K.Neelima[3], Shaik Rahmathulla[4]

Assistant Professor, Department of CSE, Gayatri Vidya Parishad College of Engineering (A), Visakhapatnam, India[1]

Student, Department of CSE, Gayatri Vidya Parishad College of Engineering (A), Visakhapatnam, India[2]

Student, Department of CSE, Gayatri Vidya Parishad College of Engineering (A), Visakhapatnam, India[3]

Student, Department of CSE, Gayatri Vidya Parishad College of Engineering (A), Visakhapatnam, India[4]

ABSTRACT: This paper describes a brief history of the research and development of information retrieval system starting with electromagnetic searching device, through the construction of indexing table and inverse document matrices through to the early adaption of computers to search for items that are relevant to user query. This includes ranking of queries based on frequency weights. This also includes vector similarity between the query and the retrieved document. The address achieved by information retrieval researches from the 1950's through to the present day of locating relevant information. The paper closes with speculation on where the future of information retrieval lies

Keywords: Information Retrieval, Ranked Retrieval, Vector Similarity, Query search, Web pages.

I. INTRODUCTION

The long history of information retrieval does not begin with the Internet. Prior to the broad public day-to-day use of search engines, information retrieval (IR) systems were found in commercial and intelligence applications as long ago as the 1960s. As with many computer technologies, the capabilities of retrieval systems grew with increases in processor speed and storage capacity. The development of such systems also reflects a rapid progression away from manual library-based approaches of acquiring, indexing, and searching information to increasingly automated methods. An IR system locates information that is relevant to a user's query. The need for an IR system occurs when a collection reaches a size where traditional cataloguing techniques can no longer cope. With the growth of digitized unstructured information and, via high-speed networks, rapid global access to enormous quantities of that information, the only viable solution to finding relevant items from these large text databases was search, and IR systems became ubiquitous. This is followed by a description of how IR moved to automatic indexing of the words in text and how complex Boolean query languages gave way to simple text queries. This review finishes with a perspective on the future challenges for IR.

II. EARLY USE OF COMPUTERS FOR IR

To discuss means of dealing with a perceived explosion in the amounts of scientific information available, a specially convened conference was held by the U.K.'s Royal Society in 1948. At it, Holmstrom described a Bmachine called the Univac capable of searching for text references associated with a subject code. It appears that this is the first reference to a computer being used to search for content. He described a project to model the use of a Univac computer to search 1,000, 000 records indexed by up to six subject codes; it was estimated that it would take 15 h to search that many records. The impact of computers in IR is highlighted when Hollywood drew public attention to the innovation with the comedy Desk Set. IR as a research discipline was starting to emerge at this time with two important developments: how to index documents and how to retrieve them.

A. Indexing, The Move Toward Words :

In the field of librarianship, the way that items were organized in a collection was a topic that as regularly debated. The classic approach was to use a hierarchical subject classification scheme, such as the Dewey Decimal system, which assigned numerical codes to collection items, which was essentially a proposal to index items by a list of keywords. His

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 2, February 2014

conclusion that Uniterms were as good as and possibly better than other approaches caused much surprise and his work came under extensive scrutiny. However, Cleverdon's experimental results were found to be correct and as a result the use of words to index the documents of an IR system became established. Many aspects of Cleverdon's test collection approach to evaluation are still used in both academic research and commercial search testing today.

B. Ranked Retrieval :

The style of search used by both the electromechanical and computer-based IR systems was so-called Boolean retrieval. A query was a logical combination of terms which resulted in a set of those documents that exactly matched the query. Luhn proposed and Maron tested an alternative
approach, where each document in the collection was assigned a score indicating its relevance to a given query. The documents were then sorted and those at the top ranks were returned to the user. The researchers manually assigned keywords to a collection of 200 documents, weighting those assignments based on the importance of the keyword to the document. The scores assigned to the documents were based on a probabilistic approach. The researchers hand tested their ranked retrieval method, showing that it outperformed Boolean search on this test collection with 39 queries. In the same year as Maron's work, Luhn suggested that the frequency of word occurrence in an article furnishes a useful measurement of word significance his approach later became known as term frequency weighting. This ranked retrieval approach to search was taken up by IR researchers, who over the following decades refined and revised the means by which documents were sorted in relation to a query. The superior effectiveness of this approach over Boolean search was demonstrated in many experiments over those years for a list of these experiments. What followed were the growth of a commercial search sector and the consolidation of IR as an increasingly important research area.

C. Vector similarity :

Once the list of documents is formed, the search engine computes a semantic similarity value between the query and each document, as follows and D be the set of all documents in the search space. Let q to Q be an RDQL query, let Vq be the set of variables in the SELECT clause of q, let w be the weight vector for these variables, where for each v to Vq. We represent each document in the search space as a document vector d to D, where dx is the weight of the annotation of the document with concept x for each x to O, if such annotation exists, and zero otherwise. Note that the sum rarely has more than one term since this would mean that the same instance appears as a satisfying value for different variables in different (or the same) result set tuples.

$$\mathrm{sim}(d,q) = \frac{d \bullet q}{|d| \bullet |q|}.$$

THE MID-1990s TO THE PRESENT

Although Berners-Lee created the World Wide Web in late 1990, the number of websites and quantity of pages was relatively small until 1993. In those initial years, conventional manual cataloging of content sufficed. In the middle of 1993, as recorded by Gray's survey,2 there were around 100 websites; six months later, there were over four times that number, and six months after that, the number had increased fourfold again. Web search engines started to emerge in late 1993 to cope with this growth. The arrival of the web initiated the study of new problems in IR. This point also marked a time when the interaction between the commercial and research-oriented IR communities was much stronger than it had been before. Ideas developed in earlier years were pushed further and implemented in the commercial search sector.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 2, February 2014

A. **Web Search :**

During early development of the web, there was a list of web servers edited by Tim Berners-Lee and hosted on the CERN webserver. One historical snapshot of the list in 1992 remains, but as more and more webserver's went online the central list could no longer keep up. The very first tool used for searching on the Internet was Archie. The name stands for "archive" without the "v". It was created in 1990 by Alan Emtage, Bill Heelan and J. Peter Deutsch, computer science students at McGill University in Montreal. When a user enters a query into a search engine, the engine examines its index and provides a listing of best-matching web pages according to its criteria, usually with a short summary containing the document's title and sometimes parts of the text. Most search engines support the use of the boolean operators AND, OR and NOT to further specify the search query. Boolean operators are for literal searches that allow the user to refine and extend the terms of the search. The engine looks for the words or phrases exactly as entered. Some search engines provide an advanced feature called proximity search, which allows users to define the distance between keywords. There is also concept-based searching where the research involves using statistical analysis on pages containing the words or phrases you search for.

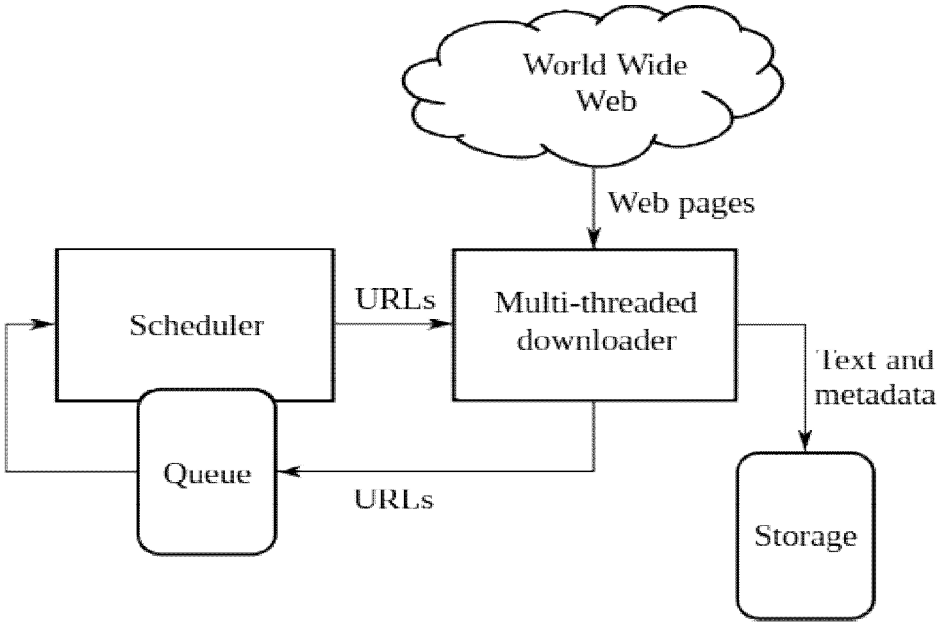

Fig 1: High-level architecture of a standard Web Crawler

B. **Exploiting Query Logs :**

Since the manually labeled training data for query classification is expensive, how to use a very large web search engine query log as a source of unlabeled data to aid in automatic query classification becomes a hot issue. These logs record the Web users' behavior when they search for information via a search engine. Over the years, query logs have become a rich resource which contains Web users' knowledge about the World Wide Web.

- Query clustering method tries to associate related queries by clustering "session data", which contain multiple queries and click-through information from a single user interaction. They take into account terms from result documents that a set of queries has in common. The use of query keywords together with session data is shown to be the most effective method of performing query clustering.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 2, February 2014

- Selectional preference based method tries to exploit some association rules between the query terms to help with the query classification. Given the training data, they exploit several classification approaches including exact-match using labeled data, N-Gram match using labeled data and classifiers based on perception.

C. **New Areas of Search :**

The applications of search and the field of IR continue to evolve as the computing environment changes. The most obvious recent example of this type of change is the rapid growth of mobile devices and social media. One response from the IR community has been the development of social search, which deals with search involving communities of users and informal information exchange. New research in a variety of areas such as user tagging, conversation retrieval, filtering and recommendation, and collaborative search is starting to provide effective new tools for managing personal and social information. An important early paper in this area dealt with desktop search which has many similar characteristics to current search applications in the mobile world.

III. CONCLUSION AND FUTURE DIRECTIONS

It is time to re-examine the evaluative methods that are used in information retrieval studies, particularly with respect to search performance on the Web. New measures can be defined that take into account hyper linking and relevance ranking. The measures presented in this paper are only a few of many such measures that could be defined. Ultimately, it would be useful to derive a set of evaluative dimensions that describe search engine performance, possibly based on a principle components or factor analysis of a large set of basic evaluative measures. More extensive studies are needed to assess the properties of different search engines and evaluative measures, particularly in realistic search tasks. Studies have shown that some measures differentiate against current search engines whereas others do not. Researchers need to determine whether these lacks of differences stem from the insensitivity of the measures or a genuine lack of difference between the search engines on the attribute being measured.

REFERENCES

1. C. Walter, Insights: Kryder's Law. Singapore: Scientific American, 2005.
2. S. Eliot and J. Rose, A Companion to the History of the Book. New York: Wiley, 2009.
3. H. E. Soper, BMeans for compiling tabular and statistical data,[U.S. Patent US00 135 169 231-1920, 1920.
4. E. Goldberg, BStatistical machine, [U.S. Patent 183 838 929-1931, 1931.
5. M. K. Buckland, Emanuel Goldberg and His Knowledge Machine: Information, Invention and Political Forces. Westport, CT: Greenwood, 2006.
6. C. N. Mooers, BThe next twenty years in information retrieval: Some goals and predictions,[Proc. Western Joint Comput. Conf., 1959, pp. 81–86.
7. K. Sparck Jones, Ed., Information Retrieval Experiment. Oxford, U.K.: Butterworth-Heinemann, 1981.
8. G. Salton, Automatic Information Organization and Retrieval. New York: McGraw-Hill, 1968.
9. J. J. Rocchio, BRelevance feedback in information retrieval,[Harvard Univ., Cambridge, MA, ISR-9, 1965.
10. M. E. Stevens, V. E. Giuliano, and L. B. Heilprin, Proc. Symp. Stat. Assoc. Methods Mechanized Documentation, Washington, DC, 1964.
11. N. Fuhr, BOptimum polynomial retrieval functions based on the probability ranking principle,[ACM Trans. Inf. Syst., vol. 7, no. 3, pp. 183–204, 1989.
12. N. Fuhr and C. Buckley, BA probabilistic learning approach for document indexing, ACM Trans. Inf. Syst., vol. 9, no. 3, pp. 223–248, 1991.

BIOGRAPHY

Mr.G.Ganesh Sriram, a well known author and an excellent teacher, received M.Tech in Software Engineering from JNT University Kakinada. He has been working as an Assistant professor in the Department of Computer Science and Engineering, Gayatri Vidya parishad College of Engineering(A), Madhurawada, Visakhapatnam. He has got many publications in reputed International Journals. He has been guiding many projects for the students of UG. His areas of interest includes C programming, Data Structures, Software Engineering, Cloud Computing and Other emerging areas in the field of Computer Science.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 2, February 2014

Pravallika Tholeti currently pursuing her B.Tech degree in Computer Science Engineering from Gayatri Vidya Parishad College of Engineering (A), Visakhapatnam. Her research interests are focused on confluence of Information Retrieval, Artificial Intelligence and Networking.

Neelima Kokkiligadda currently pursuing her B.Tech degree in Computer Science Engineering from Gayatri Vidya Parishad College of Engineering (A), Visakhapatnam. Her research interests are on Information Retrieval and Image Processing.

Shaik Rahmathulla currently pursuing his B.Tech degree in Computer Science Engineering from Gayatri Vidya Parishad College of Engineering (A), Visakhapatnam. His research interests are on Information Retrieval and Networking.