ISSN No. 0976-5697

Volume 3, No. 5, Sept-Oct 2012

International Journal of Advanced Research in Computer Science

REVIEW ARTICAL

Available Online at www.ijarcs.info

Information Retrieval: An Overview

Kuyoro Shade O.* and Awodele Oludele
Department of Computer Science
Babcock University
Ilishan-Remo, Nigeria
afolashadeng@gmail.com
delealways@yahoo.com

Ibikunle Frank A.
Department of Computer Science
Covenant University
Otta, Nigeria
faibikunle2@yahoo.co.uk

Abel Samuel B.
Department of Computer Science
Babcock University
Ilishan-Remo, Nigeria
abelsammie@yahoo.co.uk

Abstract: Information retrieval (IR) is the field of computer science that deals with the processing of documents containing free text, so that they can be rapidly retrieved based on keywords specified in a user's query. IR was born in the 1950s out of necessity to find useful information from large collections. Over the last sixty years, the field has matured considerably. IR technology is the basis of Web-based search engines, and plays a vital role in research, because it is the foundation of software that supports literature search. Several IR systems are used on an everyday basis by a wide variety of users. This article is a brief overview of Information Retrieval.

Keywords: information retrieval, Web-based search engine, universal repositories, filtering and precision

I. INTRODUCTION

Information retrieval (IR) deals with the representation, storage, organization of, and access to information items. In the past 20 years, the area of information retrieval has grown well beyond its primary goals of indexing text and searching for useful documents in a collection. Nowadays, research in IR includes modeling, document classification and categorization, systems architecture, user interfaces, data visualization, filtering, languages, etc. Despite its maturity, until recently, IR was seen as a narrow area of interest mainly to librarians and information experts. Such a tendentious vision prevailed for many years, despite the rapid dissemination, among users of modern personal computers, IR tools for multimedia and hypertext applications. In the beginning of the 1990, a single fact changed once and for all these perceptions - the introduction of the World Wide Web.[1]

The Web is becoming a universal repository of human knowledge and culture which has allowed unprecedented sharing of ideas and information in a scale never seen before. Its success is based on the conception of a standard user interface which is always the same no matter what computational environment is used to run the interface. As a result, the user is shielded from details of communication protocols, machine location, and operating systems. Further, any user can create his Web documents and make them point to any other Web documents without restrictions. This is a key aspect because it turns the Web into a new publishing medium accessible to everybody. As an immediate consequence, any Web user can push his personal agenda with little effort and almost at no cost. This universe without frontiers has attracted tremendous attention from millions of people everywhere since the very beginning. [1][2]

Despite so much success, the Web has introduced new problems of its own. Finding useful information on the Web is frequently a tedious and difficult task. For instance, to satisfy his information need, the user might navigate the space of Web links (i.e., the hyperspace) searching for information of interest. However, since the hyperspace is vast and almost unknown, such a navigation task is usually inefficient. For naive users, the problem becomes harder, which might entirely frustrate all their efforts.[3] The main obstacle is the absence of a well defined underlying data model for the Web, which implies that information definition and structure is frequently of low quality. These difficulties have attracted renewed interest in IR and its techniques as promising solutions. As a result, almost overnight, IR has gained a place with other technologies at the center of the stage.[4]

This paper is organized as follows. In section 2.0, we present a brief history of information retrieval information retrieval. Section 3.0 presents the basic IR models. Section 4.0 describes web information retrieval and systems. Section 5 presents the differences between the classic information retrieval and web information retrieval. While section 6.0 looked critically into information retrieval evaluation, and section 7.0 gives the conclusion with some future directions.

II. BRIEF HISTORY

The practice of archiving written information can be traced back to around 3000 BC, when the Sumerians designated special areas to store clay tablets with cuneiform inscriptions. Even then the Sumerians realized that proper organization and access to the archives was critical for efficient use of information. They developed special classifications to identify every tablet and its content. The need to store and retrieve written information became

increasingly important over centuries, especially with inventions like paper and the printing press. Soon after computers were invented, people realized that they could be used for storing and mechanically retrieving large amounts of information. In 1945, Vannevar Bush published a ground breaking article titled "As We May Think" that gave birth to the idea of automatic access to large amounts of stored knowledge. In the 1950s, this idea materialized into more concrete descriptions of how archives of text could be searched automatically. Several works emerged in the mid-1950s that elaborated upon the basic idea of searching text with a computer. One of the most influential methods was described by H.P. Luhn in 1957, in which (put simply) he proposed using words as indexing units for documents and measuring word overlap as a criterion for retrieval. [1][2]

Several key developments in the field happened in the 1960s. Most notable were the development of the SMART system by Gerard Salton and his students, first at Harvard University and later at Cornell University and the Cranfield evaluations done by Cyril Cleverdon and his group at the College of Aeronautics in Cranfield. The Cranfield tests developed an evaluation methodology for retrieval systems that is still in use by IR systems today. The SMART system, on the other hand, allowed researchers to experiment with ideas to improve search quality. A system for experimentation coupled with good evaluation methodology allowed rapid progress in the field, and paved way for many critical developments. The 1970s and 1980s saw many developments built on the advances of the 1960s. Various models for doing document retrieval were developed and advances were made along all dimensions of the retrieval process. These new models/techniques were experimentally proven to be effective on small text collections available to researchers at the time. However, due to lack of availability of large text collections, the question whether these models and techniques would scale to larger corpora remained unanswered. This changed in 1992 with the inception of Text Retrieval Conference (TREC). TREC is a series of evaluation conferences sponsored by various US Government agencies under the auspices of NIST, which aims at encouraging research in IR from large text collections.[1][2]

With large text collections available under TREC, many old techniques were modified, and many new techniques were developed (and are still being developed) to do effective retrieval over large collections. TREC has also branched IR into related but important fields like retrieval of spoken information, non-English language retrieval, information filtering, user interactions with a retrieval system, and so on. The algorithms developed in IR were the first ones to be employed for searching the World Wide Web from 1996 to 1998.[5]

III. INFORMATION RETRIEVAL MODELS

In order to effectively retrieve information, a number of models have been developed but the three classic models in information retrieval are called Boolean, vector, and probabilistic. In the Boolean model, documents and queries are represented as sets of index terms. Thus, the model is said to be set theoretic. In the vector model, documents and queries are represented as vectors in a t-dimensional space (algebraic model). In the probabilistic model, the framework for modeling document and query representations is based on probability theory (probabilistic model).[6]

A. Boolean Model:

The Boolean model of information retrieval is a classical information retrieval (IR) model and, at the same time, the first and most adopted one. It was proposed about 1950, and is used by virtually all commercial information retrieval systems today. The Boolean information retrieval is based on Boolean Logic and classical Sets Theory in that both the documents to be searched and the user's query are conceived as sets of terms. Retrieval is based on whether or not the documents contain the query terms. [7][8]

Advantages of the Boolean model include: (1) It is easy to implement and it is computationally efficient. Hence, it is the standard model for the current large-scale, operational retrieval systems and many of the major on-line information services use it. (2) It enables users to express structural and conceptual constraints to describe important linguistic features. Users find that synonym specifications (reflected by OR-clauses) and phrases (represented by proximity relations) are useful in the formulation of queries. (3) The Boolean approach possesses a great expressive power and clarity. Boolean retrieval is very effective if a query requires an exhaustive and unambiguous selection. (4) The Boolean method offers a multitude of techniques to broaden or narrow a query. (5) The Boolean approach can be especially effective in the later stages of the search process, because of the clarity and exactness with which relationships between concepts can be represented.[3][9] The disadvantages of Boolean model include the fact that users find it difficult to construct effective Boolean queries for several reasons. Users are using the natural language terms AND, OR or NOT that have a different meaning when used in a query. Thus, they will make errors when they form a Boolean query, because they resort to their knowledge of English.[7][10]

B. Vector Space Model:

The Vector space model, proposed in 1970, is a statistical retrieval model that represents the documents and queries as vectors in a multidimensional space, whose dimensions are the terms used to build an index to represent the documents. The creation of an index involves lexical scanning to identify the significant terms, where morphological analysis reduces different word forms to common stems, and the occurrence of those stems is computed. Query and document surrogates are compared by comparing their vectors, using, for instance, the cosine similarity measure. [2][11]

In this model, the terms of a query surrogate can be weighted to take into account their importance, and they are computed by using the statistical distributions of the terms in the collection and in the documents. The vector space model can assign a high ranking score to a document that contains only a few of the query terms if these terms occur infrequently in the collection but frequently in the document. The vector space model makes the following assumptions:

a. The more similar a document vector is to a query vector, the more likely it is that the document is relevant to that query, and
b. The words used to define the dimensions of the space are orthogonal or independent. While it is a reasonable first approximation, the assumption that words are pairwise independent is not realistic. [1][2][8][11]

The vector space model, like all statistical retrieval models has the following advantages:

c. It provides users with a relevance ranking of the retrieved documents. Hence, they enable users to control the output by setting a relevance threshold or by specifying a certain number of documents to display.
d. Queries can be easier to formulate because users do not have to learn a query language and can use natural language.
e. The uncertainty inherent in the choice of query concepts can be represented.[11]

Despite its simplicity, the vector space model is a resilient ranking strategy with general collections. It yields ranked answer sets which are difficult to improve upon without query expansion or relevance feedback within the framework of the vector model. A large variety of alternative ranking methods have been compared to the vector space model but the consensus seems to be that, in general, the vector space model is either superior or almost as good as the known alternatives. Since it is simple and fast, the vector space model is a popular retrieval model nowadays.[2][11]

C. Probabilistic Model:

The classic probabilistic model, which later became known as the binary independence retrieval (BIR) model was introduced in 1976 by Roberston and Sparck Jones. The probabilistic model attempts to capture the IR problem within a probabilistic framework based on the Probability Ranking Principle, which states that, 'an information retrieval system is supposed to rank the documents based on their probability of relevance to the query, given all the evidence available'. The principle takes into account that there is uncertainty in the representation of the information need and the documents. [11]

There can be a variety of sources of evidence that are used by the probabilistic retrieval model, and the most common one is the statistical distribution of the terms in both the relevant and non-relevant documents. Probabilistic retrieval model has the same general characteristic advantages and or shortcomings as the other statistical retrieval models.[1][2][11]

IV. WEB INFORMATION RETRIEVAL AND SYSTEMS

Retrieving information from the web is becoming a common practice for internet users. The huge size and heterogeneity of the web is no longer strange. Therefore, the web poses a dire challenge to the effectiveness of classical information retrieval systems. A critical goal of successful information retrieval on the web is to identify which pages are of high quality and relevance to a user's query. The success of the web lies in the many software tools that are available for its information retrieval. The main tools used by the web in its information retrieval include:

a. General-purpose search engines. This can be Direct (e.g. AltaVista, Excite, Google, Infoseek and Lycos) or Indirect or Meta-search (e.g. MetaCrawler, DogPile, AskJeeves, and Invisible Web).
b. Hierarchical directories. This can be manual, that is, the database is mostly built by hand or automatic. Examples of manual hierarchical directories are Yahoo, LookSmart and Open Directory. Automatic hierarchical directories are now populating hierarchy. For each node in the hierarchy, fine-tuned query are formulated and run modified HITS algorithm. The techniques used in automatic hierarchical directories are connectivity and/or text based. Another feature of automatic hierarchical directories is Categorization, here for each document the best placement is found in the hierarchy.
c. Specialized search engines. These deals with heterogeneous data sources include the Home page finder such as Ahoy, the Shopping robots such as Jango and Junglee, whose database is mostly built by hand, and Applet finders.
d. Search-by-example. Examples are Alexa's "What's related", Excite's "More like this", Google's "Googlescout", etc.
e. Collaborative filtering. Examples include Firefly and GAB and
f. Meta-information. These are Search Engine Comparisons and are used for Query log statistics. [3][6][9][12]

With the fast growth of the Internet, more and more information is available on the web and as a result, web information retrieval has become a fact of life for most Internet users. The following are some of uniqueness of web information retrieval:

a. Bulk. The bulk size of the Internet is over 400 million documents as measured in the year 2000, which is growing at the speed of 20M per month.
b. Dynamic Internet. The Internet is changing everyday while most classic information retrieval systems are designed for mostly static text databases.
c. Heterogeneity. The Internet contains a wide variety of document types: pictures, audio files, text and scripts etc.
d. Variety of Languages. The type of languages used in the Internet is more than 100.
e. Duplication. Copying is another important characteristic of the web, as it is estimated that about 30% of the web pages are duplicates.
f. High Linkage: Each document averagely has more than 8 links to other pages.
g. Ill-formed queries. Web information retrieval systems are required to service short and not particularly well represented queries from the Internet users.
h. Wide Variance in Users: Each web user varies widely in their needs, expectations and knowledge.
i. Specific Behavior. It is estimated that nearly 85% users only look at the first screen of the returned results from search engines. 78% users never modify their very first query.[6][9[11]

V. CLASSIC INFORMATION VERSUS WEB INFORMATION RETRIEVAL

Classic information retrieval constitutes all previous information retrieval techniques before and other than the web information retrieval. The input of classic information retrieval is mainly for document collection and the goal is to retrieve document or text with information content that is relevant to user's information need. Classic information retrieval involves two main aspects: (1) Processing the document collection and (2) processing queries (searching). [1][2

To determine the query results, that is, which documents to return, information retrieval models like the Boolean and Vector models are used. On the other hand, the input of web information retrieval is the publicly accessible web while the goal is to retrieve high quality pages that are relevant to user's information need. Web information retrieval can be static, in which files like text, audio and videos are retrieved, or dynamic, which is mainly database access generated on request. Two aspects of the web information retrieval are processing and representation of the document collection and processing queries. Processing and representation of document collection involves either gathering the static pages or learning about the dynamic pages. [1][2]

Web information retrieval has the following advantages over classic information retrieval: 1. User (a) Many tools are available to the user; (b) Personalization of information result given a query is better and (c) Interactivity: for instance the query can be refined or expanded as desired. 2. Collection/System (a) Hyperlinks are available to link one document to the other; (b) Statistics is easy to gather even in large sample sizes and (c) Interactivity: the system makes the users explain what they want.[5][6][11]

VI. INFORMATION RETRIEVAL EVALUATION

Much effort and research has gone into solving the problem of evaluation of information retrieval. However, it is probably fair to say that most people active in the field of information storage and retrieval still feel that the problem is far from solved. Progress in the field critically depends upon experimenting with new ideas and evaluating the effects of these ideas, especially given the experimental nature of the field. The two desired properties that have been accepted by the research community for measurement of search effectiveness are recall -the proportion of relevant documents retrieved by the system; and precision -the proportion of retrieved documents that are relevant. It is well accepted that a good IR system should retrieve as many relevant documents as possible (i.e., have a high recall), and it should retrieve very few non-relevant documents (i.e., have high precision). Unfortunately, these two goals have proven to be quite contradictory over the years. Techniques that tend to improve recall tend to hurt precision and vice-versa.[1][2]

In classic information retrieval, the performance of an Information Retrieval system can be evaluated by assessing recall and precision, but in web information retrieval, the quality of pages varies widely such that document relevance is not enough. The goal is to return both high-relevance and high-quality, that is, valuable pages. Different users may differ about the relevance or non-relevance of particular documents to given queries. Therefore, document relevance is a subjective notion. Several experiments and researches have been done to assess relevance. And it is a general assumption in the field of Information Retrieval that should a retrieval strategy fare well under a large number of experimental conditions then it is likely to perform well in an operational situation where relevance is not known in advance.[13][14] A document is relevant to an information need if and only if it contains at least one sentence which is relevant to that need. This is the true evaluation of the effectiveness of a document, since effectiveness is purely a measure of the ability of the system to satisfy the user in terms of the relevance of documents retrieved.[1][2]

VII. CONCLUSION

The field of information retrieval has come a long way in the last sixty years and has enabled easier and faster information discovery. In the early years, there were many doubts raised regarding the simple statistical techniques used in the field. However, for the task of finding information, these statistical techniques have indeed proven to be the most effective ones so far. Techniques developed in the field have been used in many other areas and have yielded many new technologies which are used by people on an everyday basis (e.g., web search engines, junk-email filters, news clipping services). Going forward, the field is attacking many critical problems that users face in today's information-ridden world. With exponential growth in the amount of information available, information retrieval will play an increasingly important role in the future.

VIII. REFERENCES

[1] G. Chowdhury Introduction To Modern Information Retrieval, 3rd Ed Facet Publishing, 2010

[2] R. Baeza-Yates, B. Ribeiro-Neto, "Modern information retrieval, ACM Press, New York, USA, 1999.

[3] Alis J. Technologies, 'Web languages hit parade. http://babel.alis.com:8080/palmares.html, 1997.

[4] K. Andrews, "Visualizing cyberspace: information visualization in the harmony internet browser," in Proceedings '95 information visualization, Atlanta, USA, October 1995, pp. 97-104.

[5] K. S. Jones, P. Willett, "Readings in information retrieval, Morgan Kaufmann., 1997.

[6] G. Kowalski, M.T. Maybury, Information storage and retrieval systems, Springer, 2005.

[7] P. Aniek, J. Brennan, R. Flynn, D. Hanssen, B. Alvey, and J. Robbins, "A direct manipulation interface for boolean information retrieval via natural language query," in Proc. of the 19th Annual International ACM/SIGIR Conference, Brussels, Belgium, 1990, pp. 135-150.

[8] A. Apostolico and Z. Galil, "Combinatorial algorithms on words. Springer-Verlag, New York, 1985.

[9] www.wikipedia.org "Information Retrieval" Retrieved 22-06-2011

[10] P. Aniek, "Adapting a full-text information retrieval system to the computer troubleshooting domain," in Proc. of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 1994, pp. 349-358,.

[11] S. Amit, "Modern information retrieval: A brief overview, 2001.

[12] ANSI/NISO Standards, Z39.50, "Information retrieval: Application service definition and protocol specification, 1995.

[13] C.T. Meadow, B.R. Boyce, D.H. Kraft, C.L. Barry. Text information retrieval systems. Academic Press, 2007.

[14] C. J. Van Risjbergen, "The geometry of information retrieval, Cambridge UP, 2004.

[15] M. Levene, Search Engines: Information Retrieval in Practice, The Computer Journal , 2011

[16] V. Sutton, Innovations in Information Retrieval: Perspectives for Theory and Practice Library Management, 2012

