IJCSN - International Journal of Computer Science and Network, Volume 3, Issue 6, December 2014
ISSN (Online): 2277-5420 **www.IJCSN.org**
Impact Factor: 0.274

A Comparison on Intelligent Web Information Retrieval Systems

[1] Anupama Prasanth, [2] Dr. M. Hemalatha

[1] Research Scholar, Karpagam University
Coimbatore, Tamilnadu, India

[2] Professor, Department of Computer Science, Karpagam University
Coimbatore, Tamilnadu, India

Abstract - The key technology for accessing relevant data from large volume is Information retrieval. Information retrieval technology gives assurance to access large data. The major challenge of information retrieval is to find and manage all existing information in the web. So it became the elementary skill behind web search tool. Knowing the relevant information at the time of requirement is important for people. They considered information as one of the most valuable and strategic goods. But the availability of information nowadays increases tremendously, so this cause information oversupplies and results in time-consumption and difficulty in accessing relevant. Aimed to overcome these difficulties in the beginning itself several automated tools are used for searching information relevant to the user needs.

The responsibility of IR is to collect and represent the information and allows retrieving the relevant information to exact problems at real time through wired or wireless devices. The intention to find as much possible as additional background information will help an information retrieval system to improve the retrieval accuracy. This scenario requires new advanced tools, which covering in a better way the various phases of the information streams and capable of surviving with the severe limitations of existing tools for information retrieval on the web. So the main intention of this research is to finding out the techniques which can improve the effectiveness of information retrieval.

Keywords - *Information Retrieval, Feedback Mechanism, term Frequency, Inward Link.*

1. Introduction

The World Wide Web has become an important communication media of business and daily life. In terms of content and usage they became more and more dynamic. Knowing the relevant information at the time of requirement is the important for people. They considered information as one of the most valuable and strategic goods. But the availability of information nowadays increases tremendously, so this cause information oversupplies [3] and results in time-consumption and difficulty in accessing relevant. Aimed to overcome these difficulties in the beginning itself several automated tools are used for searching information relevant to the user needs. The popular tools include Search engines, Meta Search engines and Directories, even though they show poor performance. Information retrieval technology gives assurance to access large data. The major challenge of information retrieval is to find and manage all existing information in the web. So it became the elementary skill behind web search tool. The responsibility of IR is to collect and represent the information and allows retrieving the relevant information to exact problems at real time through wired or wireless devices.

Collection and representation of IR is fully based on user queries, the query is normally composed of words from natural language, and to respond to this with limited clue is a very exciting mission. The intention to find as much possible as additional background information will help an information retrieval system to improve the retrieval accuracy. [7].The furthermost related information's are offered to the users, and they can assess the relevance with respect to their problems. This scenario requires new advanced tools, which covering in a better way the various phases of the information streams and capable of surviving with the severe limitations of existing tools for information retrieval on the web.

2. Information Retrieval

The web has become full-fledged with diverse information resources like personal home pages, digital libraries, bibliographies, e-commerce sites and product and service features, research publications, FTP, Usenet news and mail servers [16] and became positioned as the largest distributed information space. Several related studies put

forward that the contents of web doubles every four months [1].

The full potential of the web can be realized only through effective search and retrieval technologies. At present there are so many searching tools are available they retrieve too many documents according to the user query, but from those the relevant document for the users is very little. Moreover it is not necessarily that the relevant documents should appear at the top of the result page.

This is a research on finding out the techniques which can improve the effectiveness of information retrieval. So this research started with the study on navigational strategies for searching the web, evaluation methods for presenting the stuffing of web contents and various models for retrieving it. In order to conduct a comparative analysis we need several measurable parameters. The first step of the study is to find out various parameters and use the result for comparative analysis. After examining these find out the factors which exhibit an efficient performance in information retrieval.

2.1 Traversing the Web

Relevant documents from the web can be retrieved using Web robot. Web robot is a software program which accept user query, locate related documents and rank them according to the query relevance and return those ranked list. But because of the hugeness of documents in the web makes this approach impossible for every user query.
We can utilize the web robot in another practical way, in that create a searchable index of web documents and do the search in that index using web robot. Most of the search tool adopted a practical approach of updating the index periodically. So the index have to construct in an efficient way. The structure of the web is like directed graph. So graph traversal algorithm can be applied, also the client server communication paradigm enables the robot to start from a single computer to traverse the entire web.

The effectiveness of indexing system can be explained by two main parameters: *Recall* is the ratio of the number of relevant documents retrieved to the total number of relevant documents in the collection and *Precision* is the ratio of the number of relevant documents retrieved to the total number of documents retrieved [13]. If at all possible try to maintain high recall and high precision.

2.2 Search Tools and Services

Search tools and search services are the two major sectors of web where automated methods of information retrieval are necessary. Robots are installed for indexing of documents in search tools. The major search tools, search engine, they search in the index database to retrieve web pages relevant to the user query. Search services simplify the web search by hiding the search tools and database from users. The spider in AltaVista index the web documents based on availability of full text of the document. They update the index atleast once a day. They revisit a page according the frequency of its updation, and can sustain Boolean, phrase and case sensitive queries. Their relevance calculation is based on whether the query term appears in the first few lines of the document.

Another search tool Excite, which uses spider program for indexing full text documents. But that spider program searches only web and Usenet groups for updating the index. Using Excite we can search exact query word, or in combination of Boolean operators AND, OR, and NOT. Results are ordered according to the rank and also it provides a “similar” query.

HotBot uses Slurp for indexing web documents. Slurp update the index based on HTML data and meta-text documents. HotBot has distributed index database across several computers so it enables parallel searching process. They support search term, phrase, proper noun, or URL also supports case-sensitive and Boolean searches. Their relevance calculation is based on various factors such as frequency and document length. If the query term appears in the title or META tag has high relevance than others.

InfoSeek Guide has robot for retrieving HTML and PDF documents. They search as usual all web documents as well as especially in Web FAQs also. Its main advantage is that it support searches for symbols, phrases also it searches images based on the caption. The relevance calculation in this method is also based whether the query term appears in the beginning of the documents. Lycos the search tool that can able to retrieve documents which match to the query by even some number of terms. The user can select their option like loose, fair, close, good and strong matches, according to their requirement. In this method the relevance is calculated based on weights of matched terms in the document. If the query term shows in the title or beginning of the document has high relevance.

The robot in WebCrawler has a list of web servers and URLs, and it retrieve documents from that list. It uses the round-robin methodology to avoid repeated fetching of documents from the same server. The relevance is calculated by their regularity in the documents. Heavy weightage is given to the terms which are regular in documents and irregular in reference list. Yahoo is another tool which can be used for both browsing and searching. It also uses robots for collecting new links. It allows Boolean

operators and also phrase searching. The relevance is given based on the frequency of appearance of query term in the document.

Google also has the robot to search in the web but the relevance calculation for indexing is different, here it gives high rank to the pages that are mostly linked to. It indicates the importance of a page. Each link to a page is considered its vote, so the page which have highest vote of support will get highest rank or top most relevance. Bing also have almost similar relevance calculation, it takes all documents from the web and parses each document for word frequency. They do stemming and parsing the generate hash value for each word, that is stored in frequency table. The same processes do for each query and map with that in frequency table.

There are so many sites they support search services, they transmit user queries to several search engines and retrieve relevant documents from all those simultaneously , if any duplicated data remove that and present the information in result page. I*BM* InfoMarket and MetaCrawler are examples of search services.

3. Retrieval Effectiveness Assessment

The retrieval effectiveness of information retrieval systems are measured using two major parameters, precision and recall. In order to measure the relevance based on these parameters required a predetermined quantity of documents, a normal set of queries, and relevant and irrelevant documents for each query. But to carry out such an experiment is quite difficult task. So here the experiment to compare the effectiveness is carried out in terms of standard queries and the number of documents retrieved for that.

The result of one experiment which conducted on WEBSIFT [13] based on various search tools and services using query "latex software" is given below Table1. The query was intended to find both public-domain sources and commercial vendors for obtaining Latex software.

Table 1: Comparison of result for "latest software" query

Software Tool/ Service	***Disjunctive query***	***Conjunctive query***	***Phrase Query***
AltaVista	200,000	30,000	100
Excite	134,669	29,287	29,287
HotBot	3,696,449	61,830	17,630
InfoSeek Guide	3,111,835	427	100
Lycos	29,881	26	N/A
OpenText	481,846	2,541	6
WebCrawler	158,751	864	6
WWW Worm	4,999	2	N/A
Galaxy	6,351	20	N/A
Magellan	17,658	17,658	N/A
Yahoo	373 categories; 18,344 sites	1 category; 3 sites	N/A; 101 sites
IBM InfoMarket	100	N/A	N/A
MetaCrawler	29	32	34

From the result table it has been seen that among the search tools, Hotbot and InfoSeek has the largest retrieval power, more than 3 million; and WWWW is the least retrieval rate, 4,999. Hotbot has an important feature of searches, using search term and phrases. Also its relevance calculation is based on various factors such as frequency and document length. If the query term appears in the title or META tag has high relevance than others. InfoSeek Its main advantage is that it support searches for symbols, phrases also it searches images based on the caption. The relevance calculation in this method is also based whether the query term appears in the beginning of the documents. So these may be the reason for their dominance in the field of retrieving documents.

4. Improving Retrieval Effectiveness

The latest Search tools are developed focusing on only query-processing speed and database size. The main reason for that is old HTML versions are insufficient in presenting document contents to search tools [5, 4]. Now they introduced META TAG feature, this gives a clue of what that document is. In almost all cases where all we use search tools, for each search query it returns thousands of relevant documents. It is actually again a burden to the user to search into that and identify the appropriate one for them. So here this research is focused on the area of how can provide less or short relevant list of documents to the user according to the query. After comparing major search tools, found out major features than can give more promising results are:

4.1 Relevance Feedback Techniques

Query construction plays a vital role in retrieval efficiency, even though, it is not always possible to restrict the user. [6, 18]. But users can provide feedback about the documents retrieved for their query. These feedbacks can substantially improve the retrieval process. Instead of simple say this document retrieved is not relevant it is more preferable to give relative feedbacks. In order to use relevance feedback technique in search engines, they require changing their document representation itself. So It will become be more communicative and semantically thriving than just indexing the title. That means they need to index the entire content but not simply title. User

feedback has an important role in measuring the relevance of a document in an information retrieval system[2]. The traditional method of precision and recall treat relevance as only a two-leveled notion. So major component required in an information retrieval system is consideration of user feedback before indexing the documents. This gives a privilege to the user to judge the retrieved documents and give their option of top documents in the retrieved list according to their query [17]. This new list again classifies and re ranked, so definitely it has a major role in determining the final set of retrieved items.

4.2 Term Frequency

While analyzing the previous search tool result it is obvious that most well performed tools have considered an important feature for ranking, term frequency. So this is also another feature required for an efficient IR system. According to the concept of term frequency a document is considered as bag of words. Each word in the document is associated with a weight, and the documents are ranked based on the weight of query terms present in that particular document.

4.3 Inward Link

Google proves that ranking based on inward link is inevitable. Link structure helps to determine which web pages are to be added to the collection of relevant documents, and how to order them. Definitely there is no doubt that if we incorporate relevance feedback system with term frequency and inward link for relevance calculation certainly will show an effective information retrieval.

5. Conclusions

None of the search tools integrate the techniques relevance feedback, term frequency and inward link for relevance calculation of web pages. The quality in indexing web documents has an awesome effect on retrieval. Undoubtedly the incorporation of these techniques by a search tool will significantly dig over irrelevant documents and ranked relevant documents in the top. An efficient information retrieval algorithm should compatible with international standards and able to meet out all the challenges efficiently. This study is limited to an analysis of some of the major search tools and its methodologies, to find out some features which can incorporate to the IR system to improve its efficiency. As future work we can do a study on how we can improve the information retrieval process by considering different dimensions and develop a system.

References

[1] V.N. Gudivada; "Information Retrieval in the World Wide Web"; IEEE Internet Computing"; 1997.

[2] Patricio Galees; " Word Distribution Analysis for Relevance Ranking and Query Expansion"; Lecture notes in Cmputer Science; 2008.

[3] D, Shenk. "Data Smog: Surviving the Information Glot." NewYork: Harper and Collins, 1997.

[4] Etizioni, O. "The World Wide Web: Quagmire or Gold Mine?" ACM Communication 39 (1996): 65 - 68.

[5] Etzioni, O., and D. Weld. "Intelligent Agents on the Internet, Fact, Fiction and Forecast." IEEE Expert 10 (1995): 44 - 49.

[6] Frakes, W. B., and R. Baeza-Yates. Information Retrieval: Data Structures and Algorithms. Prentice Hall, Englewood Cliffs, 1992.

[7] Frontier, Challenges and Opportunities for Information Retrieval . The Second strategic workshop on information retreival in Lorne, SWIRL, February 2012.

[8] G, Salton. "The Smart Retreival System: Experiments in Automatic Document Processing." Prentice Hall, Englewood Cliffs, 1971.

[9] Kunhchu, Ibrahim. "Web based Evolutionary and Adaptive Information Retrieval." IEEE Transactions on Evolutionary Computation, April 2005.

[10] Lewis, David D. Active by Accident: Relevance feednback in IR. AAAI Fall Symposium on Active Learning: Unpublished working notes, 1995.

[11] Pothula, Sujatha. "A Review on Qualitative Analysis on Multilingual Information Retreival System." International Journal for REsearch in Science and ADvanced Technologies, Aug 2012.

[12] Raghavan, V., and S.K.M. Wong. "A Critical Analysis of Vector SPace Model for Information Retrieval." J. Am. Soc. Information Science 37 (1986): 279 - 287.

[13] Robng, Tane, Srivastava Jaideep, and Nert, Cooley Pang. "WEBSIFT: The Website Information Filter SYstem." Supported by NSF, ARL, 1999.

[14] S. E, Robertson, and Jones K. S. "Relevance Weighting of Serach Terms." Journal of the American Society for Information Sciences, 1976: 129 - 146.

[15] Steve, Lawrence, and Giles Lec. "Searching the Web: General and Scientific Information Access." IEEE Communication, Jan 1999.

[16] venkat, N Gudivada, Raghavan Vijay V, Grosky William L, and Kasanagothi Rajesh. "Information Retreival on the World Wide Web." IEEE Internet Computing, 1999.

[17] Xuehuashen, Bin Tan, and Cheng Xiang Zhai. "Contexr Sensitive Information REtrieval Using Implicit Feedback." SIGIR'05 August 15 - 19. ACM, 2005.

[18] Ye, Lu, Chunhui Hue, Xing Quan Zhu, Hong Jiang Zhang, and Qiang Yang. "A Unified Framework for Semantics anf Feature Based Relevance Feedback in Image Retrieval Systems." ACM Multimedia, 2000.

Ms. Anupama Prasanth, holds a Master's degree in Computer Applications from Bharatiyar University, Coimbatore and is currently pursuing her PhD from Karpagam University Coimbatore.

Dr. M. Hemalatha completed M.Sc., M.C.A., M. Phil., Ph.D (Ph.D, Mother Terasa women's University, Kodaikanal). She is Professor & Head and guiding Ph.D Scholars in Department of Computer Science at Karpagam University, Coimbatore. Twelve years of experience in teaching and published more than hundred papers in International Journals and also presented more than eighty papers in various national and international conferences. She received best researcher award in the year 2012 from Karpagam University. Her research areas include Data Mining, Image Processing, Computer Networks, Cloud Computing, Software Engineering, Bioinformatics and Neural Network. She is a reviewer in several National and International Journals.