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. 4, Issue 3, March 2016

A Study on World Wide Web Information Retrieval and Web Search Techniques

S.Balan[1], Dr. P.Ponmuthuramalingam[2]

Ph.D. Research Scholar, Department of Computer Science, Government Arts College (Autonomous), Coimbatore, Taminadu, India[1]

Associate Professor& Head**,** Department of Computer Science, Government Arts College (Autonomous), Coimbatore, Taminadu, India[2]

ABSTRACT: This paper is concerned with study and analysis of information retrieval and web search. To extract and uncover knowledge from web documents and services is said to be web mining. Generally web contains huge collection of documents plus hyperlink information to access the web pages. Information retrieval is given as source of document based on the user's query. It deals with various information retrieval techniques and web search methods, the analysis of both methods are discussed and listed some of the existing methods to find the results on this survey. The result not only proves the advantage of those methods but also compared the previous extraction and searching techniques.

KEYWORDS: Information Retrieval, Web Search, Web Mining, Web Data Mining, Searching Techniques.

I. INTRODUCTION

The World Wide Web (WWW) is commonly known as hyper media information retrieval or client server model. Hypertext was invented by Ted Nelson in 1965. Tim-Berner's Lee was invented the web in 1989. Then in 1989 Lee invented the distributed hyper text system for communication between client and server. HTTP, HTMl, URL are also started in the year of 1993. Marc Andreessen released the first mosaic for x graphical web browser for UNIX. Then in mid 1994 jm Clark collaborated with marc and then released Netscape browser [3]. In 1995 Microsoft entered in the market to challenge nets cape browser. In 1969 & 1972 ARPANET connections were made and TCP/IP was developed by Vinto Cerf & bob khan in the year of 1973. TCP/IP was birth and connected as internet in the year of 1982. In 1993 search systems was introduced by Stanford university students. Then in 1994 yahoo was created by jerry yang and David filo. Google was launched in 1998 by Sergey Brin and Larry page. Msn search engine was launched in the year of 2005.

Web data mining is said to be the amount of data on the web is still growing and contains all types of data. It makes challenging for integrating multiple pages and hyperlink with many other pages. The information on the web is noisy it contains two source i.e., a web page contains a piece of information and another one web does not have quality control of information. Data mining is also known as Knowledge Discovery Process (KDD). To discover useful patterns in the web, there are many data mining tasks some of the common tasks are classification, clustering, association, rule mining and sequential pattern mining. It is carried out in three main steps namely pre-processing, data mining and post processing. Web mining is used to discover useful information or knowledge from the web. It is categorized into three types web structure mining, web content mining and web usage mining. It is similar to data mining process, the main difference is data collection and the same three step process is as follows data pre-processing, web data mining, and post pre-processing [3].

Information retrieval helps the user find needed information from a large collection of text documents. The documents are web pages; generally the architecture of information retrieval contains the user, query operations, retrieval system, document index, indexer and document collection. The user query represents the user's information such as keyword queries, Boolean queries, phrase queries, proximity queries, full document queries, natural language

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. 4, Issue 3, March 2016

questions. The Boolean model is one of the earliest and simplest information retrieval model, both the query and retrieval are based on Boolean algebra such as document representation, Boolean queries, and document retrieval. Vector space model is widely used in information retrieval model such as document representation, term frequency scheme, inverse document frequency and queries, statistical language model based on probability and have foundations in statistical theory [3]. To improve the retrieval effectiveness there are many techniques to retrieve the result such are the rocchio method, machine learning methods and pseudo relevance feedback.

II. RELATED WORK

Collection of documents is used for retrieval. There are some traditional text documents such as stop word removal, stemming and handling of digits, hyphens, punctuations and cases of letters. The basic method of web search and information retrieval is used to find the documents based on the user query, there are various schemes available such as inverted index, index construction, index compression. Each scheme includes a method for coding and decoding. These methods are unary coding, elias gamma coding, elias delta coding, golomb coding, variable type coding. The web search method is based on crawling the pages on the web. Those pages are then parsed, indexed and stored; the operations of web search are as follows parsing, indexing, searching and ranking, occurrence type, count and position. Several search engines used together to produce meta search engine. Some techniques with ranking mechanism using combination using similarity scores and combination using rank positions.

There are some bulk websites and links are available in the web [1], to retrieve the useful information various tools and search techniques are used to fetch the information [6][7]. Keyword search is very difficult and its drawback of the existing tool is precision and recall [4][15]. There are various model bring information retrieval research based and web search engine. Searching for information is mentioned by vannevar bush in the year of 1945[11]. After that information retrieval was introduced and many techniques are brought in. web search engines are scaling up in the year of 1994-2000 to improve the search quality some goals are designed they are improved search quality, academic search engine research [8]. Mainly the information is retrieved based on the text, so the text retrieval is based on full text scanning, signature files, inversion, and vector model and clustering, using semantics information and so on.

Electronic search tools may interpret search terms using Boolean operators, phrase and proximity searching, truncation or wild card functions, case sensitivity, fields, stop words and relevance sorting. The important problem of information retrieval is user analysis the problem depends upon the information needs, how to process the information, is it closer to finding the solution for the problem, knowledge representation, processing of information and system evaluation. The general applications of information retrieval systems are as follows digital library, search engines and media search [2].

III. ANALYSIS OF INFORMATION RETRIEVAL AND SEARCH TECHNIQUES

Sno	Author Name	Title	Models	Techniques
1	Tanveer J. Siddiqui, U. S. Tiwary	Integrating Notion of Agency and Semantic in Information Retrieval multi-agent model	Intelligent multi-agent method	Combination of conceptual graph and multi-agent model
2	Yi Xiao, Ming Xiao, Fan Jhang	Intelligent Information Retrieval Model Based on Multi-Agents	Multi agent system	Analysis agent, filter agent, feedback agent.
3	Jianguo Jiang, Zhongxu Wang, Chunyan Liu, Zhiwen Tan, Xiaoze Chen, Min Li	The Technology of Intelligent Information Retrieval Based on the Semantic Web	Semantic web	Precision and recall
4	Wenjie Li,	Semantic Web-	Ontology and	Ontology web language

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. 4, Issue 3, March 2016

	Xiaohuan Zhang, Xiaofei Wei	Oriented Intelligent Information Retrieval System	multi agent model	(OWL), Simplified object access protocol (SOAP).
5	Bruno Antunes, Paulo Gomes and Nuno Seco	SRS: A Software Reuse System based on the Semantic Web	Software reuse system and semantic web	RDF Schema and software development knowledge element
6	Pan Ying, Wang Tianjiang, Jiang Xueling	Building Intelligent Information Retrieval System Based on Ontology	Ontology	Semantic retrieval
7	Urvi Shah, Tim Finin, Anupam Joshi, R. Scott Cost, James Mayfield	Information Retrieval on the Semantic Web	Ontology	Text and semantic markup language
8	Hany M. Harb, Khaled M. Fouad, Nagdy M. Nagdy	Semantic Retrieval Approach for Web Documents	Domain ontology and semantic retrieval	Semantic information retrieval techniques
9	Shi-Yi Xie, Jia-Cun Liu, Han Wang	Research Of intelligent Information Retrieval System based on Three Layers Agent Structure	Artificial intelligence, priority algorithms	Three layer agent structure
10	Sergey Brin and Lawrence Page	The Anatomy of a Large-Scale Hypertextual Web Search Engine	Page rank	Intuitive Justification, anchor text
11	Manish Sharma, Rahul Patel	A Survey on Information Retrieval Models, Techniques And Applications	Searching	Linera search, brute force search, binary search
12	Sharon Coward	Internet Search Techniques	Search tools	Search engines, meta search engine, information gateways, invisible / deep web
13	Ed Greengrass	Information Retrieval: A Survey	Information retrieval	Boolean approach, vector space approach
14	Eugene Agichtein, Eric Brill, Susan Dumais	Improving Web Search Ranking by Incorporating User Behavior Information	implicit relevance feedback	Representing User Actions as Features, Deriving a User Feedback Model

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. 4, Issue 3, March 2016

IV. CONCLUSION AND FUTURE WORK

This research aims with study and analysis of web information retrieval and web searching techniques. The information retrieval is used to find the relevant information based on the user's interest. Similarly web search contains huge amount of information connected from around the world. To extract the information retrieval methods are used to find out the solution. To improve the quality and effectiveness of search mechanism between the methods in various problems may use both the techniques and identified the results. It remains the depth of the research is interesting in future.

REFERENCES

1. Ankita Sharma, 'Intelligent Information Retrieval System: A Survey', Advance in Electronic and Electric Engineering, ISSN 2231-1297, Volume 3, Number 1 (2013), pp. 63-70
2. Algorithms for Information Retrieval – Introduction, Lab module 1.
3. Bing Liu, 'Web Data Mining, Exploring Hyperlinks, Contents, and Usage Data', ACM Computing Classification (1998): H.2, H.3, I.2, I.5, E.5, ISBN-10 3-540-37881-2, and Springer Berlin Heidelberg, New York.
4. Eddie C.L. Chan, George Baciu, s.C. Mak, 'Cognitive Location-Aware Information Retrieval by Agent-based Semantic Matching', 8th IEEE International Conference of cognitive informatics (ICCI09), IEEE 2009.
5. Hany M. Harb, Khaled M. Fouad, Nagdy M. Nagdy, 'Semantic Retrieval Approach for Web Documents', (IJACSA) International Journal of Advanced Computer Science and Applications,Volume 9, 2011 .
6. Jianguo Jiang, Zhongxu Wang, Chunyan Liu, Zhiwen Tan, Xiaoze Chen, Min Li ,'The Technology of Intelligent Information Retrieval Based on the Semantic Web"',2nd International Conference on Signal Processing Systems (ICSPS), IEEE 2010
7. Mohd Wazih Ahmed, Dr. M. A. Ansari ,'A survey: Soft computing in Intelligent Information Retrieval Systems', International Conference on Computational Science and Its Applications, IEEE 2012.
8. Oliver A. McBryan,' GENVL and WWWW: Tools for Taming the Web', First International Conference on the World Wide Web, CERN, Geneva (Switzerland), May 25-26-27 1994. http://www.cs.colorado.edu/home/mcbryan/mypapers/www94.ps.
9. Pan Ying, Wang Tianjiang, Jiang Xueling, 'Building Intelligent Information Retrieval System Based on Ontology"',The Eighth International Conference on Electronic Measurement and Instruments, IEEE 2007.
10. Shi-Yi Xie, Jia-Cun Liu, Han Wang , 'Research Of intelligent Information Retrieval System based on Three Layers Agent Structure' , Proceedings of the Second International Conference on Mache Learning and Cybernetics, Xi'an, IEEE 2003.
11. Singhal, Amit, 'Modern Information Retrieval: A Brief Overview', Bulletin of the IEEE Computer Society Technical Committee on Data Engineering, 2001.
12. Tanveer J. Siddiqui, U. S. Tiwary ,'Integrating Notion of Agency and Semantic in Information Retrieval multi-agent model', Proceeding of the 2005 5th International Conference on Intelligent Systems Design and Applications (ISDA'05) , IEEE 2005.
13. Wenjie Li, Xiaohuan Zhang, Xiaofei Wei, 'Semantic Web-Oriented Intelligent Information Retrieval System', International Conference on BioMedical Engineering and Informatics, IEEE 2008.
14. Yi Xiao, Ming Xiao, Fan Jhang ,'Intelligent Information Retrieval Model Based on Multi-Agents',IEEE 2007.
15. Youssef Bassil, 'A Survey on Information Retrieval, Text Categorization, and Web Crawling', Journal of Computer Science & Research (JCSCR) - ISSN 2227-328X, Vol. 1, No. 6, Pages. 1-11, December 2012