Volume : 5 | Issue : 10 | October 2016 • ISSN No 2277 - 8179 | IF : 3.508 | IC Value : 69.48

Original Research Paper

A Study on Working of Web Crawler

Computer Science

KEYWORDS : Web Crawlers, Seed URL, Focused Crawling, Spider, Parallel Crawling, Tropical Crawling

Niti Saxena — M.Tech., Computer Science and Engineering, Galgotias College of Engineering and Technology, Greater Noida

ABSTRACT

As with growing technological research and advancements, information on the web is increasing vastly and exponentially. Every second we have increase in valuable information. But information retrieval on time is challenge for survival with abundance of data on web and different user perspective. While searching out for some data, unlimited or limited based on the search, thousands or hundreds of results were appear as result. As far user time is precious and he or she might be impatient to look or go through each and every page listed. Search engine basically helps us getting track of information user wants to access and are interested in. It tracks the information based on user preferences and sort results within some specified time frame according to relevancy of user preferences. The top results are the main attraction for user which is a big challenge as there are number of technologies that lists out the relevant information according to user preferences as fast as possible. Crawling is one of such an important underlying technique that makes the search engine more effective, reliable and useful. Web Crawlers are techniques that traverse through web searching and fetch out the pages for relevant and preferable information inside specific subject area from millions, billions or trillions of web pages on internet.

INTRODUCTION

With technological growth and advancement of the World-Wide Web poses unpredictable challenges for general-purpose crawlers and search engines. A focused crawler or topical slacker is a web crawler that attempts to download only web pages that are related to a pre-defined or given set of keywords. Topical crawling generally assumes that only the keyword is given, while focused crawling also assumes that few labeled keywords of required and not required pages are available. A focused crawler is also described as a crawler which returns web pages relevant for a given keyword in the web. A focused crawler has the following main components:

A way to determine if a particular web page is relevant to the given keyword, and A way to determine how to proceed from a known set of pages.

An early search engine which deployed the focused crawling strategy was proposed in [1] based on the intuition that relevant pages often contain relevant links. It searches deeper when relevant pages are found, and stops searching at pages not as relevant to the keyword. Unfortunately, the above crawlers show an important drawback when the pages about a keyword are not directly connected in which case the crawling might stop pre-maturely. The attempt to apply business process modeling and mining technique to analyzing online knowledge sharing activities [2]. An initial ontology design and the architecture of a distributed information system that they were implementing [3].Among the discovered knowledge; sequential-pattern mining is used to discover the frequent subsequences from a sequence database. Most research handles the static database in batch mode to discover the desired sequential patterns [4].The proposed framework, which we call structural generative descriptions moves.

The structural time series representation to the probability domain, and hence was able to combine statistical and structural pattern recognition paradigms in a novel fashion [5].

The learned about the typical behavior of terrorists by applying a data mining algorithm to the textual content of terror-related Web sites. The resulting profile was used by the system to perform real-time detection of users suspected of being engaged in terrorist activities [6]. They implemented a prototype system of the focused crawler - a keyword-specific news gathering system which was prepared for comparative experiments on different similarity measures with the anchor text [7]. Malicious adversaries could deviate arbitrarily from their prescribed protocols. Secure protocols that are developed against malicious adversaries require utilization of complex techniques. Clearly, protocols that can withstand malicious adversaries provide more security [8]. Domain-specific internet portals are growing in popularity because they gather content from the Web and organize it for easy access, retrieval and search [9].

The URL analysis models of the existing focused crawler, and also their pros and cons, then they proposed a URL analysis model based on the improved genetic algorithm, in which the selection operator, crossover operator and mutation operator are optimized [10]. Distributed and dynamic nature of Web resources is a major problem for search engines maintain up-to-date index of the Web content as they have to crawl the Web periodically. A focused or keyword-driven crawler is a specific type of crawler that analyzes its crawl boundary to find the links that are likely to be most relevant for the crawl while avoiding irrelevant regions of the Web [11]. They compared different data mining methods and techniques for classifying students based on their Moodle usage data and the final marks obtained in their respective courses [12]. Effective relevance prediction can help avoid downloading and visiting many irrelevant pages, they proposed a new learning-based approach to improve relevance prediction in focused Web crawlers [13]. The usage of internet is immense on a large scale for the past many years. Especially more than 90% of people are using search engine. People are using search engine largely for their keywords [14]. Because of the complex Web structure, most approaches of focused crawling employ a local search algorithm, which will only search pages in a sub-graph of the Web. And the multi-keyword feature of Web pages makes it difficult to determine the relevance of a Web page [15]. An overlay-based parallel data mining architecture, which executes fully distributed data management and processing by employing the overlay network, can achieve high scalability [16]. A program that traverses the Internet by choosing relevant pages to a predefined keyword and neglecting those out of concern [17].They proposed an UBFC (URL rule based focused crawler) algorithm based on a double-crawler framework (an experimental crawler and a focused crawler) [18].

WORKING OF WEB

To view a Web page on the World Wide Web [19], the procedure starts by typing the URL into a Web browser, or by following a hyperlink to that page. The Web browser then gives some messages in order to fetch and display it. First, the server-name of the URL is resolved into an IP address that uses the domain name system, or DNS. This IP address is used to send data packets to the Web server.

Volume : 5 | Issue : 10 | October 2016 • ISSN No 2277 - 8179 | IF : 3.508 | IC Value : 69.48

The browser then requests the resource by sending an HTTP request to the Web server at that given address. In the case of a common Web page, the HTML text of the page is requested first and then parsed by the Web browser, which will then make requests for images and other files. All this searching within the Web is performed by the special engines that are known as Web Search Engines [20].

TECHNIQUE OF HOW SEARCH ENGINE PRESENTS INFORMATION TO THE USER INITIATING A SEARCH

When you ask a search engine to get the desired information, it is actually searches through the index which it has created and does not actually searches through the Web. Different search engines give different ranking results because not every search engine uses the same algorithm to search through all the indices.

THE QUESTION IS WHAT IS GOING ON BEHIND THESE SEARCH ENGINES AND WHY IS IT POSSIBLE TO GET RELEVANT DATA SO FAST?

The answer is web crawlers. The web crawler is a software program that traverses the web by downloading the pages and follows the links from page to page. Such programs are also called wanderers, robots, spiders, and worms. The structure of the World Wide Web is a graphical structure, i.e. the links of a page are used to open other web pages. Internet is a directed graph, web page as node and hyperlink as edge, so the search operation is a process of traversing the directed graph. By following the linked structure of the Web, we can traverse a number of web-pages starting from a seed page. Web crawlers are used to create a copy of all the visited pages for later processing by a search engine that will index the downloaded pages that will help in fast searches. Web search engines work by storing information about many web pages, which they retrieve from the WWW. These pages are retrieved by a Web crawler. Web crawlers are programs that use the graph structure of the web to move from page to page.

It is a simplified Web crawler in Figure 1. According to Figure 1, a Web crawler starts from a URL called the Seed URL to visit the Internet. The Page Downloader gets a URL from URL List to download the page and gives page to the Link Extractor. The Page Downloader checks whether to download pages or not. As the crawler visits these URLs, the Link Extractor identifies all the hyperlinks whether they are according to the requirements and transfers them to the URL Filter, and finally stores the results into URL list. The Crawling Parameter Assistor provides the parameter setting for the needs of all parts of the crawler.

Web crawler was internet's first search engine that has performed keyword searches in both names and texts of the page. It was developed by Brain Pinker-ton, a computer student at the University of Washington [22].

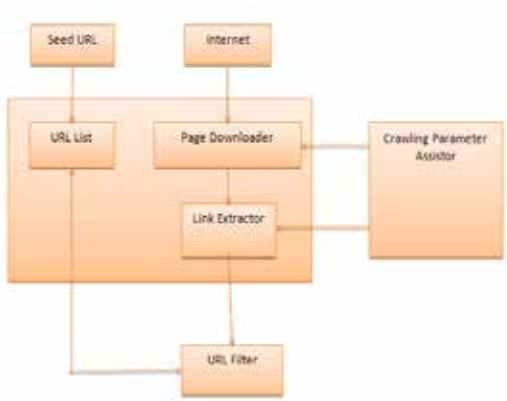

Figure 1: A simplified web crawler [21]

TYPES OF SEARCH ENGINES

The search engine belongs to 3 different categories and all are unique. All are having different rules and procedures .There is basically 3 types of search engines [20, 23].

Crawler Based Search Engine

Such search engines uses crawlers to categorize the web pages. Crawlers visit a Web Site to find information on internet and store it for search results in their databases. Crawler finds a Web page, downloads it and analyzes the information presented on web page. The web page will then be added to search engine's database. When a user performs a search, the search engine will check its database of Web pages for the keywords the user searched. The results are listed on the pages by order of which is closest.

Human-powered Search Engine

Such search engines rely on humans to give information that is indexed. Only information that is submitted by humans is indexed. This type of search engines are mostly used at small scale and rarely used at large scale. A Directory uses human editors that decide the site belongs to which category. They place Websites in 'directories' database. By focusing on particular categories, user narrows the search to those records that can be relevant.

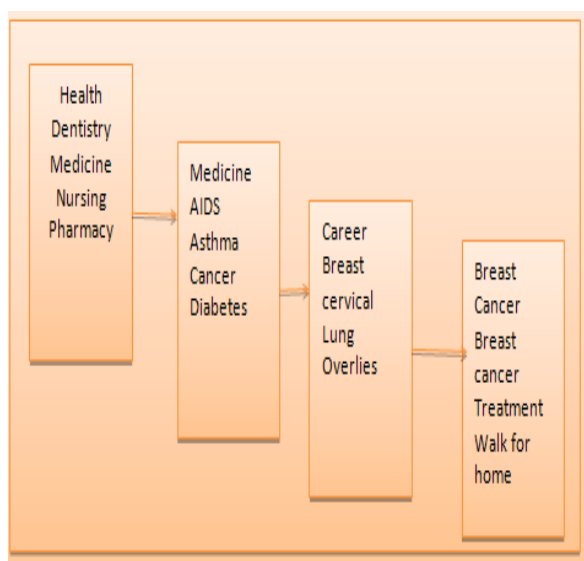

Figure 2: Directories of a search engine [24].

Hybrid Search Engine

Hybrid search engines use a combination of both crawler based results and directory results. It differs from traditional search engine such as Google or a directory based search engine such as yahoo in which the programs operates by comparing a set of metadata. Examples of hybrid search engines are: Yahoo, Google.

DEFINITION OF WEB-CRAWLER

A web-crawler is a program or automated script which browses the World Wide Web in a methodical and automated manner. To move from page to page web crawlers uses the graphical structure of the Web [20, 25]. Such programs are also called wanderers, robots, spiders, and worms. The World Wide Web has a graphical structure, i.e. the other pages are opened by traversing the links given in a page.

Actually Internet is a directed graph, web page as node and hyperlink as edge, so traversing the directed graph is

the search operation. Web crawlers are programs that exploit the graph structure of the web to move from page to page. However `crawlers' itself doesn't indicate the speed of these programs, so they are known as fast working programs [26].

A SURVEY OF WEB CRAWLERS

The original Google crawler [19, 28] was developed at Stanford. Topical crawling was first introduced by Menczer. Focused crawling was first introduced by Chakrabarti et al. [27, 29] A focused crawler has the following components: (a) How to know whether a particular web page is relevant to given keyword, and (b) way to determine how to follow the single page to retrieve multiple set of pages. A search engine which used the focused crawling strategy was proposed in [30] based on the assumption that relevant pages must contains only the relevant links. So it searches deeper where it founds relevant pages, and stops searching at pages not as relevant to the keyword. But, the above crawlers are having a drawback that when the pages about a keyword are not directly connected the crawling can stop at early stage. They keep the overall number of downloaded Web pages for processing [31] to a minimum while maximizing the percentage of relevant pages. For high performance, the seed page must be highly relevant. Seed pages can also be selected among the best results retrieved by the Web search engine [32, 33].

A standard crawler followed a breadth first strategy. If the crawler starts from a webpage which is n steps from a target document, we have to download before all the documents that are up to n-1 steps from the starting document.

A focused crawler identifies the most relevant links, and ignores the unwanted documents. If the crawler has to start from document that is n steps from target document, it downloads a subset of the documents that are maximum n-1 steps from the starting document. If the search strategy is optimal, then the crawler takes only n steps to discover the target.

A focused crawler efficiently seeks out documents about a specific keyword and guides the search based on both the content and link structure of the web [26]. A focused crawler implements a strategy that associates a score with each link in the pages it has downloaded [34, 35 and 36].

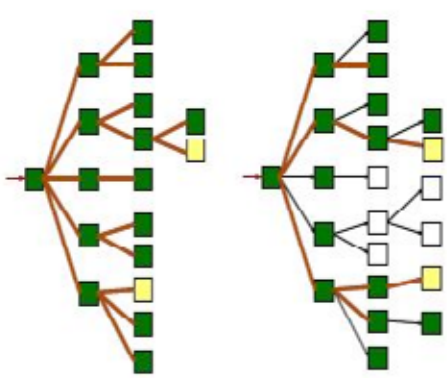

Figure 3: a) Standard Crawling b) Focused Crawling

A topical crawler ideally downloads only web pages that are relevant to a particular keyword and avoid downloading the irrelevant pages. So a topical crawler can predict the probability that a link to that page is relevant before actually downloading the page. A predictor can be the anchor text of links; and this approach was taken by Pinkerton [37]. Menczer et al. [38] show that simple strategies are very effective for short crawling, while techniques such as reinforcement learning [39] and evolutionary adaptation gives the best performance for longer crawling. Diligent et al. [40] use the complete content of the pages that are visited already to get the similarity between the query and the pages that have not been visited yet. Guan et al [41] propose a new frontier prioritizing algorithm which efficiently combines link-based and content based analysis to evaluate the priority of an uncrawled URL in a queue.

Approaches to focused crawling are Best first approach, Info spiders, Fish search and Shark search. In Best first approach, [42] we have given a Frontier of links and the next link is selected on the basis of some priority or score. So every time a best available link is opened and traversed. Info spiders use neural networks. Info Spiders [43, 44] is a multi-agent system for online, dynamic Web search. Fish search [45] is based on the assumption that relevant pages must have relevant neighbors. Thus, it searches deeper on the documents that are found relevant to the search query, and do not search in "dry" areas.

In Fish-search algorithm Internet is treated as a directed graph, webpage as node and hyperlink as edge, so the search operation is the process of traversing directed graph. For every node we judge whether it is relevant, I means the node is relevant and 0 for irrelevant. So all the relevant pages are assigned the same priority value. The list of URLs which is maintained are having different priority, the URL which are at the front of the list are more superior, and will be searched sooner than others. If relative page is found, it stands for that the food has been found by the fish. However Fish Search algorithm has some limitations, so a powerful improved version of Fish Search algorithm is developed known as- Shark Search [46]

In this algorithm, the improvement is that instead of the binary (relevant/irrelevant) evaluation, it returns a "fuzzy" score, i.e., a score between 0 and 1 (0 for no similarity and 1 for perfect "conceptual" match) rather than a binary value. In shark search we have found a threshold value which can determine the relevance of the page. However, Best first crawlers have been shown better results in case of info spiders and shark search and other non-focused breadth first crawling approaches. So, best first crawling is considered to be the most successful approach to focused crawling due to its simplicity and efficiency.

WORKING OF BASIC WEB CRAWLER

The basic working of a web-crawler can be discussed as follows:

1. Select a starting seed URL or URLs.
2. Add it to the frontier.
3. Now pick the URL from the frontier.
4. Fetch the web-page corresponding to that URL.
5. Parse that web-page to find new URL links.
6. Add all the newly found URLs into the frontier.
7. Go to step 2 and repeat while the frontier is not empty.

Note that it also depicts the 7 steps given earlier. Such crawlers are called sequential crawlers because they follow a sequential approach.

PARALLEL CRAWLERS

The size of the Web grows exponentially, so it is very difficult to retrieve the significant pages of the Web from a large number of web pages by using a single sequential

Volume : 5 | Issue : 10 | October 2016 • ISSN No 2277 - 8179 | IF : 3.508 | IC Value : 69.48

crawler. Therefore, multiple processes are run by the search engines in parallel to perform the task of getting relevant pages, in order to maximize the download rate. We call this type of crawler as a parallel crawler. Parallel crawlers as the name indicates work parallel to get the pages from the Web and add them to the database of the search engine [48].

Each parallel crawler have its own database of collected pages and own queue of un-visited URLs. Once the crawling procedure finishes, the collected pages of every crawler are added to the database of the search engine. Parallel crawling architecture no doubt increases the efficiency of any search engine.

CRAWLING TECHNIQUES [21]

Distributed Crawling

The size of web is a single crawler process even if it is a multithreading process will be insufficient for large search engines that have to fetch large amount of data in a very less time. When a single crawler is used all the fetched data passes through a single physical link. By distributing the crawling makes the system scalable and easily configurable and also makes the system fault tolerable.

Focused Crawling

The goal of a focused crawler is to seek out pages that are selective and are relevant to a desired keyword. Therefore a focused crawler can predict the probability that a link to a particular page is relevant before actually downloading the page [38]. The performance of a focused crawler depends on the richness of links in the specific keyword being searched. The keywords are specified not using keywords, but using the documents, focused crawlers try to "predict" whether or not a target URL is pointing to a relevant web page before actually fetching the page. In addition, focused crawlers visit URLs in an optimal order such that URLs pointing to relevant and high-quality Web pages are visited first, and URLs that point to low-quality or irrelevant pages are never visited. This leads to significant savings in hardware and network resources, and helps to keep the crawl more up-to-date.

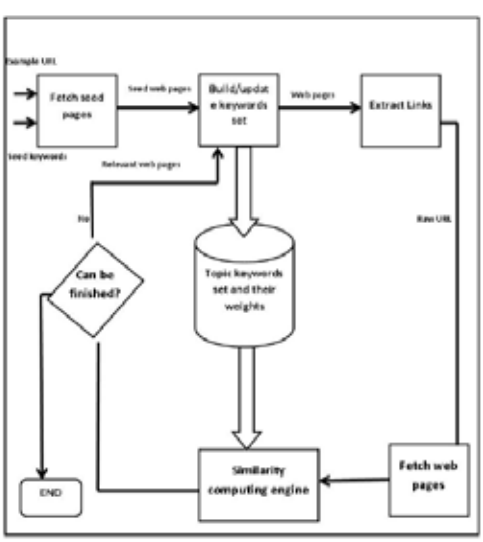

Figure 4: Focused Crawler working process [49]

Seed pages fetching subsystem

From the given seed keywords, the system searches them on a search engine. The result which is returned by search engine consists of a huge set. The top N (N<500) URL's are probably relevant to the keyword. The crawler uses these top N URLs as seed URLs and from these URLs, it fetches the seed pages.

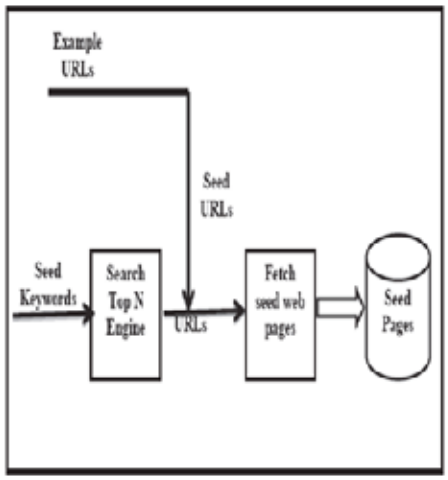

Figure 5: Fetch seed pages by seed keywords and example URLs [49]

Relevant keywords generating subsystem

If the documents are mostly relevant to the keyword, then it is easier to find the relevant keywords from them and this subsystem is designed to find relevant keywords from those documents For each word Ti in document, first the term frequency tf is counted by the system, and then retrieve its document frequency df and finally computes weight. Weight (i).The top N (N<50) highest weight keywords are outputted as keyword keywords set.

SIMILARITY COMPUTING ENGINE

When a crawler fetches a new page, it needs to judge the page whether or not the page is relevant to the keyword. The document D is that web page which has to be judged. The query Q is a set of searched keywords. The computing result is Similarity Sim (Q, D) and its float value is between 0 and 1.We have set a threshold as a standard for judgment of document relevance. If the value is higher, the precision of retrieved pages relevant to the keyword would be higher. But the recall would be lower.

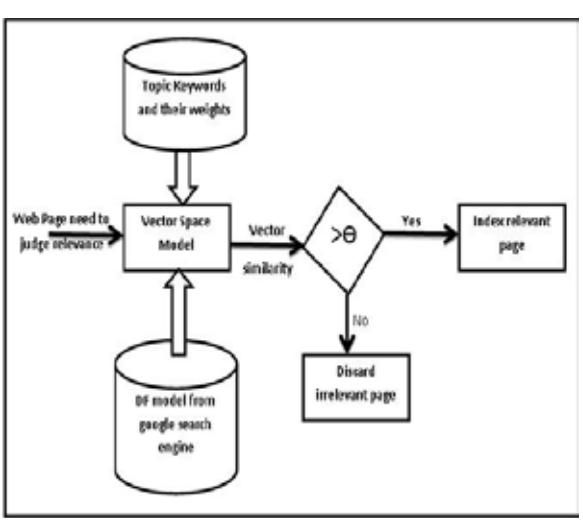

Figure 7: Similarity Computing Engine [49]

CONCLUSIONS

Web crawlers are the program that uses the graphical structure of the Web to move from page to page. A focused crawler is a crawler that targets a desired keyword and gathers only a relevant Web page which is based upon predefined set of keywords and do not waste resources on web pages that are not relevant. Best-first search is the most popular search algorithm used in focused crawlers. In best-first search, URLs are not just visited in the order they are present in the queue; instead, some rules are applied to rank these URLs .But we see there are multiple URLs and keywords on a single web page. So the time complexity of web page increases and it negatively affects the performance of focused crawling which also affects the relevancy score which describes that the web page is relevant for particular specified keyword decreases.

REFERENCES:

1. De Bra, P., Houben, G., Kornatzky, Y., and Post, R. "Information retrieval in distributed hypertexts",Proc. 4th RIAO Conference, 1994.
2. Alan Wang,G., Jiannan Wang, H., Li J. and WeiguoFan."Mining Knowledge Sharing Processes in Online Discussion Forums", 2014. Hawaii International Conference on System Science.
3. Bra, De P.M.E. and Post, R.D.J. "Information Retrieval in the World Wide Web making client based searching feasible", Computer Networks & ISDN systems, 1994 27(2) 183-192.
4. Chun-Weilin, J.WenshengGanandTzung-Pei Hong. "Efficiently Maintaining the Fast Updated Sequential Pattern Trees With Sequence Deletion", 2014.
5. Edgar, S.,García-Treviño and Javier A. Barria." Structural Generative Descriptions for Time Series Classification ", 2014. IEEE Transactions on cybernetics.
6. Elovici, Y., Kandel, A., Last, M., Shapira, B.andZaafrany, O. " Using Data Mining Techniques for Detecting Terror-Related Activities on the Web ".
7. Hong-Wei Hao, Cui-Xia Mu, Xu-Cheng Yin, Shen Li, "An improved topic relevance algorithm for focused crawling",Systems, Man, and Cybernetics (SMC), 2011 IEEE International Conference on ,9-12 Oct. 2011 ,pp: 850 - 855 .
8. Lindell, Y. andPinkas, B. "Privacy Preserving Data Mining", Journal of Cryptology 2000.
9. McCallum, A., Nigam, K., Rennie, J. and Seymore, K. "Automating the Construction of Internet Portals with Machine Learning". Information Retrieval.Vol 3, 127-163, 2000 .
10. Ning, H., Wu, H., He, Z. and Tan, Y. "Focused crawler URL analysis model based on improved genetic algorithm", Mechatronics and Automation (ICMA), 2011 International Conference on 7-10 Aug. 2011,pp: 2159 – 2164.
11. Ravakhah, M. and Kamyar, M. "Semantic Similarity Based Focused Crawling" ,Computational Intelligence, Communication Systems and Networks, 2009. CICSYN '09. First International Conference on 23-25 July 2009,pp: 448 – 453.
12. Romero, C.,Ventura, S., Pedro, G., Espejo and Hervás, C. "Data Mining Algorithms to Classify Students".
13. Safran, M.S.,Althagafi, A. and DunrenChe, "Improving Relevance Prediction for Focused Web Crawlers", IEEE/ACIS 11th International Conference on Computer and Information Science.
14. Su, C.,Gao, Y., Yang, J. and LuoB. "An efficient adaptive focused crawler based on ontology learning", Hybrid Intelligent Systems, 2005. HIS '05. Fifth International Conference on 6-9 Nov. 2005.
15. Sun, Y.,Jin,P., and Yue, L., 2008. "A framework of a Hybrid Focused Web crawler", 2nd international conference on Future Generation Communication & Networking Symposia.
16. Suto, K., Nishiyama, H., Kato, K., Mizutani, K., Akashi, O. and Takahara,A. "An Overlay-Based Data Mining Architecture Tolerant to Physical Network Disruptions",2014. IEEE Transactions on emerging topics in computing.
17. Wang, W.,Chen*, X.,Zou, Y., Wang, H. and Dai Z. "A Focused Crawler Based on Naive Bayes Classifier",Third International Symposium on Intelligent Information Technology and Security Informatics,2010 IEEE,pp:517-521.
18. Zhang, X., Zhou, T., Yu Z. and Chen, D. "URL Rule Based Focused Crawlers", IEEE International Conference on e-Business Engineering 2008.
19. Arvin Arasu, Junghoo Cho, Andreas Paepcke: "Searching the Web", Computer Science Department, Stanford University.
20. Mr.Ravinder Kumar, Sandeep Sharma: "Web Crawling Approaches in |Search Engines" CSE Department, Thapar University ,2008.
21. Brian Pinkerton,"Web Crawler: Finding what people want ", Doctor of Philosophy University of Washington, 2000.
22. Mr.Ravinder Kumar, Ravikiran Routhu: "Enrichment in Performance of Focussed Crawlers" CSE Department, Thapar University ,2010.
23. Fabrizio Silvestri, "High performance issues in web search engines: Algorithms and Techniques", PhD: Thesis: TD 5/04 ,available at :http://hpc.isti.cnr.it/_silvestr May 2004.
24. Paul De Vrieze, "Improving search engine technology, "Master thesis, 7-March-2002
25. Gautam Pant, Padmini Srinivasan, and Filippo Menczer: "Crawling the Web"available at- << http://dollar.biz.uiowa.edu/~pant/Papers/crawling.pdf >>
26. Lovekesh Kumar Desai,"A distributed approach to crawl domain specific hidden web", georgia state university, 2007.
27. Mark najork and Allan Heydon, "High Performance web crawling", SRC Research Report 173, published by COMPAQ systems research centre on Sep 26, 2001
28. Sergey Brin and Lawrence Page, "The anatomy of a large-scale hyper textual Web search engine", In Proceedings of the Seventh International World Wide Web Conference, pages 107–117, April 1998.
29. S. Chakrabarti. "Mining the Web". Morgan Kaufmann, 2003.
30. J. Kleinberg, "Authoritative sources in a hyperlinked environment." Report RJ 10076, IBM, May 1997, 1997.
31. Chang C, KayedM, Girgis MR and Shaalan KF,"A survey of Web Information Extraction systems", IEEE Transactions on knowledge and engineering.2006.
32. Kraft R and Stata R,"Finding buying guides with a web carnivore", First latin American Web Congress ,pages 84-92.,2003.
33. Pant G, Johnson J and Giles C.L," Panorama: Extending digital libraries with topical crawlers", Proceedings of the 4th ACM/IEEE-CS Joint Conference on Digital Libraries.2004.pages 142-150 [14] Kraft R and Stata R,"Finding buying guides with a web carnivore", First latin American Web Congress ,pages 84-92.,2003
34. S. Chakrabarti, B. Dom, P. Raghavan, S. Rajagopalan, D. Gibson, and J. Kleinberg
35. , "Automatic resource compilation by analyzing hyperlink structure and associated text," in Proc. 7th World Wide Web Conference, Brisbane, Australia, 1998.K. Bharat and M. Henzinger, "Improved algorithms for topic distillation in hyperlinked environments," in Proceedings 21st Int'l ACM SIGIR Conference., 1998.
36. J. Kleinberg, "Authoritative sources in a hyperlinked environment." Report RJ 10076, IBM, May 1997, 1997.
37. B. Pinkerton, "Finding what people want: Experiences with the web crawler," in Proceedings of the First International World-Wide Web Conference, Geneva,Switzerland, May 1994.
38. F. Menczer, G. Pant, and P. Srinivasan, "Topical web crawlers: Evaluating adaptive algorithms," ACM Transactions on Internet Technology (TOIT), Vol. 4, no. 4, pp. 378–419, 2004.
39. A. Grigoriadis, "Focused crawling using reinforcement learning," Master's thesis, The University of Edinburgh, 2003.
40. M. Diligenti, F. Coetzee, S. Lawrence, C. L. Giles, and M. Gori, "Focused crawling using context graphs," in 26th International Conference on Very Large Databases, VLDB 2000, Cairo, Egypt, 2000, pp. 10–14.
41. Z. Guan, C. Wang, C. Chen, J. Bu, and J. Wang, "Guide focused crawler efficiently and effectively using on-line topical importance estimation," in Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2008, Singapore, July 20-24, 2008, 2008, pp. 757–758.
42. Menczer F, Pant G and Srinivasan P," Topical Web Crawlers: Evaluating Algorithms "ACM Transactions on Internet Technology (TOIT), Nov, 2004.
43. F. Menczer and R. K. Below. Adaptive retrieval agents: "Internal-

Volume : 5 | Issue : 10 | October 2016 • ISSN No 2277 - 8179 | IF : 3.508 | IC Value : 69.48

izing local context and scaling up to the Web". Machine Learning, 39(2{3):203{242, 2000.

44. F. Menczer, G. Pant, and P. Srinivasan. Topical web crawlers: Evaluating adaptive algorithms. To appear in ACM Transactions on Internet Technologies, 2003.http://dollar.biz.uiowa.edu/Papers/TOIT.pdf. 40
45. P. DeBra and R. Post, Information retrieval in the worldwide web: making client-based searching feasible, in: Proc. of the 1st International World Wide Web Conference,Geneva, Switzerland, 1994.
46. M. Hersovici, M. Jacovi, Y.S. Maarek, D. Pelleg, M. Shtalheim and S. Ur, "The Shark- Search algorithm — an application: tailored Web site mapping," in: 7th World-Wide Web Conference, April, 1998, Brisbane, Australia, online at http://www7.scu.edu.au/programme/full papers/1849/com 1849.htm
47. Bing liu, "web data mining", Springer-Verlag Berlin Heidelberg, 2007
48. Junghoo Cho, Hector Garcia-Molina: "Parallel Crawlers", 7–11 May 2002, Honolulu, Hawaii, USA.
49. Yang Yongsheng, Wang Hui, " Implementation of Focussed Crawler" COMP 630D course Project Report,2000
50. Debashis Hati, Amritesh Kumar "Improved Focused Crawling Approach for Retrieving Relevant Pages Based on Block Partitioning". 2nd International Conference on Education Technology and Computer (ICETC) 2010.