ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 12, December 2015

Domain Specific Web Crawler: A Survey

Hetal J.Thankil[1], Pooja Sawant[2], Nisha Kasale[3], Sonam Yerawar[4], Prasad Belhe[5],

Swapnali G.Game[6]

Assistant Professor, Dept. of Computer Engineering, JSCOE, Hadapsar, Pune, India[1]

UG Student, Dept. of Computer Engineering, JSCOE, Hadapsar, Pune, India[2,3,4,5]

Assistant Professor, Dept. of Computer Engineering, JSCOE, Hadapsar, Pune, India[6]

ABSTRACT: The "Domain Specific web crawler" is a system or a program which we will design to search information from web. A crawler is an automation tool that traverses the web page for the purpose of extracting information. In domain specific Web crawler, the crawler crawls the pages, which are relevant to our domain. To find the domain related web page we need to visit a Web page and calculate the relevance value. Based on relevance value the selection of page is decided.For a crawler it is not an easy task to download the domain specific Web pages. It focus will be to identify Web pages for a particular domain in web. The usage of internet is immense on a large scale for the past many years. In this survey we are going to discuss about types of crawler ,crawling algorithm , relevancy check and machine learning.

KEYWORD: Crawler, Ant, Spider, bot, agent, Automation, Search Engine, Focused, web mining, text analysis

I. INTRODUCTION

A web crawler or spider is a computer program that browses the WWW in sequencing and automated manner. A crawler which is sometimes referred to spider, bot or agent is software whose purpose it is performed web crawling[5]. Today the size of the web is thousands of millions of web pages that is too high and the growth rate of web pages are also too high i.e. increasing exponentially due to this the main problem for search engine is deal this amount of the size of the web. The World Wide net (WWW) having billion websites and looking documents that is additional specific with the user's needs is progressively tough[3]. As more information becomes available on the Web, finding relevant information from it becomes more difficult too[23].

A web search engine is a software system that is designed to search for information on the World Wide Web. The search results are generally presented in a line of results often referred to as search engine results pages . The information may be a specialist in web pages, images, information and other types of files. Some search engines also mine data available in databases or open directories. Unlike web directories, which are maintained only by human editors, search engines also maintain real-time information by running an algorithm on a web crawler. A web crawler (also known as a robot or a spider) is a system, a program that traverses the web for the purpose of bulk downloading of web pages in an automated manner[9].

A web crawler is a program that, given one or more seed URLs, downloads the web pages associated with these URLs, extracts any hyperlinks contained in them, and recursively continues to download the web pages identified by these hyperlinks. Web crawlers are an important component of web search engines, where they are used to collect the corpus of web pages indexed by the search engine. The types of crawler includes horizontal (BFS), vertical (DFS), periodic, parallel, topical, domain specific, social network, semantic, and dynamic. Challenges of web crawler are web coverage, dynamic web, hidden web, deep web and freshness[9].

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 12, December 2015

The basic architecture of web crawler is given below (Figure1). More than 13% of the traffic to a web site is generated by web search [8]. Today the size of the web is thousands of millions of web pages that is too high and the growth rate of web pages are also too high i.e. increasing exponentially due to this the main problem for search engine is deal this amount of the size of the web. Due to this large size of web induces low coverage and search engine indexing not cover one third of the publicly available web. By analyzing various log files of different web site they found that maximum web request is generated by web crawler and it is on an average 50%.

II. RELATED WORK

Most related work in this area is associated with popular search engines and their crawling algorithms and detailed architecture is kept as a business secret. However, web crawlers such as RBSE, the WebCrawler, the World Wide Web Worm, the crawler of the Internet Archive, an early version of the Google crawler, Mercator, Salticus, the WebFountain and the WIRE have published descriptions of their architecture. Besides structural issues, research about Web crawling has focused on parallelism, discovery and control of crawlers for Website administrators, accessing content behind forms (the "hidden" web), detecting mirrors, keeping the freshness of the search engine copy high, long-term scheduling , and focused crawling. In addition, there have been studies on characteristics of the Web, that affect directly the performance of a Web crawler such as detecting communities, characterizing server response time, studying the distribution of web page changes and proposing protocols for web servers to cooperate with crawlers.

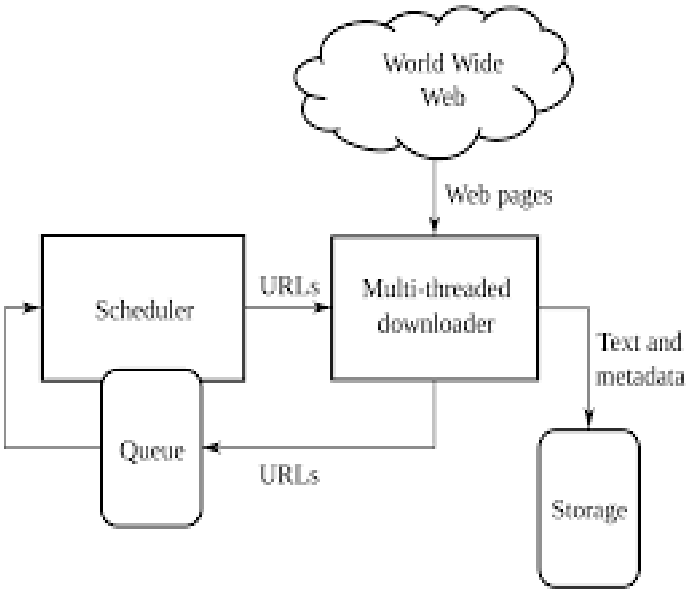

Figure1: Architecture of Crawler

Crawling the web is not a programming task, but an algorithm design and system design challenge because of the web content is very large. At present, only Google claims to have indexed over 3 billion web pages. The web has doubled every 9-12 months and the changing rate is very high.About 40% web pages change weekly when we consider slightly change,

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 12, December 2015

but when we consider changing by one third or more than the changing rate is about 7% weekly.

MAIN TYPES OF CRAWLER

1.Focused Web Crawler

Focused Crawler is the Web crawler that tries to download pages that are related to each other. It collects documents which are specific and relevant to the given topic. It is also known as a Topic Crawler because of its way of working . The focused crawler determines the following – Relevancy, Way forward. It determines how far the given page is relevant to the particular topic and how to proceed forward. The benefits of focused web crawler is that it is economically feasible in terms of hardware and network resources, it can reduce the amount of network traffic and downloads . The search exposure of focused web crawler is also huge.

2.Incremental Crawler

A traditional crawler, in order to refresh its collection, periodically replaces the old documents with the newly downloaded documents. On the contrary, an incremental crawler incrementally refreshes the existing collection of pages by visiting them frequently; based upon the estimate as to how often pages change. It also exchanges less important pages by new and more important pages. It resolves the problem of the freshness of the pages. The benefit of incremental crawler is that only the valuable data is provided to the user, thus network bandwidth is saved and data enrichment is achieved.

3.Distributed Crawler

Distributed web crawling is a distributed computing technique. Many crawlers are working to
distribute in the process of web crawling, in order to have the most coverage of the web. A central server manages the communication and synchronization of the nodes, as it is geographically distributed. It basically uses Page rank algorithm for its increased efficiency and quality search. The benefit of distributed web crawler is that it is robust against system crashes and other events, and can be adapted to various crawling applications.

4. Parallel Crawler

Multiple crawlers are often run in parallel, which are referred as Parallel crawlers. A parallel crawler consists of multiple crawling Processes called as C-procs which can run on network of workstations. The Parallel crawlers depend on Page freshness and Page Selection . A Parallel crawler can be on local network or be distributed at geographically distant locations. Parallelization of crawling system is very vital from the point of view of downloading documents in a reasonable amount of time.

III. LITERATURE SURVEY

Pp no.	Paper title	Techniques	Parameters achieved
1	Smart Crawler: A Two-stage Crawler for Efficiently Harvesting Deep-Web Interfaces (IEEE-2015)	Site Locating, In-site Exploring.	Coverage for deep web interfaces and maintains highly efficient crawling.
2	Topical Web Crawling for Domain-Specific Resource Discovery Enhanced by Selectively using Link-Context.(IJARI-2015)	Heuristic approach for selectively using link-context and ignore Anchor text.	Retrieving domain specific resources or more relevant web pages.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 12, December 2015

3	A Keyword Focused Web Crawler Using Domain Engineering and Ontology(IJARCCE-2015)	Using keyword extract the URL.	The amount of extracted documents was reduced. Link analyzed, and deleted a good deal of irrelevant websites.
4	Self-Adaptive Semantic Focused Crawler for Mining Services Information Discovery(IEEE-2014)	Heterogeneity, ubiquity, and Ambiguity.	Precisely and efficiently Discovering, formatting, and indexing mining service information over the Internet.
5	Study of Web Crawler and its Different Types(IOSR-JCE-2014)	Focused Web Crawler, Incremental Crawler, Distributed Crawler, Parallel Crawler.	Does not waste resources on irrelevant material. Not have dependency on previous crawler.
6	A Comparative Study of Hidden Web Crawlers(IJCTT-2014)	Surfacing, Virtual Data Integration.	Achieves high coverage with just a small number of queries
7	Web Crawling Algorithms(2014)	Breadth First Search, Best First Search, Fish Search, A* Search, Adaptive A* Search.	The algorithm can work efficiently in static as well as less dynamic environments.
8	SURVEY OF WEB CRAWLING ALGORITHMS(2014).	BFS, DFS, Page rank, Batch-page rank.	Reduce the network traffic and crawling costs.
9	Design and Development of a Domain Specific Focused Crawler Using Support Vector Learning Strategy (IJIRCCE-2014).	SVM, classification & regression Analysis.	Machine learning, flexibility.
10	Focused web crawling using named entity recognition for narrow domains(IJRET-2013)	Harvest ratio, Tunneling.	Making relevance judgments.
11	Enhancing the performance of web Focused CRAWLER using ontology (IJCT-2013)	Relevance calculation.	It reduces the number of results extracted.
12	An Extended Model For Effective Migrating Parallel Web Crawling With Domain Specific And Incremental Crawling.	Firewall mode, Cross-over mode, Exchange mode.	Yield high quality pages.
13	Design and Implementation of a Simple Web Search Engine	Page Rank, HITS, Salsa.	Anchor text analysis.
14	Novel Architecture of Web Crawler for URL Distribution (IJCST-2011)	cluster computing, load balancing.	URLs can be distributed for balancing the load on web crawler and improving its performance.
15	WEB CRAWLER - AN OVERVIEW(IJCSC-2011)	Selection policy, Re-visit policy Politeness policy, Parallelization policy.	Paper describes about Different types of Web crawler and the policies used in the web crawlers.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 12, December 2015

16	A New Approach to Design Graph Based Search Engine for Multiple Domains Using Different Ontologies (2011)	Relevance Calculation for All Domains, Checking Domain of a Web page.	Eliminate the problem of irrelevant results, highly scalable.
17	Combining text and link analysis for focused crawling—An application for vertical search engines(SD-2006)	Web information retrieval, Vertical search engines and focused web crawlers, Hypertext combined latent analysis (HCLA).	Ignores any understanding of language semantics.
18	Focused crawling: a new approach to topic-specific Web resource discovery	Classification,Distillation, Integration with the crawler.	Goal-directed crawling.
19	Crawling the Web	Naive Best-First Crawler, Shark Search Context Focused Crawler, Info Spiders.	Benefit of using the DOM structure and scalability benefits of distributed topical crawling.
20	Domain-Specific Web Site Identification: The CROSSMARC Focused Web Crawler	CROSSMARC, Site navigation, Page-filtering, Link-scoring.	Present the extraction results according to the user's personal references and constraints.

IV. CONCLUSION AND FUTURE WORK

A focused crawler is the web crawler that tries to download the pages that are related to each other. The relevancy in data decides the performance of crawler. Many researchers implemented different crawlers using different algorithm and methods to improve the QoS. In this paper the technique which are useful to speedup and give the relevant result are considered. The given table contains different techniques, methods and their achievement. The QoS is achieved by proper threading and rapid text mining. We came to know the advantages and limitation of different algorithm of web crawling. It is possible to create a Hybrid Crawler which will overcome the limitation of algorithm to improve the quality of crawler.

REFERENCES

1. Feng Zhao, Jingyu Zhou, Chang Nie, Heqing Huang, Hai Jin."Smart Crawler: A Two-stage Crawler for Efficiently Harvesting Deep-Web Interfaces", IEEE 2015.
2. Lu Liu, Tao Peng and Wanli Zuo. "Topical Web Crawling for Domain-Specific Resource Discovery Enhanced by Selectively using Link-Context", IJARI,Vol. 12, No. 2, March 2015.
3. Gunjan Agre, Snehlata Dongre. "A Keyword Focused Web Crawler Using Domain Engineering and Ontology", IJARCCE, Vol. 4, Issue 3, March 2015.
4. Hai Dong, Member, IEEE, and Farookh Khadeer Hussain. "Self-:Adaptive Semantic Focused Crawler for Mining Services Information Discovery" IEEE, VOL. 10, NO. 2, MAY 2014
5. Trupti V. Udapure, Ravindra D. Kale, Rajesh C. Dharmik."Study of Web Crawler and its Different Types", IOSR-JCE, e-ISSN: 2278-0661, p-ISSN: 2278-8727Vol. 16, Issue 1, Ver. VI (Feb. 2014), PP 01-05.
6. Sonali Gupta, Komal Kumar Bhatia. "A Comparative Study of Hidden Web Crawlers", IJCTT, Vol. 12, No. 3 , June 2011.
7. Aviral Nigam. "Web Crawling Algorithms", Vol. 4 Iss. 3, PP. 63-67, Sept. 2014.
8. Rahul kumar, Anurag Jain and Chetan Agrawal. "Survey Of Web Crawling Algorithms", AVC,Vol.1, No.2/3, September 2014.
9. M. Selvakumar, Dr. A.Vijaya."Design and Development of a Domain Specific Focused Crawler Using Support Vector Learning Strategy", JIRCCE, Vol.2, Special Issue 5, October 2014.
10. Sameendra Samarawickrama, Lakshman Jayaratne. "Focused Web Crawling Using Named Entity Recognition For Narrow Domains", IJRET, ISSN:

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 12, December 2015

2319-1163, 2013.
11. Prasant Singh Yadav, Mrs Mala Kalra, Dr. K.P Yada."Enhancing the performance of web Focused CRAWLer using ontology", IJCT, Vol. 4 No. 2, March-April 2013, ISSN 2277-3061.
12. Md. Faizan Farooqui, Dr. Md. Rizwan Beg and Dr. Md. Qasim Rafiq. "An Extended Model For Effective Migrating Parallel Web Crawling With Domain Specific And Incremental Crawling.", IJWSC, Vol.3, No.3, September 2012.
13. Andri Mirzal. "Design and Implementation of a Simple Web Search Engine", Vol. 7, No. 1, January, 2012.
14. Avanish Kumar Singh, Amit Kumar Pathak . "Novel Architecture of Web Crawler for URL Distribution", IJCST, Vol. 2, Issue 3, September 2011.
15. S.S. Dhenakaran and K. Thirugnana Sambanthan."WEB CRAWLER - AN OVERVIEW", pp. 265-267, Vol. 2, No. 1, January-June 2011.
16. Debajyoti Mukhopadhyay, Sukanta Sinha."A New Approach to Design Graph Based Search Engine for Multiple Domains Using Different Ontologies", IEEE-2008.
17. G. Almpanidis, C. KotropoulosÃ, I. Pitas. "Combining text and link analysis for focused crawling—An application for vertical search engines", 29 September 2006.
18. Soumen Chakrabarti a,Ł,1, Martin van den Berg b,2 , Byron Dom. "Focused crawling: a new approach to topic-specific Web resource discovery".
19. Gautam Pant, Padmini Srinivasan and Filippo Menczer. "Crawling the Web".
20. "Domain-Specific Web Site Identification:The CROSSMARC Focused Web Crawler".
21. "A New Approach To Design a Domain Specific Web Search Crawler Using Multilevel Domain Classifier", pp.476-487.
22. Vladislav Shkapenyuk, Torsten Suel."Design and Implementation of a High-Performance Distributed Web Crawler"
23. Maryam Hazman. "A Survey Of Focused Crawler Approaches" , Vol 3, No.4, JGRCS-April 2012.
24. Joel Acevedo, Giselle Agosto, Maria Ortiz de Zuniga, Professor Morley Mao."Web Crawler" , University of Michigan EECS 489.