International Journal of Pure and Applied Mathematics
Volume 120 No. 5 2018, 1163-1173
ISSN: 1314-3395 (on-line version)
url: http://www.acadpubl.eu/hub/
Special Issue

A Study of Various Semantic Web Crawlers and Semantic Web Mining

*K.Lokeshwaran[1]
Research Scholar, Department of Computer Science and Engineering
Sri ChandrasekharendraSaraswathiViswaMahavidyalaya University, Kanchipuram. India.
Email: k.lokeshwaran@gmail.com,

A. Rajesh[2]
Principal, C. Abdul Hakeem College of Engineering and Technology,Melvisharam, India.
Email: amrajesh73@gmail.com

ABSTRACT

The principle reason for this study is to examine and comprehend the ideas of different semantic web crawlers and about semantic web mining. Numerous pages are persistently being added each day in World Wide Web, and information is updating more frequently. To extract profitable information from the WWW search engines are utilized. The most imperative thing of search engine is to generate quality outcomes and capacity to crawl seamlessly, and index the web proficiently. This paper quickly surveys how the data's are extracted from the semantic web by utilizing crawler and concentrate on the various research areas of semantic web mining.

Keywords: Ontology,Crawling techniques, Search engine, Semantic Web Mining

1. INTRODUCTION

In present scenario, usageof web is growing at a rapid rate. The World Wide Web gives a huge volume of data of all types. Currently people utilize search engines once in a while; substantial volumes of information can be explored effectively through search engines, to extricate important data from web. However in the ocean of web, discerning through all the Web Servers and the pages isn't wise one. Regularlyevery day number of web pages are increasing and nature of data changes every second [4]. Because of the expansive number of pages present on Web, the search engine relies on crawlers for the gathering of required pages [6].

The two quick emerging research areas, Semantic Web Crawlers and semantic Web Mining congregate both on the accomplishment of the World Wide Web (WWW). They supplement each other well since they address one another test postured by the considerable accomplishment

of the present WWW: In WWW most of the information's are unstructured to the point that they must be comprehended by people, however the amount of information is huge so that they must be processed proficiently by machines [18].

The Semantic Web crawler addressesthe initial segment of this challenge by endeavoring to influence the information to machine understandable, while Semantic Web Mining focuses the second part via naturally extracting the valuable knowledge covered up in these data, and making it accessible as an accumulation of sensible extents. This paper is organised into four segments. Section-1 covers introduction of Semantic Web, Section-2 contains outline of the different Semantic web scrawling technique, Section-3 contains presents review of Semantic Web Mining and Section-4 incorporates conclusion and future work while references are appeared in the last segment.

2. SEMANTIC WEB

The Semantic Web is built on the base of World Wide Web, in which for every information semantics and services are defined, making it conceivable to "comprehend" and fulfillthe solicitations of people and machines to utilize the web content.The key establishments of the Semantic Web are ontologies and annotations. Ontologies, which are typically represented using RDF Schema or OWL languages, portray formally shared conceptualizations of an area (e.g., individuals, gatherings, etc.). Annotationsrestrainontology-based descriptions to existing Web assets (e.g., saying that the data contained in website page alludes to a man) and are typically portrayed in RDF. In synopsis, HTML depicts documents and the connections between them, while RDF, RDF Schema and OWL, by opposite, can portray subjective things, for example, individuals, gatherings, and so forth.

In recent times, the idea of the Web of Linked Data has showed up as a system for making all RDF information accessible utilizing the HTTP protocol, as it is finished with HTML documents. This idea has picked up footing after the distribution of huge Semantic Web resources, for example,DBPedia, Bio2RDF, etc., and by the declaration by a few Governments of their choice to make open information open, in an arrangement of Open Government activities (e.g., data.gov, data.gov.uk).

2.1 Semantic Web Crawling

Web crawler is a highly proficient programme that surfs the World Wide Web in a deliberate, mechanized way. The primary goal of web crawler is to fetch the web pages and embed them to local archive. Crawlers are essentially used to make an imitation of all the visited pages that are later handled by asearch engine that will index the downloaded pages that assist in fast search. Searchengines work is to persist information about various web pages. The automated programme which retrievesthese web pages and its links associated with these web pages [2].

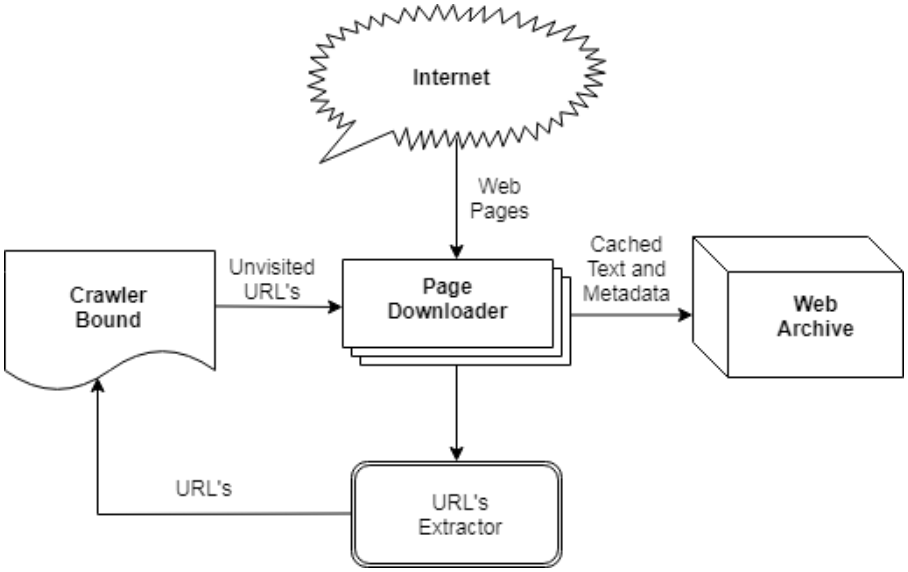

Figure 1: Architecture of Semantic Web Crawler

Building a practical crawler is not an easy task as we like. Due to efficiency and many other concerns, it involves a great deal of engineering. There are two types of crawlers: universal crawlers and topic crawlers[3].A universal crawler downloads all pages irrespective of their contents, while a topic crawler downloads only pages of certain topics. The difficulty in topic crawling is how to recognize such pages. Web crawler is an Internet that systematically browses the World Wide Web, typically for the purpose of Web indexing. It also called as Web spider, an ant, an automatic indexer, Web Scutter.Figure 1 shows the generalized architecture of web crawler.

Looks theoretically simple, constructing a reasonable crawler is by no means simple. Because of productivity and numerous different concerns, constructing crawler needs more time and involves lot of designing. There are two kinds of crawlers: universal crawlers and

topic(keyword) crawlers [3]. Irrespective of the page content a universal crawler downloads all the pages, while a topic crawler downloads just pages related to specific keywords. The trouble in topic crawling is the means by which to perceive such pages. Web crawler is an Internet that efficiently surfs the World Wide Web, primarily with the end goal of Web Indexing. The working of a web crawler is expressed below.

Step 1: *Assigning the starting URL or URLs*
Step 2: *Initializing the bound with starting URL*
Step 3: *Opting the URL from the bound*
Step 4: *retrieving the web-page related to that URL's*
Step 5: *Interpreting the retrieved page to extract the URLs*
Step 6: *appending all the unvisited links to the record of URL i.e. into the bound*
Step 7: *Restart with step 2 and repeat till the bound is vacant.*

Crawler Bound: - It module maintains the list of unvisited URLs. The record is set with initial URLs which might be conveyed by a user or another program [14]. Basically it's simply the accumulation of URLs. The working of the crawler begins with the initialURL. The lists of unvisited URL's present in the bound are crawled along the process. The corresponding page of the URL is retrieved from the Web, and the list of unvisited URL's from the pages is appended to the bound [13]. The process of retrieving and extracting the URL's preceded until the point when the bound is vacant or some other protocols which is used to stop it. To obtain URLs from the bound isbased on some prioritization conspire [8].

Page downloader: - The primary task of the page downloader is to retrieve the pages from the web related to the URL's which is maintained in the crawler Bound. To retrieve the pages, page downloader requires a HTTP client programme to send and receive the HTTP request and responses. HTTP client shouldfix timeout period to avoid unnecessary time delay to read large files from the slow server. Practically, the HTTP client is allowed to download the initial 10KB of a web document. [9].

Web archive: - It is used to persist and organize with alarge volume of data "objects," [23] the data objects refers to web pages. The archiveonly persist a standard HTML pages. Rest of thedocument types are omitted by the crawler [19]. It is hypothetically not that unique in relation to different frameworks that store data objects. It stores the crawled pages as easily

distinguishable files. The up-to-date version of the web pages, which are retrieved by crawler, is stored in this archive.

2.2 Types of Semantic Web Crawler

A. Distributed Crawler

The most challenging task is indexing the web because of its rapid evolving nature. As the span of the Web is developing it has turned out to be basic to parallelize the crawling procedure keeping in mind, the end goal to downloading the pages in a sensible measure of time. A solitary crawling process is lacking for extensive – scale engines that need to bring a lot of data quickly.

If the centralized semantic web crawler is utilized all the information which are retrieved through crawling will goes through a fixed physical connection. Distributing the crawling activity through numerous processes can help to fabricate a versatile, effortlessly configurable system. Unnecessary hardware load can be reduced by dividing it to the various workstations. This increases theoverall download speed and relentless quality. Each task is performed in a completely circulated manner, that is, no focal facilitator exists [15].

B. Focused Crawler

A universally useful Web crawler accumulates more number of pages from a specific set of URL's, where as an focused crawler is intended to just accumulate pages on a specific topic, in this way decreasing the network traffic and download. The objective of the focused crawler is to specifically search out pages that are significant to a pre-characterized set of points.

The topics are determined not utilizing keywords, but rather utilizing praiseworthy reports. Instead of gathering and indexing all open web documents to have the capacity to answer all ad-hoc queries, an focused crawler breaks down its creep limit to discover the links that are probably going to be most important for the crawl, and keeps away from unessential regions of the web.

This prompts huge savings in network resources and hardware. The focused crawler has three fundamental parts: a classifier, which makes significance judgments on pages, crawled to settle on link elaboration, a distiller which decides a proportion of centrality of crawled pages to decide visit priorities, and a crawler with progressively reconfigurable need controls which is represented by the classifier and distiller [10].

C. Incremental Crawler

A conventional crawler, to keep its collection up to date, now and then overrides the old documents with the recently downloaded documents. Despite what might be expected, an incremental crawler step by step invigorates the current collection of pages by visiting them regularly; basedon the gauge with respect to how frequently pages change [19]. It additionally trades less important pages by new and more vital pages. It settles the issue of the freshness of the pages. The advantage of incremental crawler is that only the important information is given to the user, in this way network bandwidth is spared and information enhancement is accomplished [12][11].

D. Parallel Crawler

Various crawlers are regularly kept running in parallel, which are alluded as Parallel crawlers [17]. A parallel crawler comprises of numerous crawling processes [17] referred as C-procsdistributedamong workstations [7]. Page selection and Page freshnessis the primitive factors in which the parallel crawlers relies on [16]. A Parallel crawler can be on nearby system or be conveyed at geologically distant areas [5].Parallelization of crawling framework is extremely indispensable from the perspective of fetching documents in a sensible measure of time [7].

3.SEMANTIC WEB MINING

As the Semantic Web upgrades the original of the WWW with formal semantics, it offers a decent premise to improve Web Mining: The sorts of (hyper) links are portrayed unequivocally, permitting the knowledge engineer to increase further bits of knowledge in Web structure mining; and the contents of the pages represented along with formal semantics, enabling her to apply mining methods which require more structured input.

3.1 Extracting Semantics from the Web

The exertion behind the Semantic Web is to add semantic annotation to Web pages with a specific end goal to get to knowledge rather than unstructured material, enabling information to be overseen in a programmed way.

A. Ontology Learning

Ontology extraction is a challenging task. One route is to design the ontology by hand, yet this is a significant costly. The articulation towards Semantic Web Mining, 269 Ontology Learning was authored for the self-loader extraction of semantics from the Web so as to make ontology [22]. There, machine learning methods were utilized to enhance the ontology building process.

Ontology learning uses a considerable measure of existing resources, similar to dictionary, thesauri etc. It tailors the techniques of a few research area e. g., from machine learning, data recovery, or agents [20], and applies them to find the semantics in the data

B. Mapping and Merging Ontologies

With the growing utilization of ontologies, the issue of covering knowlege in a typical domain happens more frequently and ends up critical. Domain Specific Ontologies are engineered by numerous authors in different settings. These ontologies establish the framework for building new domain specific ontologies in related domains and expanding numerous ontologies from archives. The procedure of ontology consolidating takes input from two (at least two) source ontologies and returns a combined ontology in light of the given source ontologies. In this manner, recently for knowledge engineer many frameworks for merging ontologies have been proposed [21 and 22].

C. Instance Learning

It is likely sensible to anticipate that user will physically comment on new documents to a specific degree, yet this does not take care of the issue of old records containing unstructured material. Regardless we can't anticipate that everybody will physically annotate each delivered mail or document, as this would be outlandish. Besides a few user may need to extract and utilize various or extra data from the one given by the creator. For the reasons specified above it is essential for the Semantic Web to deliver automatic or semi-automatic techniques for extracting data from Web-related reports, either to help in commenting on new records or to extract extra data from existing unstructured or incompletely structured pages.

3.2Semantic Web Content and Structure Mining

In the Semantic Web, content and structure are emphatically bonded. Subsequently, the qualification amongst content and structure mining vanishes. Notwithstanding, the appropriations of the semantic annotations may give extra implicit knowledge.

A vital group of techniques which can without much of a stretch be adjusted to Semantic Web content/Structure Mining are the methodologies examined as Relational Data Mining (once called Inductive Logic Programming (ILP). RelationalData Mining searches for patterns that include numerous relations in relational database.

It includes systems for classification, regression, clustering, and association analysis. It is very clear to change the algorithms so they can manage information depicted in RDF or by ontologies. There are two major logical difficulties in this endeavor. The first is the extent of the information to be handled (i.e., the versatility of the calculations), and the second is the way that the information are appropriated over the Semantic Web, as there is no central database server.

4. CONCLUSION AND FUTURE WORK

This study is concerned about the investigation and analysis of semantic web crawling strategies, procedures, and semantic web Mining and research challenges in the web. In comparison with various crawling methods the Focused Crawling techniquemainly focuses on cutting edge web users concentrate around specific topic and it doesn't waste resources on unimportant material. This research work provides scope for exploring the semantic web in various directions, for example, maintaining semantic web link integrity, exploiting the information based on semantics from the web and learning the domain ontologies, applications of semantic web mining and its features.

REFERENCES

1. Ayar Pranav, Sandip Chauhan, "Efficient Focused Web Crawling Approach for Search Engine", International Journal of Computer Science and Mobile Computing, Vol. 4, No. 5, pp. 545-551, May 2015.

2. Md. Abu Kausar, V. S. Dhaka, Sanjeev Kumar Singh, *"Web Crawler: A Review",* International Journal of Computer Applications (0975 – 8887),Volume 63– No.2, February 2013.

3. DhirajKhurana, SatishKumar., Web Crawler: A Review IJCSMS International Journal of Computer Science & Management Studies, Vol. 12, Issue 01, January 2012.

4. Bharat Bhushan, Narender Kumar, " Intelligent Crawling On Open Web for Business Prospects", IJCSNS International Journal of Computer Science and Network Security, VOL.12 No.6, June 2012

5. Pavalam S. M., S. V. Kasmir Raja, Jawahar M., and Felix K. Akorli, *"Web Crawler in Mobile Systems",* International Journal of Machine Learning and Computing, Vol. 2, No. 4, August 2012

6. Keerthi S. Shetty, SwarajBhat and Sanjay Singh, *"Symbolic Verification of Web Crawler Functionality and Its Properties",* International Conference on Computer

Communication and Informatics (ICCCI -2012), Coimbatore, INDIA, IEEE Conference Publications,2012

7. Shruti Sharma, A.K.Sharma, J.P.Gupta, *"A Novel Architecture of a Parallel Web Crawler",* International Journal of Computer Applications (0975 – 8887) Volume 14– No.4, January 2011.

8. IoannisAvraam, IoannisAnagnostopoulos*, "A Comparison over Focused Web Crawling Strategies"*, Panhellenic Conference on Informatics, IEEE Conference Publications, 2011.

9. Wenxian Wang, Xingshu Chen, YongbinZou, Haizhou Wang, Zongkun Dai, "A Focused Crawler Based on Naive Bayes Classifier", Third International Symposium on Intelligent Information Technology and Security Informatics, IEEE Conference Publications,2010.

10. Christopher Olston and Marc Najork., Web Crawling, Foundations and Trends in Information Retrieval Vol. 4, No. 3 175–246, 2010.

11. NirajSinghal, Ashutosh Dixit, and Dr. A. K. Sharma, *"Design of a Priority Based Frequency Regulated Incremental Crawler",* International Journal of Computer Applications (0975 – 8887) vol.1, no. 1, pp. 42-47, 2010.

12. A. K. Sharma and Ashutosh Dixit*, "Self Adjusting Refresh Time Based Architecture for Incremental Web Crawler"* International Journal of Computer Science and Network Security, vol.8 no.12, pp. 349-354 ,2008.

13. Gautam Pant, Padmini Srinivasan, *"Learning to Crawl: Comparing Classification Schemes",* ACM Transactions on Information Systems, Vol. 23, No. 4, Pages 430–462, October 2005.

14. Pant Gautam, Srinivasan Padmini, MenczerFilippo, *"Crawling the Web"* In Levene, Mark; Poulovassilis, Alexandra. Web Dynamics: Adapting to Change in Content, Size, Topology and Use. Springer. pp. 153-178. 2004

15. Baldi, Pierre, Modeling the Internet and the Web: Probabilistic Methods and Algorithms, willey Publications, pp-158, 2003.

16. AH Chung Tsol, Daniele Forsali, Marco Gori, Markus Hagenbuchner, Franco Scarselli, *"A Simple Focused Crawler"* Proceeding 12th International WWW Conference, pp. 1, 2003.

17. Junghoo Cho, Hector Garcia-Molina, *"Parallel Crawlers", WWW2002, Honolulu, Hawaii, USA, 2002.*

18. Hans Chalupsky., Ontomorph: A translation system for symbolic knowledge. In Principles of Knowledge Representation and Reasoning: Proceedings of the Seventh International Conference (KR2000), pages 471–482, 2000.

19. Junghoo Cho and Hector Garcia-Molina.*"The evolution of the web and implications for an incremental crawler"*, In Proceedings of the 26th International Conference on Very Large Databases, 2000.

20. Williams A Band Tsatsoulis C., An instance-based approach for identifying candidate ontology relations within a multi-agent system, In Proceedings of the First Workshop on Ontology Learning OL'2000, Berlin, Germany, Fourteenth European Conference on Artificial Intelligence, 2000.

21. Hans Chalupsky., Ontomorph: A translation system for symbolic knowledge. In Principles of Knowledge Representation and Reasoning: Proceedings of the Seventh International Conference (KR2000), pages 471–482, 2000.

22. Hovy E.H., Combining and standardizing large-scale, practical ontologies for machine translation and other uses, In Proc. 1st Intl. Conf. on Language Resources and Evaluation (LREC), Granada, 1998.

23. Jun Hirai SriramRaghavan Hector Garcia-Molina Andreas Paepcke, *"WebBase : A repository of web pages"* , available: http://ilpubs.stanford.edu:8090/380/1/1999-26.pdf