Volume 5, Issue 5, May 2015 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Web Crawlers for Semantic Web

Akshaya Kubba
Computer Science Department
Dronacharya Government College,
Gurgaon, Haryana, India

Abstract: Web mining is an important concept of data mining that works on both structured and unstructured data. Search engine initiates a search by starting a crawler to search the World Wide Web (WWW) for documents .Web crawler works in a ordered way to mine the data from the huge repository. The data on which the crawlers were working was written in HTML tags, that data lags the meaning. It was a technique of text mapping.
Semantic web is not a normal text written in HTML tags that are mapped to the search result, these are written in Resource description language. The meta tags associated with the text are extracted and the meaning of content is find.
The data on the web is increasing day by day and the web pages are no more static in nature instead have become dynamic so it is important for a crawler to revisit the page and find for the updated information and give us the efficient result in no time. The primary goal of the research is to see how the crawlers work on semantic weband how optimized they display the results. This research is aimed at explaining how the web crawlers work on semantic data.

Keywords- Semantic web, crawlers, web crawlers, ontology, Swoogle, Wrappers.

I. INTRODUCTION

Syntax and Semantics the two common words in any vocabulary, what does they really mean? Syntax of a language means the proper format of writing it e.g. For English vocabulary the syntax is Subject, verb, object.Semantics of a language refers to the meaning of the sentence. *Web mining*is a part of data mining that extracts the structure and content of the web page, *Semantic mining* on the other hand finds the meaning of the data and extract the relevant web page. Web mining can work on Unstructured and Structured data while Semantic mining works only on structured data.We exploit the semantics of the web content and apply some ontology heuristics. This concept came into existence by *Berners-lee* to enrich the web with machine understandable data for which we need a knowledge base called ontology.

When we talk about search engine how does it search the desired result for us, the main focus goes to the crawler or the spider that crawls from one web page to another web page and so on until and unless it finds the desired result.

Now how does the search engine understand what are we looking for? An agent works for us. It takes the requirements from us and then scratches the repository and gives us with the desired result.E.g.suppose we want to buy a laptop of a particular configuration from a shop which is in the radius of our home about 4km and should have a warranty of 2yrs. So you just need to set up your agent to do all these work for you and provide it with your requirements, it will quickly response to you with the results.Search Engine date backs to 1990's and a lot of research has been done on crawlers and searching but here in this paper I am trying to explain the new era of search that has came into existence in recent years" semantic web" ,The working of crawlers on the Semantic web.

This paper is organized into different sections the section I of this paper gives the Introduction; section II describes the Web Crawlers; section III describes Semantic Technology; section IV gives conclusion.

II. WEB CRAWLERS

A web crawler is a computer program written with the intent of finding a result on the web.Web crawler has many synonyms: spiders scutter or ants etc. It is nothing more than a script written to browse the WWW (World Wide Web). There is a large pool of unstructured information in the form of Hypertext documents which is difficult to be accessed manually, hence there is a need for a good crawler to work efficiently for us.

The basic algorithm of a crawler is:

```
{
Start with seed page
        {
        Create a parse tree
        Extract all URL's of the seed tree (front links)
        Create a queue of extracted URL's
                {
```


Fetch the URL"s from the queue and repeat
}
}
}Terminate with success

There are different types of web crawlers available:

i) *Restricted webCrawler*: This crawler is restricted to a certain domain. It crawls just some specified portion of the web page and access too many links is restricted.
ii) *Path ascending crawler*: It extract all the links from a particular website and then crawls all those URL's and them again in the same way it keeps and crawling the consecutive pages that follows the web page and spans the entire web. It also follows the revisit policy to check the expiration of content.
iii) *Focused crawler:* It allows crawling web pages in accordance with the given query. So the search is focused to some portion of the web and irrelevant web pages are not crawled.

A. Searching the web page

The first step in working of the crawler is searching the desired document. When we input a query to the search space the crawlers start their action, they crawls through millions of web pages to find the desired result. It starts from a root pageor seed page and then extract the URL'sfrom that seed page and then keep it in a queue and then keep on following the external links until the desired threshold is reached or some other stopping criteria is reached. It then keeps the data into a local repository.

Local repository keeps all the crawled pageswhich are later re-visited by the crawler later to check for any updates.

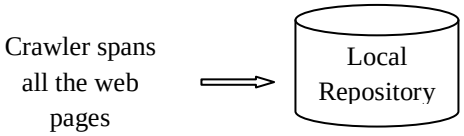

Figure 1: Searching by crawler

Once a page is fetched it is thenkept into a local repository and itsURL is normalized. This avoidsrevisiting to the same page. *Normalization* of URL's is done in many ways like the URL is converted to lower case or the initials of the URL's are in upper case.The crawler has to crawl many local servers with the same algorithm hence it is really difficult as it has to deal with many problems. URL Normalization is important step in Optimization so as web search engines only pick the meaningful page. Normalization or*Canonicalization* is also a major issue as search as many website points to same link, so to select the best among them is required to be done. This is done according to the bro There are many normalization that preserves the semantics of the web page and some normalization disturbs the semantics of web, it is important to select the best normalization technique.

B. Page Rank and revisiting

A search engine searches a page by its page rank. i.e. by its value in the WWW. Page rank means "how much worth a web page is". It uses some page rank algorithm for it other than just text matching the query. Page rank may be assigned by the frequency of time the search term appears on the web page or the no. of sites that link to the web page. The more optimized a web page is the better is its page rank.

Page rank of a page is calculated as

$$PR(x) = (1-d) + d(\frac{PR(t1)}{C(t1)} + \cdots \ldots \frac{PR(tn)}{C(tn)}) \qquad (1)$$

Where d is damping factor whose value is taken between 0-1. t1….tn are the no of links to our page and C(tn) is the no. of links going out of the page. So a page having better page rank is fetched by the crawler and displayed.

Also another factor called P value determines the importance of the page it is calculated as

$$PV = No.of\ F\ links - No.of\ B\ links \qquad (2)$$

Where PV is the page Value and F links are the Forward links and B links are the backward links. The highest PV is given the priority , a page can even have 0 PV. i.e. least priority.

Again some points that the crawler has to deal with is as the web is very dynamic in nature so it is required to **re-visit** the page and check for updating. So some of the crawlers cost is included to re-visit the page. This comprise the age of the page and the freshness of the page. The page should not be outdated and should have maximum freshness .This revisit policy is used in incremental crawler**.**There can be two types of changes: "*change inthe structure*" of the page or" *change in the content*" or "*Change Image"*

1) *"Change in structure*"refers to how the HTML document is structured i.e. how texts and images are formatted on a web page. This is done with the help of HTML tags. An algorithm would run that would store the positions of the tags and the characters associated with it. So crawlers would compare the previous tags location and if the page is updated it will updated the index page at the time of re visit.
2) *"Change in content"* As the crawlers scans the document for the first time it assigns a code to the text content of the web page so at the time of re visit it compares the text code of the document. It finds the ASCII count of the page and the total character count.

3) *"Change Image" Each* image is assigned an I value so whenever the revisit is done the I value of the web page is compared for any update.

C. Handling overhead

The main problem with the crawler is to handle the large directory for which it has to run n number of servers at a time, there can also be multiple crawlers working in parallel. As the web is Dynamic and its size has increased to a huge extend so there is need to run the crawlers in parallel to give the desired result in a definite time. Running the crawlers in parallel can ease their work as it does two things *first*it increases the network scalability and *second* network load is dispersed at different geographic location. After the crawlers collect the data it goes to a central location in compressed form.Crawler can crawl 100 pages in 1 second using 4 crawlers.

Parallelization of crawler also has some disadvantages *first* overhead is too high,they can even harm the network traffic and crash the web servers, it is mandatory that crawler should be written good to handle the load*second* and they must communicate with each other periodically. This is done so they know that each one is downloading a different page to minimize the overlap.

III. SEMANTIC TECHNOLOGY

In recent years a new technique of search semantic technology has came into existence.This technology basically indents to find the" meaning "of the content rather than the text mapping techniques. So it is more efficient and relevant method of searching. It helps the machine to better understand the content and hence benefits the user to give a better result output.

Swoogle[1] is a search engine for semantic web that index the semantic content it's first and second version was built in 2004 and third version in 2005[1]

Google new algorithm"humming bird" came in September 2013 for semantic web. The word humming bird actually means "precise and fast" does conversational search [6].

It does not only see the keywords or the Meta tags rather it searches in a conversational way. For e.g. suppose I want to find a shoe house that is open on Sunday till 9 p.m and has all good brands available, the search engine will read my location and read the entire sentence in the query and provide me the results.

It actually gives you precise and accurate results than the previous search engines. We get the result of the query only not any unnecessary information. For e.g. if you search for a "CBSE results class X" it will give you the direct link for class X results of CBSE website.

Some other semantic web search engines include "Powerset" for natural language search.

A. Semantic wrappers

Wrappers are basically meant to extract XML documents from the HTML documents. They basically convert the unstructured data to a more structured form. So it is important to write a wrapper but there are some draw bags to it.

Drawbags of writing a wrapper:

i) As all the information on web is written in HTML tags and they lack semantic content so it is difficult for wrapper to understand the content of interest.
ii) Manually written wrappers are robust.
iii) As the web is dynamic so it is difficult for a wrapper to work on heterogeneous data.

The web content is changing day by day so it is difficult to write a wrapper manually for the non homogeneous data. Only a small part of the wrapper code refers to the content rest most of the code of wrapper is same for all data. Wrappers are able to integrate information from different web pages to work efficiently.

So a semi automatic wrapper construction called XWRAP was created. It works by extracting the Meta tags of the HTML document and convert them into XML tags in wrapped text. It does two functions firstly it generates information extracting rules and then secondly using that information it constructs a wrapper program.

B. Semantic Vocabulary

Semantic Vocabularyalso called Ontology is the base of Semantic Technology. It determines the instance- of or relationship between the entities. Ontology's are basically the vocabularies of the semantic web with some difference Ontology refers to what kinds of things exist and what relationships do they have , on the other hand Vocabulary is just a mere list of words. Ontology is defined by the scope of the data or the class and the relationships between the classes. Ontology is the basic building blocks of semantic web. Ontology's basically help in a decision support system e.g. suppose we have an ontology related to health concern. Doctors and other medical professionals would use this ontology to inform about the diseases and symptoms while for a pharmaceuticals that ontology provide us the list of drugs and dosage. So if in all this ontology is integrated it can help a patient to treat an ailment, hence it supports Decision Support system.Ontology basically helps to link the data. The vocabularies created for semantic web are not in HTML format in fact they follow RDF (Resource Description Framework) schema or OWL (Web Ontology Language).

Ontology building steps follows:

i) *Know the scope and Domain of ontology*: Scope explains for whom we are making ontology for and what all will it cover, what queries it will answer
ii) *Enumerate the important terms* : it defines all the important terms or classes used in the ontology building

iii) *Define the class hierarchy used*: Which hierarchy are we using top down or bottom up to create classes and subclasses

IV. CONCLUSION

This paper describesSemantic Web mining which is nothing more than knowledge extraction. We create knowledge from the unstructured database which is understood by machines. This technologyhas increased the popularity of search engine and this knowledge base machine learning method is fruitful to the users. We have explained how the web Crawlers workson Semantic data and how the ranking of the page is calculated to index the web page.

REFERENCES

[1] Dhiraj Khurana, Satish Kumar, "Web Crawler: A Review"IJCSMS International Journal of Computer Science &Management Studies, Vol. 12, Issue 01, January 2012ISSN (Online): 2231 –5268

[2] M. Hersovici, M. Jacovi, Y. S. Maarek, D. Pelleg, M. Shtalhaim, and S. Ur. " The sharksearchalgorithm—An application: TailoredWeb site mapping", In *Proceedings of the 7th International World Wide Web Conference*, Brisbane, Australia, 1998.

[3] S. Chakrabarti, M. van den Berg, and B. Dom.," Focused crawling:Anewapproach to topic specific Web resource discovery",*Computer Networks*, 31(11–16):1623–1640, 1999.

[4] S.S Dhenakaran and K. Thirugnana Sambanthan, "Web Crawler - An Overview", International Journal of Computer Science and Communication Vol. 2, No. 1, January-June 2011, pp. 265-267

[5] Swati Mali, B.B. Meshram "Focused Web Crawler with Page Change Detection Policy", 2nd International Conference and workshop on Emerging Trends in Technology (ICWET) 2011.

[6] Maedche., "Ontology Learning for the Semantic Web,Kluwer" , 2002.

[7] Maedche and Staab S., "Ontology learning for the semanticweb". IEEE Intelligent Systems, 16(2):72 –79, 2001

[8] S.Balan, Dr.P.Ponmuthuramalingam, "A Study on Semantic Web Mining And Web Crawler" , International Journal Of Engineering And Computer Science ISSN:2319-7242, Volume 2 ,Issue 9 Sept., 2013 Page No. 2659-266

[9] http://searchengineland.com/google-hummingbird-172816

[10] http://swoogle.umbc.edu/index.php?option=com_swoogle_manual&manual=introduction

[11] http://computer.howstuffworks.com/internet/basics/search-engine4.htm

