ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 3, March 2015.

A Keyword Focused Web Crawler Using Domain Engineering and Ontology

Gunjan Agre[1], Snehlata Dongre[2]

Student, Department of Computer Science and Engineering, G.H.R.C.E College, Nagpur, Maharashtra[1]

Assistant Professor, Department of Computer Science and Engineering, G.H.R.C.E College, Nagpur, Maharashtra[2]

Abstract: As the number of users on internet grows the number of accessible web page also grows which causes more troublesome for users to find relevant or specific data according to their needs. Web crawler is that the method utilized by search engines to collect pages from the net. The necessity of an online crawler that downloads most relevant web content from such an oversized internet remains a serious challenge within the field of Information Retrieval Systems. Most internet crawlers use keyword base approach for retrieving the knowledge from Web. However they retrieve several irrelevant web contents as well. With the utilization of linguistics additional relevant pages can be downloaded. Linguistics will be provided by ontology. This paper proposed algorithm on ontology based internet crawler specified such that only relevant sites can be retrieved and estimate best path for crawling which uses for improving the crawling performance.

Keywords: Web Crawler, Focused web crawler, Importance-metrics, Ontology, domain knowledge.

I. INTRODUCTION

The World Wide net (WWW) having billion websites and looking documents that is additional specific with the user's needs is progressively tough. The World Wide Web supports dynamic content that is growing progressively as well as news, current problems, new technology, business info, finance, marketing, recreation, education become cosmopolitan over a large space of net.

The web crawler largely downloads solely the relevant or specific websites in keeping with the user needs instead of downloading all websites sort of ancient search engines. Therefore the basic goal of focused crawler is to pick out and hunt down the net pages that fulfil user's demand. The link analysis algorithmic programs like page ranking algorithm and different metrics area unit use to range the URLs supported their ranking and choice policies for downloading most specific websites.

In this paper, the keyword focused web crawler has been projected. The keyword focused web crawler algorithmic program seeks out the URLs of websites supported their priority and domain ontology. Additionally the information path plays vital role to find relevant websites.

The web crawler is the software program which act as a main component of search engine. Crawler is additionally known as spider or a computer code agent.

In general web crawler starts its operating victimization seed address that act as associate initial address for creep method. Once visiting to the net page of seed address it transfer that online page and so extracts all the hyperlinks present therein downloaded online page and stores all that links to the queue that is additionally known as frontier and recursively repeat the procedure till it gets the relevant results.

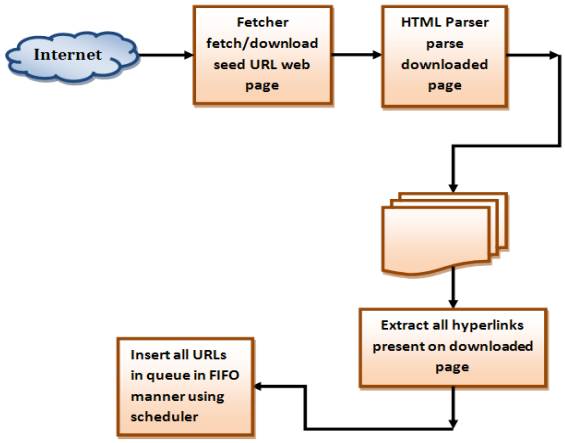

Figure1 .Architecture of Simple web crawler

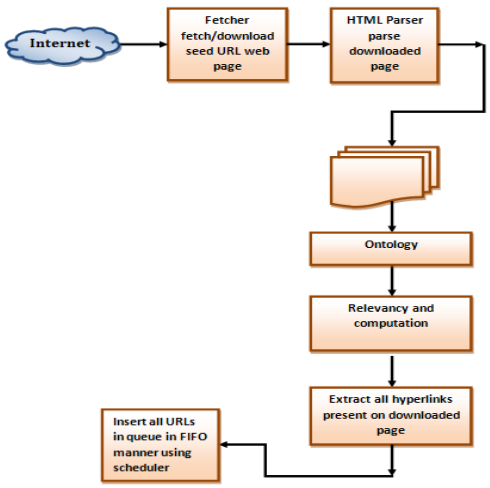

Figure2 .Architecture of web Crawler using ontology

DOI 10.17148/IJARCCE.2015.43111

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 3, March 2015.

The main intension of web crawler is to download only important pages from web and to visit the important pages according to their priorities it placed in queue (frontier).

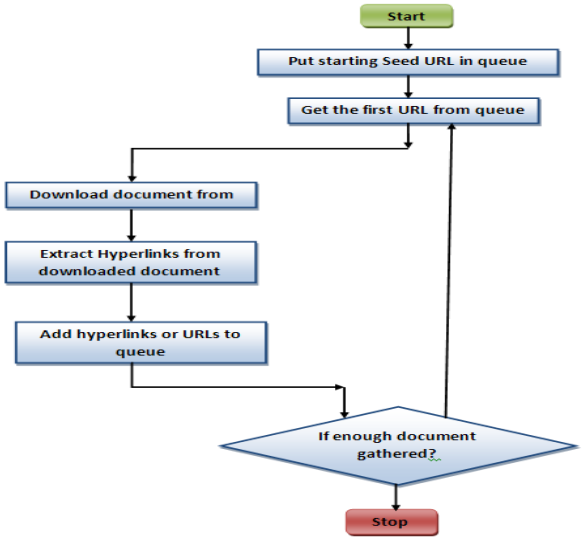

Figure 3.Implementation of Frontier (Queue) in Keyword Focused Web Crawler.

The main aim of this paper is to deals with the domain ontology and knowledge path to finds out the most relevant web pages according to the user requirements.

Section 1 deals introduction to domain engineering, robot.txt file and crawling policies.

Section 2 includes the methodology containing algorithm, flowchart, precision calculation formula.

Section 3 includes result and section 4 includes conclusion.

A. Domain engineering

A Domain engineering primarily based search has nice potential to enhance structuring and looking out in element libraries [1]. Ontology will be used for structuring and Filtering the knowledge repository (in our case, a universal resource locator Queue i. e. Frontier). Ontology is that the method of Domain data illustration [2] and those we will use ontology engineering for filtering the URLs from the Frontier. it should be used as a abstract framework to developers.

In domain engineering, ontology will play many roles. In step with Uschold, "ontology could take a spread of forms, however essentially it'll embody a vocabulary of terms, and a few specification of that means. This includes definitions and a sign of however ideas area unit inter-related that together impose a structure on the domain and constrain the doable interpretations of terms"[3]. Thus, associate ontology consists of ideas and relations, and their definitions, properties and constrains expressed as axioms. Associate ontology isn't solely associate hierarchy of terms, however a completely axiomatized theory regarding the domain

B. Robot.txt

It is great when search engines frequently visit your site and index your content but often there are cases when indexing parts of your online content is not what you want. For instance, if you have two versions of a page (one for viewing in the browser and one for printing), you'd rather have the printing version excluded from crawling, otherwise you risk being imposed a duplicate content penalty. Also, if you happen to have sensitive data on your site that you do not want the world to see, you will also prefer that search engines do not index these pages (although in this case the only sure way for not indexing sensitive data is to keep it offline on a separate machine).

One way to tell search engines which files and folders on your Web site to avoid is with the use of the Robots Meta tag. But since not all search engines read Meta tags, the Robots Meta tag can simply go unnoticed. A better way to inform search engines about your will is to use a robots.txt file.

Robots.txt is a text (not html) file you put on your site to tell search robots which pages you would like them not to visit. Robots.txt is by no means mandatory for search engines but generally search engines obey what they are asked not to do. It is important to clarify that robots.txt is not a way from preventing search engines from crawling your site (i.e. it is not a firewall, or a kind of password protection) and the fact that you put a robots.txt file is something like putting a note "Please, do not enter" on an unlocked door – e.g. you cannot prevent thieves from coming in but the good guys will not open to door and enter. That is why we say that if you have really sensitive data, it is too naïve to rely on robots.txt to protect it from being indexed and displayed in search results.

The location of robots.txt is very important. It must be in the main directory because otherwise user agents (search engines) will not be able to find it – they do not search the whole site for a file named robots.txt. Instead, they look first in the main directory (i.e. http://mydomain.com/robots.txt) and if they don't find it there, they simply assume that this site does not have a robots.txt file and therefore they index everything they find along the way.
The structure of a robots.txt is pretty simple (and barely flexible) – it is an endless list of user agents and disallowed files and directories. Basically, the syntax is as follows:

User-agent:
Disallow:
"*User-agent"* is search engines' crawlers and *disallows:* lists the files and directories to be excluded from indexing. In addition to "user-agent:" and "disallow:" entries, you can include comment lines – just put the # sign at the beginning of the line:
All user agents are disallowed to see the /temp directory.
User-agent: *
Disallow: /temp/

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 3, March 2015.

C. Crawling Policies:

Now days the size of web is increasing vastly and information changing in high range. The output and the behaviour of web crawler is depend upon different policies as.

- Selection policy
- Re-visit policy
- Politeness policy
- Parallelization policy

Our work is only with the selection policy.

In selection policy Crawler downloads net pages within a fraction from principally gettable data that contain mostly relevant online page it cannot downloads all pages from net. The importance of online page is relying upon its quality in terms of links or visits. arising with associate degree honest alternative policy is hard if the whole set of online page is not known throughout travel.

II. METHODOLOGY

The basic algorithm (formula) dead by any web crawler take a list of seed URLs as its input and repeatedly execute the following steps. Exclude an address from the address list, verify the science address of its host name, transfer the corresponding document, and extract any links contained in it. for each of the extracted links, certify it's Associate absolute address, and add it to the list of URLs to transfer, provided it is not been encountered before.

A. Algorithm steps

Step 1: Take input as a seed URL from which the crawling process starts.

Step 2: Create ontology tree and then find out the knowledge path.

Step 3: Downloads all the URL's that are associated with input URL.

Step 4: Extract all links present in downloaded web page and insert into URL frontier.

Step 5: To find more relevant URL, downloads page associate with this URL and extract all links present on that downloaded pages and insert URL links as a new URL into frontier.

Step 6: Repeat these steps until to get more relevant result.

B. Flowchart

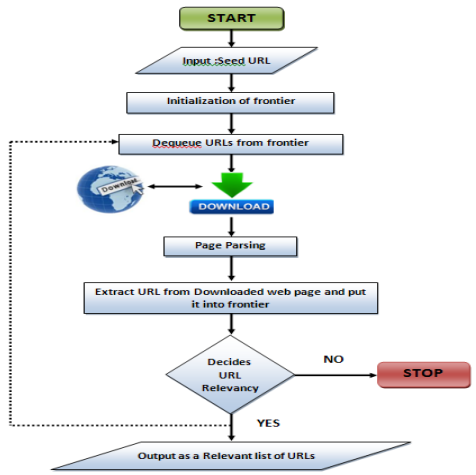

Figure 4 .Flowchart for keyword focused web crawler.

C. Calculation of precision:

The precision is calculated as:

Total number of relevant pages extracted / Total number of web pages extracted =0

An easy way to comply with the conference paper formatting requirements is to use this document as a template and simply type your text into it.

III. RESULT

TABLE I
COMPARISON OF TWO CRAWLERS

Type	Comparison of two crawlers	
	Traditional web crawler	**Keyword focused web crawler**
Total number of extracted links	740	360
Relevant number of links	500	160
Crawling Time	600 sec	200sec

IV. CONCLUSION

The main advantage of keyword focused web crawler over the available traditional web Crawlers is that it doesn't want any connectedness feedback or training for internal details that how the processing is going on (coaching procedure) so as to act intelligently.

Two types of amendment were found when examination results of each the crawlers:

I) the amount of extracted documents was reduced. Link analyzed, and deleted a good deal of irrelevant websites.

II) Turnaround time for crawling process is reduced. When a good deal of irrelevant website is deleted, crawl load is reduced.

REFERENCE

[1] Debajyoti Mukhopadhyay, Arup Biswas and Sukanta Sinha , "A New Approach to Design Domain Specific Ontology Based Web Crawler",Procedings of 10th International Conference on Information Technology, 2007.

[2] Markus Hagenbuchner ,Milly Kc, and Ah Chung Tsoi," Quality Information Retrieval for the WorldWideWeb" proceedings of International Conference on Web Intelligence and Intelligent Agent Technology IEEE/WIC/ACM in 2008.

[3] Alexandre Alvaro1, Vinicius Eduardo Santana de Almeida1, Cardoso Garcia1, Daniel Lucredio2,Silvio Romero de Lemos Meira1, "An Experimental Study in Domain Engineering" in proceedings of 33rd EUROMICRO Conference on Software Engineering and Advanced Applications ,SEAA in 2007.

[4] Arup Biswas, Sukanta and Debajyoti, "A New Approach to Design Domain Specific Ontology Based Web Crawler", 10th International Conference on Information Technology in 2007 IEEE.

[5] Ganesh, S; Jayaraj M, Aghila G "Ontology Based Web Crawler" Information Technology; Coding &Computing, 2004 volume 2,2004 IEEE.

[6] Rosella "An information guided spidering: A domainspecific case study"-2007.