www.ijecs.in
International Journal Of Engineering And Computer Science ISSN:2319-7242
Volume 4 Issue 7 July 2015, Page No. 13184-13186

Multidomain Optimised Hidden Web Crawler

Khushboo Gupta[1], Ankush Goyal[2]

[1]Student, Shri Ram College of engineering and Management, CSE Department,
Palwal
g.Khushbu1491@gmail.com

[2]Assit. Prof. Shri Ram College of engineering and Management, CSE Department,
Palwal
Ankush4989@gmail.com

Abstract *:- Although hidden web crawlers seems to be in great use while searching for hidden data, but the curent hidden web crawlers works on only one domain which proves to be a major drawback in the compatibilty of the searching process. This paper highlights the extended working of the hidden crawlers by providing multiple domain feature which not only helps the user to get high quality content but also makes searching process faster and more broader. The multiple domain system can work on two or more than two domains depending upon the requirement. The multiple domain system also provides a fast linkage between various webpages using multiple links. Thus the project proves to be an useful extension of this extensive tool.*

Keywords: URL, WWW, HiWE, Multidomain

TRADITIONAL SEARCH ENGINE:

Search engine is a program that searches information for specified keywords enterd by users and returns a list of the documents as a result from where the keywords are matched. The term is frequently used to precisely describe systems like Google, Yahoo and HotBot that allows users to search for information on the WWW.

Search engines are motorized systems that methodically separate the web for servers and web pages. As the result page is found, the system reads the words from the web pages and stores it to its database for later retrieval when asked by a user. The search engine purely checks its database of words found by the ‘“bot” on the web and returns with a set of URLs having those words to the user. This process assumed to be good until we realize that search engines are having some limitations. Some of these limitations are as follows:

1.Search engines more propabally use to index sites which have more links to them or we can say the sites which are more popular.

2.Searching websites which more likely index commercial sites than educatiional sites.

3. Indexing of new or modified pages by just one of the major search engines that takes more time.

4. Search engines are designed to read flat web pages

And even if the search engine crawlers could get in the databases used by dynamically generated web sites, some dynamically made web pages have variable URLs and some of them have same URL for all queries input by users. Thus, a search engine can not rely on the URL found to be precise on the next phase of search. Thus it is not possible for traditional search engines to index the Hidden web pages using traditional approach and hence there is a need to build the “Hidden Web Search Engine”.

DESIGNING OF HIDDEN WEB CRAWLER:

The whole process of designing a Hidden Web Crawler is divided into two phases. These are:

1.Crawling process
2.Searching process

Now to explain these two processes we have taken example of car and book domain.

Crawling Process:

Crawling process starts by collecting the web pages for particular domain. Since this work is based on used car domain and book domain, "used car" keyword or "book name" is pushed into the search box of Google and result index is extracted. Now, out of this large set of results, sampling process has randomly chosen four websites. These sites are downloaded to extract the search interfaces. The information collected in crawling process will be stored in database in the form of templates

Algorithm for Crawling process

Step 1. Domain Selection after preprocessing which includes
Tokenization
Stopword Removal
Stemming
Step 2. Retreive URL's based on domain and query input
Step 3. Extract template from URL by exploring its Source Code
Step 4. Store the extracted templates in database

SEARCHING PROCESS

In searching process when user will input the query, the keyword entered will be used to extract the result after finding the domain to which it belongs. Wordnet dictionary will be used for finding proper meaning of keyword entered by user. After entering the keyword the URLs got from searching process which are stored in database will be retrieved and will be displayed as result. User can access any of the URL from set of URLs displayed. As this system is providing feature of two domain at the time of entering query one need to select the domain to which the query will be entered. Like in this project there are two domain one is for CAR and another is for BOOK so if a person wants to purchase a used car he/she need to select Car domain and if a person wants to purchase any book than he/she need to select Book domain.

Algorithm for Searching :-

Step 1. Input Query and select the domain
Step 2. Fill provided data in template URLs of selected domain
Step 3. Resulted template Url will navigate us to direct Result page

MULTI-DOMAIN SYSTEM

Although hidden web crawlers seems to be in great use while searching for hidden data, but the curent hidden web crawlers works on only one domain which proves to be a major drawback in the compatibilty of the searching process. This project extends the working of the hidden crawlers by providing multiple domains which not only helps the user to get high quality content but also makes searching process faster and more broader. The multiple domain system can work on two or more than two domains depending upon the requirement. The multiple domain system also provides a fast linkage between various webpages using multiple links. Thus the project proves to be an useful extension of this extensive tool.

Presently Hidden web crawler is tested for two domains that is used cars and book. So it will be look like shown below figure.

Now Suppose a person wants to purchase a used car then he/she will select car domain. After selecting the car domain 3 fields will be shown that is Make, Model and City. After filling these fields user will get the set of URLs and the user can select any one from it.Likewise same process will be followed by book domain also.

Advantages of Multidomain Hidden Web Crawlers

1.User can search for more than two domains in short span of time.
2.It provides flexibility to the users.
3.Provides hidden data from web for particular domain at a time.

CONCLUSION:

Hidden Web data integration is a major challenge nowadays. Because of autonomous and heterogeneous nature of hidden web content, traditional search engine has now become an ineffective way to search this kind of data. They can neither integrate the data nor they can query the hidden web sites. Hidden Web data needs syntactic and semantic matching to achieve fully automatic integration.

In this work, fully automatic and domain dependent prototype system is proposed that extracts and integrates the data lying behind the search forms along with the feature of multi domain access.

FUTURE SCOPE:

In this research work, various challenges in the area of Hidden web data extraction and their possible solutions have been discussed. Although this system extracts, collects and integrates the data from various hidden websites successfully, this work could be extended in near future. In this work, a search engine shell has been created which was tested on a two domain. This work could be extended for other domains by integrating this work with the unified search interface so to get more enhanced system performance.

REFERENCES

1. http://www.voelspriet.nl/documenten/invisibleweb
2. http://www.w3.org/WWW.
3. www.abebooks.com/InlineOnline-Fundamentals-Internet-World-Wide.
4. http://searchenginewatch.com/article/2065173/How-Search-Engines-Work.
5. http://www.webopedia.com/DidYouKnow/Internet/2003/HowWebSearchEngines