ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 8, August 2016

Extracting the Web Data through Deep Web Interfaces

Namish A. Diwate[1;] Kanchan Varpe[2]

M.E., Dept. of Computer, RMD Sinhgad School of Engineering, Pune, India[1]

Assistant Professor, Dept. of Computer, RMD Sinhgad School of Engineering, Pune, India[2]

ABSTRACT: As deep web develops at a quick pace, there has been expanded enthusiasm for methods that assistance effectively find deep web interfaces. Nonetheless, because of the vast volume of web assets and the dynamic way of deep web, accomplishing wide scope and high proficiency is a testing issue. We propose a Two Stage Crawler, for productive collecting deep web interfaces. In the first stage, Two Stage Crawler performs site-based searching for center pages with the help of search engines, avoiding visiting a large number of pages. To obtain more accurate results for a focused crawl, Two Stage Crawler ranks websites to prioritize highly relevant ones for a given topic. In the second stage, it achieves fast in-site searching by finding most relevant links with an adaptive link-ranking process. To eliminate bias on visiting some highly relevant links in hidden web directories, we design a link tree data structure to achieve wider coverage for a website. The experimental results shows us that the harvest rates of Two Stage Crawler are better than the existing crawlers and efficiently retrieves the data from deep web.

KEYWORDS: Two Stage Crawler , Ranking, Reverse Searching, Adaptive Learning, Deep Web.

I. INTRODUCTION

A Web Crawler (also known as a robot or a spider) is a system for the bulk downloading of web pages. Web crawlers are used for a variety of purposes. Most prominently, they are one of the main components of web search engines, systems that assemble a corpus of web pages, index them, and allow users to issue queries against the index and find the web pages that match the queries. A related use is web archiving (a service provided by e.g., the Internet archive [3]), where large sets of web pages are periodically collected and archived for posterity. A third use is web data mining, where web pages are analyzed for statistical properties, or where data analytics is performed on them (an example would be Attributor [5], a company that monitors the web for copyright and trademark infringements). Finally, web monitoring services allow their clients to submit standing queries, or triggers, and they continuously crawl the web and notify clients of pages that match those queries. The deep (or hidden) web refers to the contents lie behind searchable web interfaces that cannot be indexed by search engines. Based on extrapolations from a study done at University of California, Berkeley, it is estimated that the deep web contains approximately 91,850 terabytes and the surface web is only about 167 terabytes in 2003, [1]. More recent studies estimated that 1.9 zettabytes were reached and 0.3 zettabytes were consumed worldwide in 2007, [2], [3]. An IDC report estimates that the total of all digital data created, replicated, and consumed will reach 6 zettabytes in 2014 [4]. A significant portion of this huge amount of data is estimated to be stored as structured or relational data in web databases — deep web makes 96% of all the content on the Internet, which is 500-550 times larger than the surface web [4], [3]. These data contain a vast amount of valuable information and entities such as Infomine [5], Clusty [3], Books In Print [4] may be interested in building an index of the deep web sources in a given domain (such as book). Because these entities cannot access the proprietary web indices of search engines (e.g., Google and Baidu).

II. OBJECTIVES

1) The Objective is to record learned patterns of deep web sites and form paths for incremental crawling.
2) Ranks site URLs to prioritize potential deep sites of a given topic. To this end, two features, site similarity and site frequency, are considered for ranking.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 8, August 2016

3) Focused crawler consisting of two stages: efficient site locating and balanced in-site exploring. Two Stage Crawler performs site-based locating by reversely searching the known deep web sites for center pages, which can effectively find many data sources for sparse domains.
4) Two Stage Crawler has an adaptive learning strategy that updates and leverages information collected successfully during crawling.
5) The Aho Corasic algorithm used finds the most relevant forms and links for the given site and discards the malicious links.

III. 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. Web crawlers are prominently one of the main components of web search engines that assemble a corpus of web pages or creates a copy of all the visited pages, index them, and allow users to issue queries against the index, provide fast searches and find the web pages that match the queries. Interacting with hundreds of thousands of web servers and name servers, crawling is considered as the most fragile application since it is beyond the control of the system.

Crawler follows very simple steps yet very effective work in maintenance, checking of the downloaded links and also the validation of HTML codes as follows It starts with the list of URL's to visit, called seeds and downloads the web page.

IV. RELATED WORK

The previous strategies used to deal with creation of a single profile per user but the conflicts occurred in them. Considering that if a users interest varies for same user query example, that is it may happen that the user is interested in banking exams for the query bank he entered and he may not at all interested for the blood bank. So the chances of occurring a conflict arises and we are dealing with the negative preferences to obtain the fine grain between the interested results and not interested results. Hence consider the following aspects :

1) Document Based Method :
These methods are interested in for capturing the users clicking and browsing behavior. The click through data is obtained from the user that is the documents the user has clicked on. The click through data in search engines is made up of triplets (q,r,c).

Where,
q = query
r = ranking
c = set of links clicked by user.
The document based method is used to search only the documents and it cannot handle the other queries for images , videos etc[8].

2) Concept Based Method :
These type of methods are used for capturing the users conceptual needs, browsed documents and searched histories. The web pages are crawled according to the high relevancy of the users interest. An effective approach of context graph is used which is based on the formal concept analysis for the problem which is to be solved. The concept lattice is formed for the visited pages and the context graph is formed on the basis of upper concept lattice. The system measures relevancy of the expected pages for a given topic and finds the sequence in which the pages should be visited first. The user profiles are used to represent users interest and to infer their intentions for new queries [11].

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 8, August 2016

3) Image Based Searching :
The image based searching is based on mobile images based on images and text. The set of images is obtained by web crawling in a certain restricted domain and the selected domain is based on the expected application. The various web pages across the web are covered by exploiting the keyword based search and the content based filtering method. The location is recognized by finding the images which exactly gets matched with the image of the location. The keywords are extracted from web pages that are matching the images from the set . Hence this searching is totally based on the images in which millions of web pages are crawled, and the most relevant image is found [9].

4) Audio Based Searching :
This type of searching is based on required music stored in the musical database. A Tag Based Semantic Annotation method is used on the basis of tag based approach. This approach categorizes the music on the basis of instruments, genre, artists, language, music company etc. A collaborative filtering method is used in which Tag Based Semantic Annotation is used as an input. A music generator is used to store the music files using the metadata. The audio search engine is used in order to find the required music for the user[10].

VI. SYSTEM ARCHITECTURE: (TWO STAGE ARCHITECTURE)

To efficiently and effectively discover deep web data sources, Two Stage Crawler is designed with two stage architecture, site locating and in-site exploring, as shown in Figure 1. Site locating stage is used for finding the most relevant site for a given topic and in-site exploring stage uncovers searchable forms from the site. The user enters a certain query or keyword which acts as the seed sites. The sites which are previously been searched, stored in site database. So the seed sites acts as an input to the reverse searching process. The reverse searching process is used to find the deep web pages or centre pages. These deep web pages are then parsed to extract various links and forms. The pages gets downloaded where the user clicks. If fetched deep web sites are greater than certain threshold value, it determines the relevancy of various pages. In adaptive learning process one or more searchable forms are found for a particular site. Spy NB classifier is used to classify the previously searched links and the new links to find the most relevant link from it. Ranking is done with the help of feature selection space for various relevant links. The feature selection space is determined by stemming, removal of stopwords, frequency of most frequent term in URL. When the most relevant site is obtained, in second stage of in-site exploring the various forms on that web page are fetched and stored in database. The most relevant form is obtained as an user output with the help of ranking.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 8, August 2016

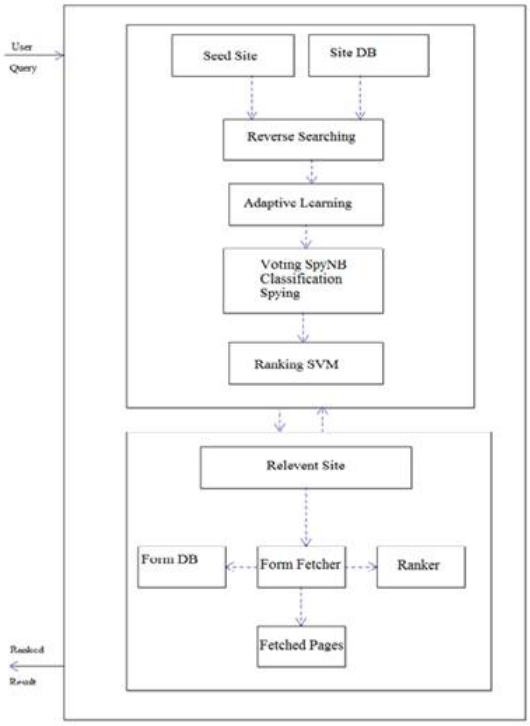

Fig.1 System Architecture: (Two stage Architecture)

VII. ALGORITHMS & TECHNIQUES USED:

Algorithm 1: Reverse Searching
Input: seed sites and harvested deep websites
Output: relevant sites
while of candidate sites less than a threshold **do**
// pick a deep website
site = getDeepWebSite(siteDatabase,
seedSites)
resultP age = reverseSearch(site)
links = extractLinks(resultP age)
for each link in links **do**
page = downloadPage(link)
relevant = classify(page)
if relevant **then**
relevantSites=extractUnvisitedSite(page)
Output relevantSites
end
end
end

Algorithm 2: Aho Corasic Algorithm
Input: relevant sites
Output: extracted relevant forms and links for relevant sites.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 8, August 2016

While queue ! = empty **Do**
Begin
Let relevant link be the next state in queue
Queue < — queue − (relevant link)
For each a such that g (relevant link, a) = s ! =
fail **do**
Begin Queue < — queue U (s)
State < — f (relevant link)
While g (state, a) = fail **Do**
State < — f (state)
f(s) < — g(state, a)
Output < — output(s) U output (f(s))
End
End
End

goto Function
Begin
New state < — 0
For i < — 1 until k **do**
enter(yi)
For all a such that g(0, a) = fail **do**
g(0, a)< ----0
End

AhoCorasic algorithm is used for extracting the most relevant URL. The input provided to the algorithm is extracted URL link from relevant link whereas set of URLs from relevant link is stored in the queue. There is a construction of goto function where the input provided is the set of keywords. It is assumed that the output(s) is empty when state s is first created, whereas g(s, a) = fail. If a is undefined or if g(s, a) has not yet been defined. The procedure enters(y) inserts into the goto graph a path that spells out y.

VIII. RESULTS

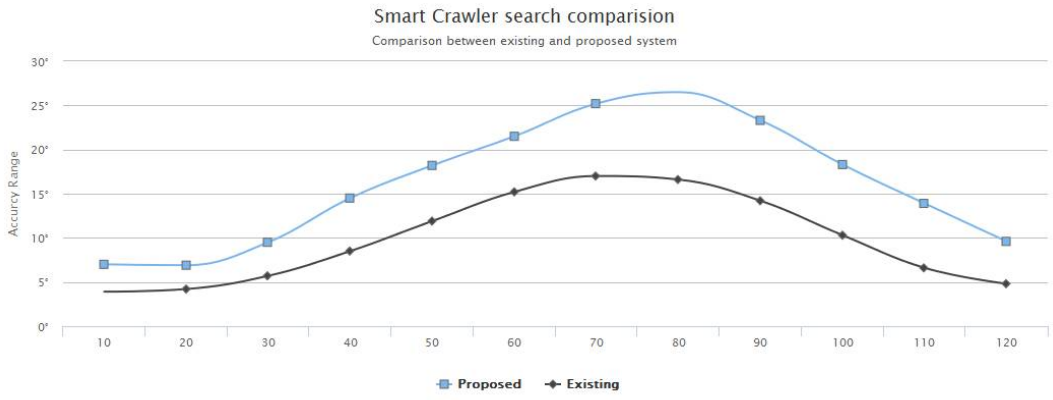

Fig.2 Graphical Comparison

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 8, August 2016

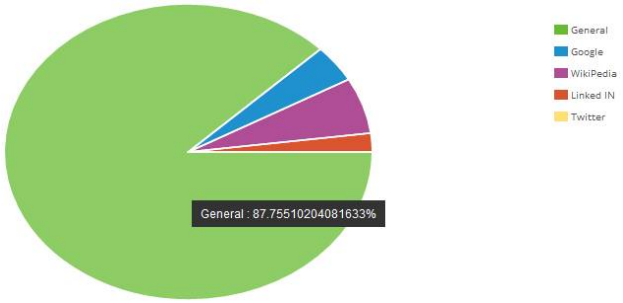

Fig.3 Pie-Chart (For Various Queries)

The above graph shows us that the proposed system of Two Stage Crawler does efficient harvesting of the deep web and gives efficient results than the existing strategies. The Aho Corasic algorithm used in the proposed system gives us the most relevant results for a query entered by the user. Also the Pie-chart tells us the percentage of crawled web pages by the user through Google, Wikipedia or the General Links he has searched.

X. CONCLUSION

The paper proposed a effective web crawler which efficiently harvests the deep web interfaces. The technique of incremental crawler used in the system performs better and is powerful which allows re-visitation of pages at different rates. The smart crawler obtains more accurate results and achieves higher harvest rates than other crawlers. The accuracy of the proposed system is higher than the existing systems. The AhoCorasic algorithm used determines the most relevant URLs. The peer-to- peer issue while crawling at other environment is the future issue to dealt with.

REFERENCES

1. Peter Lyman and Hal R. Varian. How much information? 2003. Technical report, UC Berkeley, 2003.
2. Roger E. Bohn and James E. Short. How much information? 2009 report on american consumers. Technical report, University of California, San Diego, 2009.
3. Martin Hilbert. How much information is there in the"information society"? Significance, 2012.
4. Idc worldwide predictions 2014: Battles for dominance – and survival on the 3rd platform. http://www.idc.com/research/Predictions14/index.jsp, 2014.
5. Michael K. Bergman. White paper: The deep web: Surfacing hidden value. Journal of electronic publishing, 2001.
6. Yeye He, Dong Xin, VenkateshGanti, SriramRajaraman, and Nirav Shah. Crawling deep web entity pages. In Proceedings of the sixth ACM international conference on Web search and data mining, 2013.
7. Feng Zhao, Jingyu Zhou, Chang Nie, Heqing Huang, Hai Jin. Two Stage Crawler: A Two-stage Crawler for Efficiently Harvesting Deep-Web Interfaces. Services Computing, IEEE Transactions on (Volume:PP , Issue:99).
8. Jun Xu1, Yunbo Cao, Hang Li, Nick Craswell, and Yalou Huang. Searching Documents Based on Relevance and Type., ECIR 2007, LNCS 4425, pp. 629 636, 2007.
9. Tom Yeh, Konrad Tollmar, Trevor Darrell. Searching the Web with Mobile Images for Location Recognition. , Computer Vision and Pattern Recognition,2004. Proceedings of 2004 IEEE Computer Society (Volume:2,pp: II-76-II-81).
10. Parameswaran Vellachu1 and Dr. Sunitha Abburu2. Tag Based Audio Search Engine., . IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 2, No 3, March 2012
11. Qiang Peng, Yajun Du, Yufeng Hai, Shaoming Chen, Topic Specific Crawling on the Web with Concept Context Graph Based on FCA. , Management and Service Science ,2009. MASS 09.(20-22 Sept.2009)