A Review on Extracting Underlying Content from Deep Web Interfaces

Unnati N. Bhakare
M. Tech. Scholar
Department of Computer Science and Engineering
Government College of Engineering
Amravati (MH) India
unnatibhakare13@gmail.com

Dr. Prashant N. Chatur
Head
Department of Computer Science and Engineering
Government College of Engineering
Amravati (MH) India
prashant_chatur@rediffmail.com

***Abstract*-Information retrieval and integration of web data is recent trend in today's world of technology. Huge amount of data is available in online repositories but most of it is hidden under deep web interfaces. As deep web is growing at a very fast rate it is becoming difficult to efficiently locate the deep-web interfaces and retrieving the required data. The large volume of web resources and the dynamic nature of deep web are available. This provides the wide coverage and efficient information availability. And thus pose a challenging issue in the sector of information retrieval. The rapid growth of the World-Wide Web poses extraordinary scaling tasks for general-purpose crawlers and search engines. Current-day crawlers recover content only from the openly indexed web, i.e., the set of web pages that are directly or easily reachable by hypertext links, without considering search forms and pages with prerequisites like authorization or earlier registration. This paper reviews the methodology of content extraction form deep web interface.**

Keywords- Crawlers, deep web, hidden web classification, ranking.

I. INTRODUCTION

Deep-web crawl defines the problem of evolving hidden content behind search interfaces available on the Web. As far as the deep-web sites are concerned it maintains the document-oriented text based content (e.g., Wikipedia, PubMed, Twitter, etc.), which has traditionally been the focused on the deep-web literature, we have noticed that a major portion of deep-web sites, including almost all online shopping sites, curate structured objects as divergent to text documents. The objectives of the study on content extraction [2] and the deep web interfacing are as follows:

- To measure the size and importance of Web to know the depth.
- Characterization of the deep Web's content with ease, excellence that content have, and relevance to information explorers.
- Discovering automated techniques for identifying deep web search sites and applying queries over them to get desired results.
- Initiating the process of teaching the Internet searching. Purpose of Crawler is automatically traversing the web data, to retrieve the pages and to build a local repository of the portion of the web that they visit.

Deep web crawling is the process of data collection from interfaces by issuing queries [3]. It can be analyzed from two viewpoints. One is the analyzing the macroscopic views for deep web, like the number of the web data sources, the attributes in the HTML form and the total pages included in the deep web. While deep web crawling is performed, which consists of billions of HTML forms, typically the focus is on the wide coverage of those data sources rather than exhaustively harvesting the content inside one particular data source. That means, only breadth, rather than the depth, of the deep web is preferred when the computing resource of a crawler is limited. In breadth-oriented crawling, data source location, to learn and understand the interface and the return results so that query submission and data extraction are performed automatically are challenging issues.

Another class of crawling is depth oriented, which focus on one selected web data source, with the objective to garner most of the documents from the given data source. Different types of crawling and web harvesting techniques are available some of the are as follows

A. Generic Crawlers

Only searchable forms are focused in this type of crawler and specific domain or topic is ignored.

B. Focused Crawlers

Focused crawlers are nothing but the form-focused crawler (FFC) which focuses on web forms.

C. *ACHE*

Adaptive Crawler for Hidden-web Entries perform automatic search on online databases for specific topic. This method is extension to the FFC.

The crawling and extraction technique can be explained from fig 1. First keyword is passed according to keyword the related URL's are found in second and next levels URL's are crawled until best results are found and the documents are downloaded.

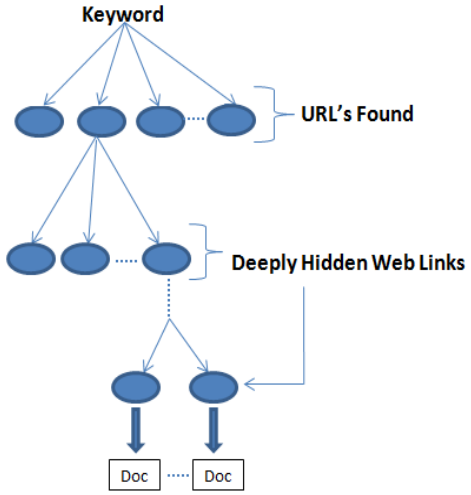

Fig. 1. Content extraction from deep web

II. FUNDAMENTALS

From the overall study we have seen that the above topic put emphasis on some important methods or techniques that are useful to get the deep web data. This can be explained as follows.

A. Finding Web Pages

The strategy for visiting web pages which more likely to relate with the given topic [5]. This helps to find the root page or center page and then the hierarchy of related deep web pages.

B. Link Classification

Forms are found sparsely distributed on web. The crawler can lose worthy target pages that are reachable by performing additional steps due to selection of only that links that bring instant return. Thus, the aim of classifier is to identify links that will be beneficial, that means the links that at the end progresses towards the pages containing forms [7]. It analyzes the features of links as follows: URL, anchor, and text closely related to URL; after this score is assigned to the links corresponding to the distance between the relevant page and a link that is reachable from that link [6]. This is worked out by Backward Crawling. It requires a good path to jump from one link to other. If the distance between two paths is shortest the path is considered as a good path. This can be understood by the example of connectivity graph. Say nodes along the path are to be considered as a links. Exhausting crawling is required over the sites to build this graph [5]. But this is possible only for a finite set of sites. It would be extraordinarily expensive to perform this task on large no of sites. Backward crawl by using the "link" facility of Google is another option for extract connectivity graph. Breadth first search technique is used in backward crawling. It starts from the page containing searchable form and next levels are constructed by finding the previous document in the current.

C. Page Classification

Crawler retrieves pages from the harvested links [6]. Let's say page P is being retrieved. Now its job of classifier to analyze the page and assigning relevance score to it. This will reflect the probability that P belong to focused topic. Page is considered as a relevant if probability crosses the threshold value.

D. Site Ranking

Site ranking is the mechanism of site URLs ranking. Site ranking is performed by two perspectives one is site similarity and site frequency [1]. Topic similarity between known deep web site and new site is measured by site similarity technique. More availability of site in other sites, its popularity gives the frequency of tat site. In this ways using site similarity and site frequency site ranking is performed.

E. Link Ranking

Link ranking technique is used for ranking links of the site. The way site similarity is performed, link similarity is also performed in the same way [1]. The only difference is i) feature space of link is required for link prioritizing. ii) Frequency of link is not considered in link ranking.

F. Retrieving And Integrating Hidden Web Data

System named MetaQuerier [4] is available which helps to integrate the large scale hidden web data. Several components for various aspects of the integration are there. One of these components is the Database Crawler which is crawled for locating online databases. Basically database crawlers do not focus on search for specific topic and also do not find the promising links. Rather it takes the valid IP address of the web server to crawl and find the root page. Once the root page is found then crawling proceeds by breadth first search technique up to fixed depth.

G. Data Extraction

Hidden web is a huge source of structured data also it is not indexed publicly; so it is challenging task to access the pages which are dynamically created through search interfaces [8]. Files containing hidden web data with structured form in different domains are read. After this by using a search query interface, query is input and the related data of specific domain and topic is returned. Several online databases provide option for static URL link by providing dynamic query access through interface. Query interface playa important role to reach towards the hidden web. A huge amount of information can be retrieved from the searchable forms. Traditional crawlers were unable to submit the query replicated by human being. Consider the example of e-commerce sites. These sites consist of HTML forms for interfaces. Components like text fields, radio buttons, checkbox, selection list etc. This interface provides access to the data of user's interest in specific domain. Here the topic entered by the user is considered as a query. Query processor extracts this query and tokens are matched with the attributes present in the repository. And finally the result requested by user is displayed on the page.

CONCLUSIONS

In this paper we have done a survey of content extraction methodology by finding deep web interface. We have studied different crawling techniques along with their basic fundamentals and found that content extraction from web can be improved significantly if deep web interfaces considered. We conclude that using deep web interfaces for content extraction will increase the availability of additional resources hidden under deep web.

REFERENCES

[1] Feng Zhao, Jingyu Zhou, Chang Nie, Heqing Huang, and Hai Jin, "SmartCrawler: A Two-Stage Crawler for Efficiently Harvesting Deep-Web Interfaces" in *IEEE Transactions on Services Computing, Vol. 9, No. 4, July/August 2016.*

[2] M. K. Bergman, "White paper: The deep web: Surfacing hidden value," In *J. Electron. Publishing, volume 7, no. 1, pp. 1–17, 2001.*

[3] Y. He, D. Xin, V. Ganti, S. Rajaraman, and N. Shah, "Crawling deep web entity pages," in *Proc. 6th ACM Int. Conf. Web Search Data Mining, 2013, pp. 355–364.*

[4] K. C.-C. Chang, B. He, and Z. Zhang, "Toward large scale integration: Building a metaquerier over databases on the web," in *Proc. 2nd Biennial Conf. Innovative Data Syst. Res., 2005, pp. 44–55.*

[5] S. Denis, "On building a search interface discovery system," in *Proc. 2nd Int. Conf. Resource Discovery, 2010, pp. 81–93.*

[6] L. Barbosa and J. Freire, "Searching for hidden-web databases," in *Proc. 8th Int. Workshop Web Databases, 2005, pp. 1–6.*

[7] L. Barbosa and J. Freire, "An adaptive crawler for locating hidden web entry points," in *Proc. 16th Int. Conf. World Wide Web, 2007, pp. 441–450.*

[8] Priyanka Jain Megha Bansal, "Efficient Crawling the Deep Web", in *International Journal of Advanced Research in Computer Science and Software Engineering, Volume 4, Issue 5, May 2014.*

[9] S. Chakrabarti, M. V. den Berg, and B. Dom, "Focused crawling: A new approach to topic-specific web resource discovery," in *Compute. Newt, vol. 31, no. 11, pp. 1623–1640, 1999.*

[10] J. Madhavan, D. Ko, º. Kot, V. Ganapathy, A. Rasmussen, and A. Halevy, "Google's deep web crawl," in *Proc. VLDB Endowment, vol. 1, no. 2, pp. 1241–1252, 2008.*

[11] O. Christopher and N. Marc, "Web crawling," *in Found. Trends Inf. Retrieval, vol. 4, no. 3, pp. 175–246, 2010.*

[12] B. Raju and K. Subbarao, "Sourcerank: Relevance and trust assessment for deep web sources based on inter-source agreement," in *Proc. 20th Int. Conf. World Wide Web, 2011, pp. 227–236.*

[13] B. Raju, K. Subbarao, and J. Manishkumar, "Assessing relevance and trust of the deep web sources and results based on intersource agreement," in *ACM Trans. Web, vol. 7, no. 2, pp. Article 11, 1–32, 2013.*

[14] M. E. Dincturk, G. vincent Jourdan, G. V. Bochmann, and I. V. Onut, "A model-based approach for crawling rich internet applications," in *ACM Trans. Web, vol. 8, no. 3, pp. Article 19, 1–39, 2014.*

[15] K. C.-C. Chang, B. He, C. Li, M. Patel, and Z. Zhang, "Structured databases on the web: Observations and implications," in *ACM SIGMOD Rec., vol. 33, no. 3, pp. 61–70, 2004.*

[16] W. Wu, C. Yu, A. Doan, and W. Meng, "An interactive clustering based approach to integrating source query interfaces on the deep web," in *Proc. ACM SIGMOD Int. Conf. Manage. Data, 2004, pp. 95–106.*

[17] E. C. Dragut, T. Kabisch, C. Yu, and U. Leser. (2009, Aug.). A hierarchical approach to model web query interfaces for web source integration. In *Proc. VLDB Endowment [Online]. 2(1), pp. 325–336.*

[18] T. Kabisch, E. C. Dragut, C. Yu, and U. Leser, "Deep web integration with visqi", in *Proc. VLDB Endowment, vol. 3, nos. 1/2, pp. 1613–1616, 2010.*

[19] E. C. Dragut, W. Meng, and C. Yu, 'Deep Web Query Interface Understanding and Integration", *ser. Synthesis Lectures on Data Management. San Rafael, CA, USA: Morgan & Claypool Publishers, 2012.*

[20] A. Bergholz and B. Childlovskii, "Crawling for domain-specific hidden web resources," in *Proc. 4th Int. Conf. Web Inf. Syst. Eng., 2003, pp. 125–133.*

[21] S. Raghavan and H. Garcia-Molina, "Crawling the hidden web," in *Proc. 27th Int. Conf. Very Large Data Bases, 2000, pp. 129–138.*

[22] C. Sheng, N. Zhang, Y. Tao, and X. Jin, "Optimal algorithms for crawling a hidden database in the web," *Proc. VLDB Endowment, vol. 5, no. 11, pp. 1112–1123, 2012.*

[23] P. G. Ipeirotis and L. Gravano, "Distributed search over the hidden web: Hierarchical database sampling and selection," in *Proc. 28th Int. Conf. Very Large Data Bases, 2002, pp. 394–405.*

[24] N. Dalvi, R. Kumar, A. Machanavajjhala, and V. Rastogi, "Sampling hidden objects using nearest-neighbor oracles," in *Proc. 17th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, 2011, pp. 1325–1333.*

[25] J. Madhavan, S. R. Jeffery, S. Cohen, X. Dong, D. Ko, C. Yu, and A. Halevy, "Web-scale data integration: You can only afford to pay as you go," in *Proc. Biennial Conf. Innovative Data Syst. Res., 2007, pp. 342–350.*

[26] M. Khelghati, D. Hiemstra, and M. V. Keulen, "Deep web entity monitoring," in *Proc. 22nd Int. Conf. World Wide Web Companion, 2013, pp. 377–382*.

[27] S. Chakrabarti, K. Punera, and M. Subramanyam, "Accelerated focused crawling through online relevance feedback," in *Proc. 11th Int. Conf. World Wide Web, 2002, pp. 148–159*.

[28] L. Barbosa and J. Freire, "Combining classifiers to identify online databases," in *Proc. 16th Int. Conf. World Wide Web, 2007, pp. 431–440.*

[29] J. Cope, N. Craswell, and D. Hawking, "Automated discovery of search interfaces on the web," in *Proc. 14th Australasian Database Conf.-Volume 17, 2003, pp. 181–189*.

[30] D. Susan and C. Hao, "Hierarchical classification ofWeb content," in *Proc. 23rd Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval, 2000, pp. 256–263*.

[31] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The weka data mining software: An update," in *SIGKDD Explorations Newsletter, vol. 11, no. 1, pp. 10–18, Nov. 2009*.