International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 05 Issue: 03 | Mar-2018 www.irjet.net p-ISSN: 2395-0072

Multi -Stage Smart Deep Web Crawling Systems: A Review

Ms. Sonali J. Rathod[1] , Prof. P. D. Thakare[2]

[1]PG Student, Department of CSE, JCOET Yavatmal, Maharashtra, India
[2]Professors, Department of CSE, JCOET Yavatmal, Maharashtra, India

---

Abstract - *As deep web grows at a very fast pace, there has been increased interest in techniques that help efficiently locate deep-web interfaces. However, due to the large volume of web resources and the dynamic nature of deep web, achieving wide coverage and high efficiency is a challenging issue. In this project propose a three-stage framework, for efficient harvesting deep web interfaces. In the first stage, web crawler performs site-based searching for center pages with the help of search engines, avoiding visiting a large number of pages. In this paper we have made a survey on how web crawler works and what are the methodologies available in existing system from different researchers.*

***Key Words*: Deep web, web mining, feature selection, ranking**

1. INTRODUCTION

The deep (or hidden) web refers to the contents lie behind searchable web interfaces that cannot be indexed by searching 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. More recent studies estimated that 1.9 petabytes were reached and 0.3 petabytes were consumed worldwide in 2007. An IDC report estimates that the total of all digital data created, replicated, and consumed will reach 6 petabytes in 2014. 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 up about 96% of all the content on the Internet, which is 500-550 times larger than the surface web. These data contain a vast amount of valuable information and entities such as Infomine, Clusty, Books In Print 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), there is a need for an efficient crawler that is able to accurately and quickly explore the deep web databases.

It is challenging to locate the deep web databases, because they are not registered with any search engines, are usually sparsely distributed, and keep constantly changing. To address this problem, previous work has proposed two types of crawlers, generic crawlers and focused crawlers. Generic crawlers, fetch all searchable forms and cannot focus on a specific topic. Focused crawlers such as Form-Focused Crawler (FFC) and Adaptive Crawler for Hidden-web Entries (ACHE) can automatically search online databases on a specific topic. FFC is designed with link, page, and form classifiers for focused crawling of web forms, and is extended by ACHE with additional components for form filtering and adaptive link learner.

The link classifiers in these crawlers play a pivotal role in achieving higher crawling efficiency than the best-first crawler. However, these link classifiers are used to predict the distance to the page containing searchable forms, which is difficult to estimate, especially for the delayed benefit links (links eventually lead to pages with forms). As a result, the crawler can be inefficiently led to pages without targeted forms. Besides efficiency, quality and coverage on relevant deep web sources are also challenging. Crawler must produce a large quantity of high-quality results from the most relevant content sources. For assessing source quality, Source Rank ranks the results from the selected sources by computing the agreement between them.

When selecting a relevant subset from the available content sources, FFC and ACHE prioritize links that bring immediate return (links directly point to pages containing searchable forms) and delayed benefit links. But the set of retrieved forms is very heterogeneous. For example, from a set of representative domains, on average only 16% of forms retrieved by FFC are relevant. Furthermore, little work has been done on the source selection problem when crawling more content sources. Thus it is crucial to develop smart crawling strategies that are able to quickly discover relevant content sources from the deep web as much as possible. In this project, I propose an effective deep web harvesting framework, namely Smart Crawler, for achieving both wide coverage and high efficiency for a focused crawler. Based on the observation that deep websites usually contain a few searchable forms and most of them are within a depth of three our crawler is divided into two stages: site locating and in-site exploring. The site locating stage helps achieve wide coverage of sites for a focused crawler, and the in-site exploring stage can efficiently perform searches for web forms within a site. Our main contributions are:

In this project propose a novel three-stage framework to address the problem of searching for hidden-web resources. Our site locating technique employs a reverse searching technique (e.g., using Google's "link:" facility to get pages pointing to a given link) and incremental three-level site prioritizing technique for unearthing relevant sites, achieving more data sources. During the in-site exploring

stage, i design a link tree for balanced link prioritizing, eliminating bias toward web pages in popular directories.

In this project propose an adaptive learning algorithm that performs online feature selection and uses these features to automatically construct link rankers. In the site locating stage, high relevant sites are prioritized and the crawling is focused on a topic using the contents of the root page of sites, achieving more accurate results. During the in site exploring stage, relevant links are prioritized for fast in-site searching. In this project will performed an extensive performance evaluation of Smart Crawler over real web data in 12 representative domains and compared with ACHE and a site-based crawler. Evaluation will shows that our crawling framework is very effective, achieving substantially higher harvest rates than the state-of-the-art ACHE crawler. The results also show the effectiveness of the reverse searching and adaptive learning.

2. LITERATURE SURVEY

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

In this paper, author proposed, deep web grows at a very fast pace, there has been increased interest in techniques that help efficiently locate deep-web interfaces. However, due to the large volume of web resources and the dynamic nature of deep web, achieving wide coverage and high efficiency is a challenging issue. Here propose a two-stage framework, namely SmartCrawler, for efficient harvesting deep web interfaces. In the first stage, SmartCrawler performs site-based searching for center pages with the help of search engines, avoiding visiting a large number of pages. To achieve more accurate results for a focused crawl, SmartCrawler ranks websites to prioritize highly relevant ones for a given topic. In the second stage, SmartCrawler achieves fast in-site searching by excavating most relevant links with an adaptive link-ranking.

2. Jianxiao Liu, Zonglin Tian, Panbiao Liu, Jiawei Jiang, "An Approach of Semantic Web Service Classification Based on Naive Bayes" in 2016 IEEE International Conference On Services Computing, September 2016 [2]

In this paper, author proposed, How to classify and organize the semantic Web services to help users find the services to meet their needs quickly and accurately is a key issue to be solved in the era of service-oriented software engineering. This paper makes full use the characteristics of solid mathematical foundation and stable classification efficiency of naive bayes classification method. It proposes a semantic Web service classification method based on the theory of naive bayes. It elaborates the concrete process of how to use the three stages of bayesian classification to classify the semantic Web services in the consideration of service interface and execution capacity.

3. Bo Tang, Student Member, IEEE, Steven Kay, Fellow, IEEE, And Haibo He, Senior Member, IEEE "Toward Optimal Feature Selection In Naive Bayes For Text Categorization" In IEEE Transactions On Knowledge And Data Engineering, 9 Feb 2016.[3]

In this paper, author proposed, automated feature selection is important for text categorization to reduce the feature size and to speed up the learning process of classifiers. In this paper, author present a novel and efficient feature selection framework based on the Information Theory, which aims to rank the features with their discriminative capacity for classification. Author first revisit two information measures: Kullback-Leibler divergence and Jeffreys divergence for binary hypothesis testing, and analyze their asymptotic properties relating to type I and type II errors of a Bayesian classifier.

4. Amruta Pandit , Prof. Manisha Naoghare, "Efficiently Harvesting Deep Web Interface with Reranking and Clustering", in International Journal of Advanced Research in Computer and Communication Engineering Vol. 5, Issue 1, January 2016.[4]

In this paper, author proposed, the rapid growth of the deep web poses predefine scaling challenges for general purpose crawler and search engines. There are increasing numbers of data sources now become available on the web, but often their contents are only accessible through query interface. Here proposed a framework to deal with this problem, for harvesting deep web interface. Here Parsing process takes place. To achieve more accurate result crawler calculate page rank and Binary vector of pages which is extracted from the crawler to achieve more accurate result for a focused crawler give most relevant links with an ranking. This experimental result on a set of representative domain show the agility and accuracy of this proposed crawler framework which efficiently retrieves web interface from large scale sites.

5. Anand Kumar , Rahul Kumar, Sachin Nigle, Minal Shahakar, "Review on Extracting the Web Data through Deep Web Interfaces, Mechanism", in International Journal of Innovative Research in Computer and Communication Engineering, Vol. 4, Issue 1, January 2016. [5]

In this paper, author proposed, web develops at a quick pace, there has been expanded enthusiasm for procedures that assistance effectively find profound web interfaces. Be that as it may, because of the expansive volume of web assets and the dynamic way of profound web, accomplishing wide scope and high proficiency is a testing issue. Author propose a two-phase system, to be specific SmartCrawler, for productive gathering profound web interfaces. In the primary stage, SmartCrawler performs site-based hunting

down focus pages with the assistance of web crawlers, abstaining from going to a substantial number of pages.

6. Sayali D. Jadhav, H. P. Channe "Comparative Study of K-NN, Naive Bayes and Decision Tree Classification Techniques" in International Journal of Science and Research, Volume 5 Issue 1, January 2016.[6]

In this paper, author proposed, Classification is a data mining technique used to predict group membership for data instances within a given dataset. It is used for classifying data into different classes by considering some constrains. The problem of data classification has many applications in various fields of data mining. This is because the problem aims at learning the relationship between a set of feature variables and a target variable of interest. Classification is considered as an example of supervised learning as training data associated with class labels is given as input. This paper focuses on study of various classification techniques, their advantages and disadvantages.

7. Akshaya Kubba, "Web Crawlers for Semantic Web" in International Journal of Advanced Research in Computer Science and Software Engineering, Volume 5, Issue 5, May 2015.[7]

In this paper, author proposed, 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 for the updated information and give us the efficient result in no time.

8. Monika Bhide, M. A. Shaikh, Amruta Patil, Sunita Kerure,"Extracting the Web Data Through Deep Web Interfaces" in INCIEST-2015. [8]

In this paper, author proposed, the web stores huge amount of data on different topics. The users accessing web data vastly in now days. The main goal of this paper is to locating deep web interfaces. To locating deep web interfaces uses techniques and methods. This paper is focus on accessing relevant web data and represents significant algorithm i.e. adaptive learning algorithm, reverse searching and classifier. The locating deep web interfaces system works in two stages. In the first stage apply reverse search engine algorithm and classifies the sites and the second stage ranking mechanism use to rank the relevant sites and display different ranking pages.

9. Raju Balakrishnan, Subbarao Kambhampati, "SourceRank: Relevance and Trust Assessment for Deep Web Sources Based on Inter-Source Agreement" in WWW 2011, March 28–April 1, 2011. [10]

In this paper, author proposed, selecting the most relevant web databases for answering a given query. The existing database selection methods (both text and relational) assess the source quality based on the query-similarity-based relevance assessment. When applied to the deep web these methods have two deficiencies. First is that the methods are agnostic to the correctness (trustworthiness) of the sources. Secondly, the query based relevance does not consider the importance of the results. These two considerations are essential for the open collections like the deep web. Since a number of sources provide answers to any query, author conjuncture that the agreements between these answers are likely to be helpful in assessing the importance and the trustworthiness of the sources.

10. Luciano Barbosa, Juliana Freire "An Adaptive Crawler for Locating Hidden Web Entry Points" in WWW 2007. [14]

In this paper, author proposed, describe new adaptive crawling strategies to efficiently locate the entry points to hidden-Web sources. The fact that hidden-Web sources are very sparsely distributed makes the problem of locating them especially challenging. Author deal with this problem by using the contents of pages to focus the crawl on a topic; by prioritizing promising links within the topic; and by also following links that may not lead to immediate benefit. Author propose a new framework whereby crawlers automatically learn patterns of promising links and adapt their focus as the crawl progresses, thus greatly reducing the amount of required manual setup and tuning.

3. CONCLUSIONS

This Paper study on various strategies proposes on deep web interface and crawlers to optimize search engine. In past frameworks have numerous issues and difficulties, for example, productivity, parcel conveyance proportion, endto-end delay, connect quality. It is trying to find the deep web databases, since they are not enrolled with any web indexes, are generally scantily disseminated, and keep always showing signs of change. To address this issue, past work has proposed two sorts of crawlers, nonexclusive crawlers and centered crawlers. Nonspecific crawlers get every single searchable frame and can't concentrate on a particular point. This framework actualizing new classifier Naïve Bayes rather than SVM for searchable shape classifier (SFC) and a space particular shape classifier (DSFC). Proposed framework is contributing new module in light of client login for chose enrolled clients who can surf the particular area as indicated by given contribution by the client. This is module is likewise utilized for sifting the outcomes. Pre-Query recognizes web databases by dissecting the wide variety in

substance and structure of structures. To join pre-question and post-inquiry approaches for classifying deep-web structures to assist enhance the precision of the shape classifier.

REFERENCES

[1]. Feng Zhao, Jingyu Zhou, Chang Nie, Heqing Huang, Hai Jin "Smart Crawler: A Two-stage Crawler for Efficiently Harvesting Deep-Web Interfaces" in IEEE TRANSACTIONS ON SERVICES COMPUTING, VOL. 9, NO. 4, JULY/AUGUST 2016.

[2]. Jianxiao Liu, Zonglin Tian, Panbiao Liu, Jiawei Jiang, "An Approach of Semantic Web Service Classification Based on Naive Bayes" in 2016 IEEE International Conference on Services Computing, SEPTEMBER 2016.

[3]. Bo Tang, Student Member, IEEE, Steven Kay, Fellow, IEEE, and Haibo He, Senior Member, IEEE "Toward Optimal Feature Selection in Naive Bayes for Text Categorization" in IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, 9 Feb 2016.

[4]. Amruta Pandit , Prof. Manisha Naoghare, "Efficiently Harvesting Deep Web Interface with Reranking and Clustering", in International Journal of Advanced Research in Computer and Communication Engineering Vol. 5, Issue 1, January 2016.

[5]. Anand Kumar , Rahul Kumar, Sachin Nigle, Minal Shahakar, "Review on Extracting the Web Data through Deep Web Interfaces, Mechanism", in International Journal of Innovative Research in Computer and Communication Engineering, Vol. 4, Issue 1, January 2016

[6]. Sayali D. Jadhav, H. P. Channe "Comparative Study of K-NN, Naive Bayes and Decision Tree Classification Techniques" in International Journal of Science and Research, Volume 5 Issue 1, January 2016.

[7]. Akshaya Kubba, "Web Crawlers for Semantic Web" in International Journal of Advanced Research in Computer Science and Software Engineering, Volume 5, Issue 5, May 2015.

[8]. Monika Bhide, M. A. Shaikh, Amruta Patil, Sunita Kerure,"Extracting the Web Data Through Deep Web Interfaces" in INCIEST-2015.

[9]. 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.

[10]. Raju Balakrishnan, Subbarao Kambhampati, "SourceRank: Relevance and Trust Assessment for Deep Web Sources Based on Inter-Source Agreement" in WWW 2011, March 28–April 1, 2011.

[11]. D. Shestakov, "Databases on the web: National web domain survey," in Proc. 15th Symp. Int. Database Eng. Appl., 2011, pp. 179–184. [12] D. Shestakov and T. Salakoski, "Host-ip clustering technique for deep web characterization," in Proc. 12th Int. Asia-Pacific Web Conf., 2010, pp. 378–380.

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

[13]. D. Shestakov and T. Salakoski, "On estimating the scale of national deep web," in Database and Expert Systems Applications. New York, NY, USA: Springer, 2007, pp. 780–789.

[14]. Luciano Barbosa, Juliana Freire "An Adaptive Crawler for Locating Hidden Web Entry Points" in WWW 2007

[15]. 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.

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