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 5, May 2016

Two Stage Crawler for Discovering Deep or Hidden Web Services

V.Geetha, S. Mathan, K. Ravikumar

Post Graduate Student, Department of CSE, RRASE College of Engineering, Chennai, India

Assistant Professor, Department of CSE, RRASE College of Engineering, Chennai, India

Professor, Department of CSE, RRASE College of Engineering, Chennai, India

ABSTRACT: The web contains enormous amount of information. From that enormous information only small amount of that information is visible to users and a huge portion of the information is not visible to the users. This is because traditional search engines are not able to index or access all information. The information which can be retrieved by following hypertext links are accessed by such traditional search engines. The forms which are not accessed by traditional search engines include login or authorization process. Hidden web refers to that part of the web which is not accessed by traditional web crawlers. An important problem of retrieving desired and good quality of information from huge hidden web database is how to find out and identify the entry points of hidden web database in the Web. The traditional web crawlers may be unable to retrieve all information from deep web databases. Therefore it is the main cause of motivation for retrieving information from deep web. Issues and challenges related to the problem are also discussed. An architecture for accessing hidden web databases that uses an intelligent agent technology through reinforcement learning is proposed. The experimental results show that the reinforcement learning helps in overcoming existing problems and out performs the existing hidden web crawlers in terms of precision and recall.

KEYWORDS: Deep web, web crawler, ranking, adaptive learning.

I. INTRODUCTION

Non-trivial extraction of implicit, previously unknown and potentially useful information from data – Exploration & analysis, by automatic or semi-automatic means, of large quantities of data in order to discover meaningful patterns. We have been collecting a myriad of data, from simple numerical measurements and text documents, to more complex information such as spatial data, multimedia channels, and hypertext documents. Here is a non-exclusive list of a variety of information collected in digital form in databases and in flat files. Some of them are 1. Business transactions 2. Scientific data 3.Medical and personal data 4. Surveillance video and pictures 5. Satellite sensing 6.Games 7. Digital media 8.CAD and Software engineering data 9.Virtual Worlds Text reports and memos 10.The World Wide Web repositories

II. RELATED WORKS

To leverage the large volume information buried in deep web, previous work has proposed a number of techniques and tools, including deep web understanding and integration hidden web crawlers and deep web samplers. For all these approaches, the ability to crawl deep web is a key challenge. Olston and Najork systematically present that crawling deep web has three steps,

1. Locating deep web content sources
2. Selecting relevant sources
3. extracting underlying content

III. PROPOSED SYSTEM

I use an effective deep web harvesting framework, for achieving both wide coverage and high efficiency for a focused crawler. 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:

1. **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
2. **The in-site exploring stage** to design a link tree for balanced link prioritizing, eliminating bias toward WebPages in popular directories.

The output form of an information system should accomplish one or more of the following objectives.

- Convey information about past activities, current status or projections of the
- Future.
- Signal important events, opportunities, problems, or warnings.
- Trigger an action.
- Confirm an action.

Service Collection

To minimize the number of visited URLs, and at the same time maximizes the number of deep websites. To achieve these goals, using the links in downloaded webpages is not enough. This is because a website usually contains a small number of links to other sites, even for some large sites. Two crawling strategies reverse searching and incremental two-level site prioritizing, to find more sites.

Service Extraction:

Once the Site Frontier has enough sites, the challenge is how to select the most relevant one for crawling. In the Site Ranker assigns a score for each unvisited site that corresponds to its relevance to the already discovered deep web sites.

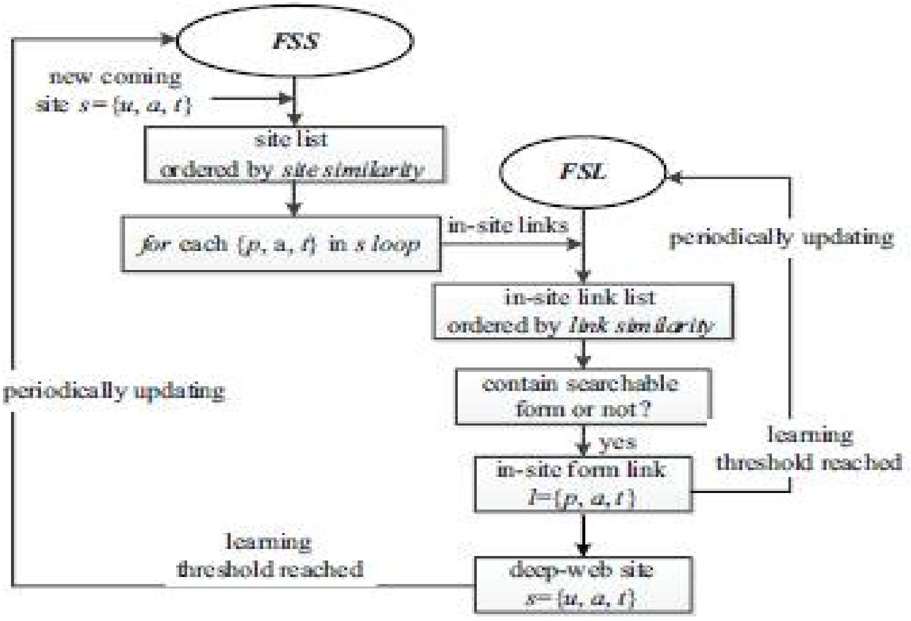
Service Ranking:

After ranking Site Classifier categorizes the site as topic relevant or irrelevant for a focused crawl, which is similar to page classifiers in FFC and ACHE. If a site is classified as topic relevant, a site crawling process is launched. Otherwise, the site is ignored and a new site is picked from the frontier. It determines the topical relevance of a site based on the contents of its homepage. When a new site comes, the homepage content of the site is extracted and parsed by removing stop words and stemming. Then we construct a feature vector for the site and the resulting vector is fed into a Naıve Bayes classifier to determine if the page is topic-relevant or not.

Adaptive Learning

Smart Crawler has an adaptive learning strategy that updates and leverages information collected successfully during crawling. The diagrammatic representation of adaptive learning process of smart crawler is shown below

DOI: 10.15680/IJIRCCE.2016. 0405204



IV. PSEUDO CODE

Reverse searching for more sites.
input : seed sites and harvested deep websites
output: relevant sites

```
while # of candidate sites less than a threshold do
site = getDeepWebSite(siteDatabase, seedSites)
resultP age = reverseSearch(site)
links = extractLinks(resultP age)
foreach link in links do
page = downloadPage(link)
relevant = classify(page)
if relevant then
    relevantSites =extractUnvisitedSite(page)
    Output relevantSites
End
End
End
```

Incremental Site Prioritizing:
input : siteFrontier
output: searchable forms and out-of-site links

```
HQueue=SiteFrontier.CreateQueue(HighPriority)
LQueue=SiteFrontier.CreateQueue(LowPriority)
while siteFrontier is not empty do
if HQueue is empty then
    HQueue.addAll(LQueue)
    LQueue.clear()
end
site = HQueue.poll()
relevant = classifySite(site)
```

```
if relevant then
    performInSiteExploring(site)
    Output forms and OutOfSiteLinks
    siteRanker.rank(OutOfSiteLinks)
if forms is not empty then
    HQueue.add (OutOfSiteLinks)
end
else
    LQueue.add(OutOfSiteLinks)
end
end
```

V. SIMULATION RESULT

The input design is the link between the information system and the user. It comprises the developing specification and procedures for data preparation and those steps are necessary to put transaction data in to a usable form for processing can be achieved by inspecting the computer to read data from a written or printed document. The design of input focuses on controlling the amount of input required, controlling the errors, avoiding delay, avoiding extra steps and keeping the process simple. The input is designed in such a way so that it provides security and ease of use with retaining the privacy. .In output design it is determined how the information is to be displaced for immediate need and also the hard copy output. Efficient and intelligent output design improves the system's relationship to help user decision-making. I have implemented Smart Crawler in Java and evaluated the approach over 12 different domains described in Table shown below,

Domain Description	
Airfare	airfare search
Auto	used cars search
Book	books search
Hotel	hotel search
Job	job search
Movie	movie titles and DVDs search
Music	music CDs search
Rental	car rental search
Apartment	apartment search
Route	map and airline search
Product	household product search
People	sports stars search

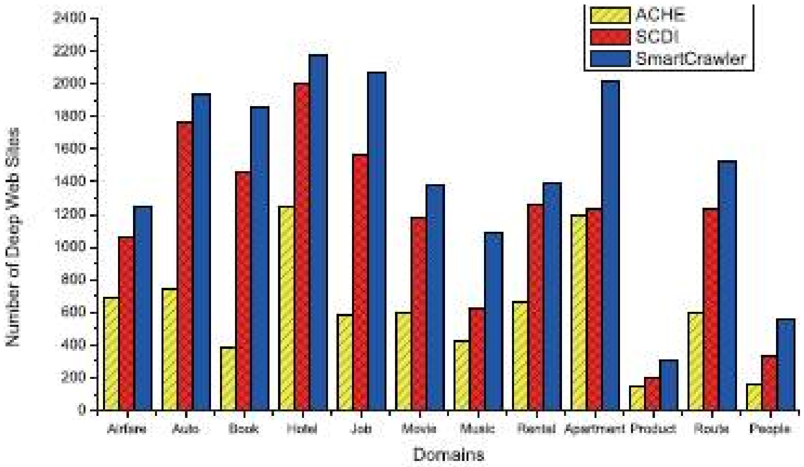


Fig1.The numbers of relevant deep websites harvestedby ACHE, SCDI and *SmartCrawler*

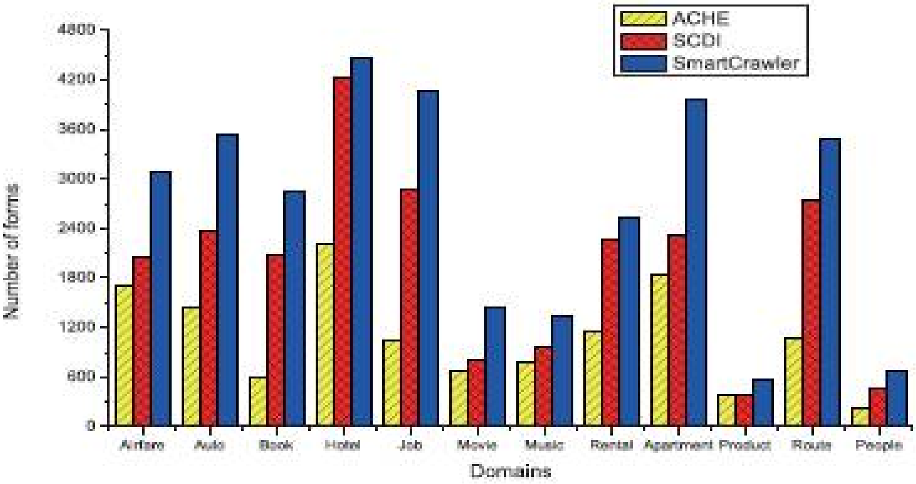


Fig 2.The numbers of relevant forms harvested by ACHE, SCDI and SmartCrawler

VI. CONCLUSION AND FUTURE ENHANCEMENT

An effective harvesting framework for deep-web interfaces, namely Smart Crawler. I have shown that approach achieves both wide coverage for deep web interfaces and maintains highly efficient crawling. Smart Crawler is a focused crawler consisting of two stages: efficient site locating and balanced in-site exploring. Smart 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. By ranking collected sites and by focusing the crawling on a topic, Smart Crawler achieves more accurate results. The in-site exploring stage uses adaptive link-ranking to search within a site; and we design a link tree for eliminating bias toward certain directories of a website for wider coverage of web directories. Experimental results on a representative set of domains show the effectiveness of the proposed two-stage crawler, which achieves higher harvest rates than other crawlers. In future work, I plan to combine pre-query and post-query approaches for classifying deep-web forms to further improve the accuracy of the form classifier.

DOI: 10.15680/IJIRCCE.2016. 0405204

REFERENCES

1. Kevin Chen-Chuan Chang, Bin He, Chengkai Li, Mitesh Patel, and Zhen Zhang. Structured databases on the web: Observations and implications. ACM SIGMOD Record, 33(3):61–70, 2004.
2. Wensheng Wu, Clement Yu, AnHai Doan, and WeiyiMeng.An interactive clustering-based approach to integrating source query interfaces on the deep web. In Proceedings of the 2004ACM SIGMOD international conference on Management of data, pages 95–106. ACM, 2004.
3. Eduard C. Dragut, Thomas Kabisch, Clement Yu, and UlfLeser. A hierarchical approach to model web query interfaces for web source integration. Proc. VLDB Endow. 2(1):325–336,August 2009.
4. Thomas Kabisch, Eduard C. Dragut, Clement Yu, and UlfLeser. Deep web integration with visqi. Proceedings of the VLDBEndowment, 3(1-2):1613–1616, 2010.
5. Eduard C. Dragut, WeiyiMeng, and Clement Yu. Deep WebQuery Interface Understanding and Integration. Synthesis Lectures on Data Management. Morgan & Claypool Publishers,2012.
6. Andr ´ e Bergholz and Boris Childlovskii. Crawling for domain specific hidden web resources. In Web Information Systems Engineering, 2003. WISE 2003. Proceedings of the Fourth International Conference on, pages 125–133. IEEE, 2003.
7. SriramRaghavan and Hector Garcia-Molina. Crawling the hidden web. In Proceedings of the 27th International Conference on Very Large Data Bases, pages 129–138, 2000.
8. Cheng Sheng, Nan Zhang, Yufei Tao, and Xin Jin. Optimalalgorithms for crawling a hidden database in the web. Proceedings of the VLDB Endowment, 5(11):1112–1123, 2012.
9. Panagiotis G Ipeirotis and Luis Gravano. Distributed searchover the hidden web: Hierarchical database sampling and selection. In Proceedings of the 28th international conference on Very Large Data Bases, pages 394–405. VLDB Endowment, 2002.
10. NileshDalvi, Ravi Kumar, AshwinMachanavajjhala, and VibhorRastogi. Sampling hidden objects using nearest-neighbor oracles. In Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining, pages 1325–1333. ACM, 2011.
11. JayantMadhavan, Shawn R. Jeffery, Shirley Cohen, Xin Dong,David Ko, Cong Yu, and Alon Halevy. Web-scale data integration: You can only afford to pay as you go. In Proceedings of CIDR, pages 342–350, 2007.
12. MohamamdrezaKhelghati, DjoerdHiemstra, and Maurice Van Keulen. Deep web entity monitoring. In Proceedings of the 22nd international conference on World Wide Web companion, pages 377–382. International World Wide Web Conferences Steering Committee, 2013.

BIOGRAPHY

V.Geetha is a post graduate student in the department of computer science engineering, Rrase College of engineering, Anna University. Received Bachelor of Engineering degree (B.E) in 2014 from ARN College of engineering technology, Anna University, Chennai, India. Her research interests are networking, cloud computing, Information Security, data mining etc.

DOI: 10.15680/IJIRCCE.2016. 0405204