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 10, October 2016

A Review Paper on Smart Web Crawler

Virkar Snehal Dattatray[1], Shinde Rahul Machindra[2], Kaphare Shradha Subhash[3], Prof. Wavhal D.N[4]

Final Year Student, Department of Computer Engineering, Jaihind College of Engineering, Kuran, India[1]

Final Year Student, Department of Computer Engineering, Jaihind College of Engineering, Kuran, India[2]

Final Year Student, Department of Computer Engineering, Jaihind College of Engineering, Kuran, India[3]

Assistant Professor, Department of Computer Engineering, Jaihind College of Engineering, Kuran, India[4]

ABSTRACT: It's a challenging issue to achieve wide coverage and high efficiency Due to the large volume of web resources and the dynamic nature of deep web.So we propose A Smart Web Crawler which search and discovers Number of centre pages from deep web and focus its trajectory towards that topic in first stage i.e Site locating due to which it avoids visiting a large number of pages. Smart Web Crawler ranks websites to prioritize highly relevant ones for a given topic. After searching centre pages in first stage it performs in-site exploration by excavating most relevant links with an adaptive link-ranking in second stage. Also there was confliction occurrences according to users interest due to single user so this drawback is also avoided in personalized Web Search engine.In this paper page refresh policy is used which redownloads the previously stored pages in the repository due to which HTTP requests are minimised so energy consumption and total staleness of pages are automaticlly decreases The existing crawler issues a large number of HTTP request to web server due to which there is more energy consumption and carbon footprint of web servers.

KEYWORDS: Deep web, two-stage crawler, carbon footprint, ranking, adaptive learning**,** personalization (profile based), staleness, greenness.

I. INTRODUCTION

A Web Crawler also known as a robot or a spider is a system for the bulk downloading of web pages. To solve this heterogeneity problem and to facilitate better sharing of geospatial information, standards have been developed by a variety of organizations [1]. 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), where large sets of web pages are periodically collected and archived for posterity [5]. 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, 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 .

It is challenging to locate the deep web databases because they are deeply distributed and changes continuosly also there was name conflict with single user if any user has profession doctor and he fired a query for bank then using single user profile there was name confliction,because the doctor is interested in the blood bank so he fires the query only bank but due to single user profile it shows the results like national bank, money bank,banking exams insted of blood bank so this drawback is also covered in this paper.We provided two types of user first is normal user and second is professional user i.e user with profession due to which confliction reduces between the query which is fired [4]. So PWS, is a search technique used for providing the better search results to the individual user. PWS also improve the quality of web search with increasing use of individual user profiled information through user's query history and user's documents [6].

Also we focus on the greenness of the web crawling which includes the two types of operations such as page discovery and page refresh operation [7]. Page Discovery operation increases the size of web repository When there is large number HTTP request then energy cosumption as well as carbon footprint of web server also more so to reduce the energy consumption and HTTP request there is a neccessity of green crawling. Page refresh operation decreases the total staleness of pages in the web repository and also carbon footprint reduction takes place [8]. So green crawling is more efficient and providing the better search resullt to the individual user.

The main **Goal and Objective** of this paper is:

- The Objective is to record learned patterns of deep web sites and form paths for incremental crawling.
- 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.
- 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 centre pages, which can effectively find many data sources for sparse domains.
- Smart Crawler has an adaptive learning strategy that updates and leverages information collected successfully during crawling.

II. LITERATURE SURVEY

There are several number of literature surveys are done in this paper

"An active crawler for discovering geospatial Web services and their distribution pattern – A case study of OGC Web Map Service."

The proposed crawler achieves good performance in crawling efficiency and results' coverage. In addition, an interesting finding regarding the distribution pattern of WMSs is discussed. We expect this research to contribute to automatic GWS discovery over the large-scale and dynamic World Wide Web and the promotion of operational interoperable distributed geospatial services.

From this paper we have referred:-

- Concept of web crawling for search.
- It describes process of crawling.

"Search Engines going beyond Keyword Search: A Survey"

This paper tries to identify the major challenges for today's keyword search engines to adapt with the fast growth of web and support comprehensive user demands in quick time. Then it surveys different non-keyword based paradigms proposed, developed or implemented by researchers and different search engines and also classifies those approaches according to the features focused by the different search engines to deliver results.

From this paper we have referred:-

- Concept of keyword based search.
- Semantic web search and question and answering system.

"Diachronic Linked Data: Towards Long-Term Preservationof Structured Interrelated Information."

In this article we focus on the key problem of preserving evolving structured interlinked data. We argue that a number of issues, which hinder applications and users, are related to the temporal aspect that is intrinsic in Linked Data. We present three use cases to motivate our approach, we discuss problems that occur, and
Propose a direction for a solution.

From this paper we have referred:-

- How to get structured information from web search.
- Acquire related information**.**

"Personalization on E-Content Retrieval Based on Semantic Web Services."

This model proposes a new approach to filtering the educational content retrieved based on Case-Based Reasoning .It is based on the model AIREH a multi-agent architecture that can search and integrate heterogeneous educational content through a recovery model that uses a federated search. The advantages of the proposed architecture, as outlined in this article, are its flexibility, customization, integrative solution and efficiency.

DOI: 10.15680/IJIRCCE.2016. 0410085

From this paper we have referred:-

- How to make personalize web search .

"Internet Applications: The Emerging Global Computer."

The initiative includes a sort of grammar and vocabulary that provide information about a document's components; this information will enable Web software to act on the meaning of Web content. Semantic Web software includes a special set of Extensible Markup Language (XML) tags that includes Uniform Resource Identifiers (URIs), a Resource Description Framework (RDF), and a Web Ontology Language (OWL).

From this paper we have referred:-

- How to make deep web search.

III. PROPOSED SYSTEM ARCHITECTURE

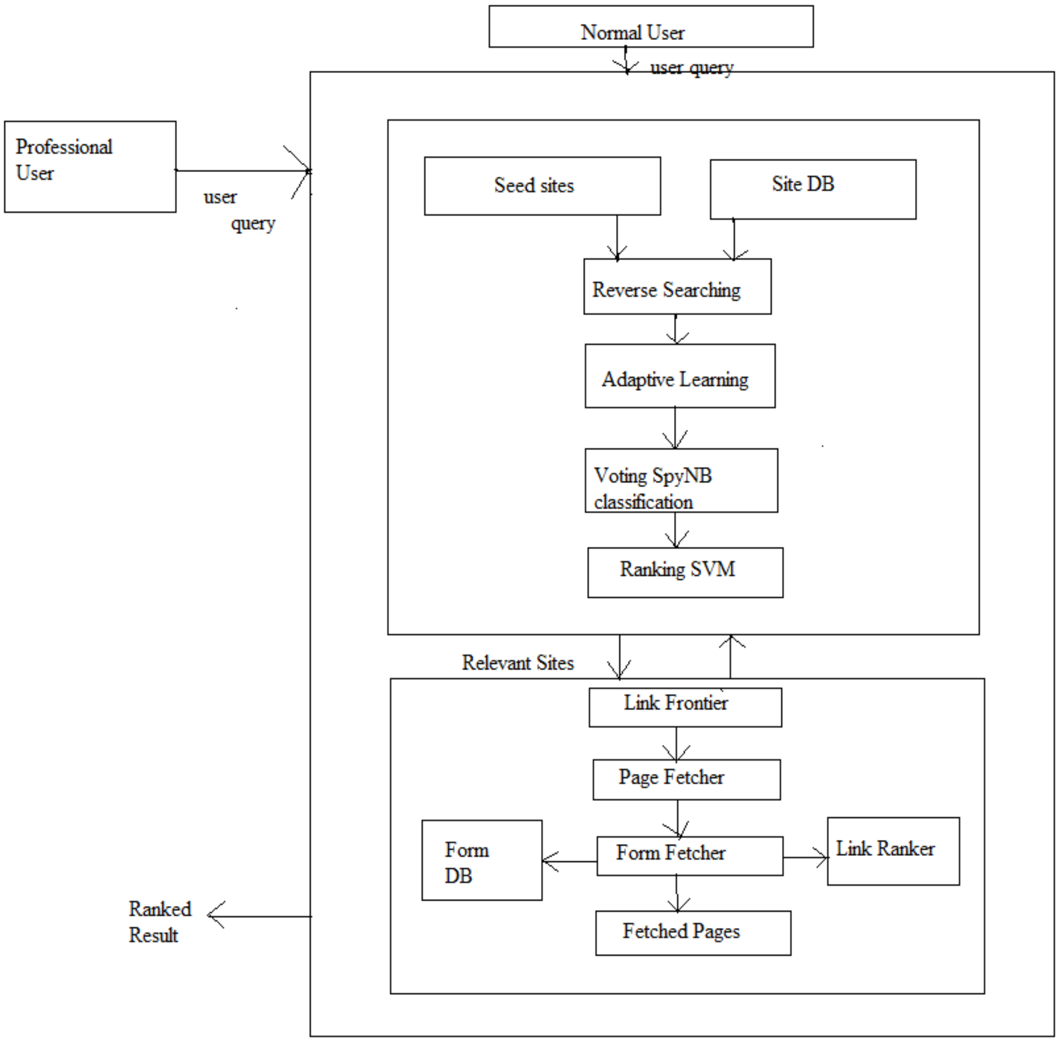

Fig 1-Proposed System Architecture (Two Stage Architecture)

DOI: 10.15680/IJIRCCE.2016. 0410085

To efficiently and effectively discover deep web data sources, Smart Web Crawler is designed with a two stage architecture, site locating and in-site exploring, The first site locating stage finds the most relevant site for a given topic, and then the second in-site exploring stage uncovers searchable forms from the site. Specifically, the site locating stage starts with a seed set of sites in a site database . Seeds sites are candidate sites given for Smart Web Crawler to start crawling, which begins by following URLs from chosen seed sites to explore other pages and other domains. When the number of unvisited URLs in the database is less than a threshold during the crawling process, Smart Web Crawler performs ”reverse searching” of known deep web sites for centre pages (highly ranked pages that have many links to other domains) and feeds these pages back to the site database. Site Frontier fetches homepage URLs from the site database, we going to rank the relevant information also it minimize the query conflict using personalization(profile based).The number of HTTP requests to web crawler minimized and also carbon footprint decreased in proposed s/m using page refreshing policy.

Advantages of Proposed system:-

- Smart Crawler has an adaptive learning strategy that updates and leverages information collected successfully during crawling.
- The first site locating stage finds the most relevant site for a given topic, and then the second in-site exploring stage uncovers searchable forms from the site.
- Two crawling strategies reverse searching and incremental two-level site prioritizing, to find more sites.
- Avoid Deep-web interfaces issues.
- Achieving wide coverage and high efficiency is a challenging issue

IV. CONCLUSION AND FUTURE WORK

Propose an effective harvesting framework for deep-web interfaces, namely Smart Web Crawler. Smart Web Crawler is a focused crawler consisting of two stages: efficient site locating and balanced in-site exploring with neglecting confliction in the users query i.e personalization and minimizing the HTTP request to web server and minimizing total staleness of pages in the repository of web crawler. In future work, we plan to combine pre-query and post-query approaches for classifying deep-web forms to further improve the accuracy of the form classifier.

REFERENCES

[1] An active crawler for discovering geospatial Web services and their distribution pattern – A case study of OGC Web Map Service-WenwenLia*, Chaowei Yang.
[2] Search Engines going beyond Keyword Search: ASurvey- MahmudurRahman.
[3] Diachronic Linked Data: Towards Long-Term PreservationOf Structured Interrelated Information: Sören Auer, François Bancilhon, Peter Buneman, VassilisChristophides.
[4] Personalization on E-Content Retrieval Based on Semantic Web Services -A.B. Gil1, S. Rodríguez1, F. de la Prieta1 and De Paz J.F.1.
[5] Internet Applications: The Emerging Global Computer
[6] J.Teevan, S.T. Dumais, and E.Horvitz, "Personalizing Search via Automated Analysis of interests and Activities," Proc.28th Ann.Int'l ACM SIGIR conf.Research and developement in information retrieval,pp.449-456, 2005.
[7] V.Hatzi, B.B. Cambazoglu, and I.Kousopoulos ,"Web page download scheduling policies for green web crawling," in Proc. 22nd Int. Conf. Software Telecommun. Comput. Netw. (SoftCom), 2014, pp. 56-60.
[8] Vertatique. (Oct. 15, 2009). Carbon Footprints Of Servers Can Vary By 10X [online]. Available: http://www.vertatique.com/carbon-footprints-servers-can-vary-10X.