

Deep Web Crawling Efficiently using Dynamic Focused Web Crawler

Patil Ashwini Madhusudan¹, Prof. Lambhate Poonam D.²

¹Department of Computer Engineering, JSCOE hadapsar Pune, Pune University, India

²Department of Computer Engineering, JSCOE hadapsar pune, Pune University, India

Abstract - The deep web also called invisible web may have the valuable contents which cannot be easily indexed by a search engine. Thus, to locate the deep web or hidden web contents a need of web crawler arise. The opposite term to the deep web is surface web that can be easily seen by a search engine. The deep web is made up of Academic information, medical records, scientific reports, government resources and many more web contents. The web pages changes swiftly and dynamically. The size of the deep web is enormous. Due to this wide and vast nature of deep web, accessing the deep web contents becomes difficult. Current approaches lack to index the deep web pages for a search engine. To address these problems, in this paper, we propose a focused semantic web crawler. The crawler will help users to efficiently access the valuable and relevant deep web contents easily. The crawler works in two stages, first will fetch the relevant sites. The second stage will retrieve the relevant sites through deep search by in-site exploring.

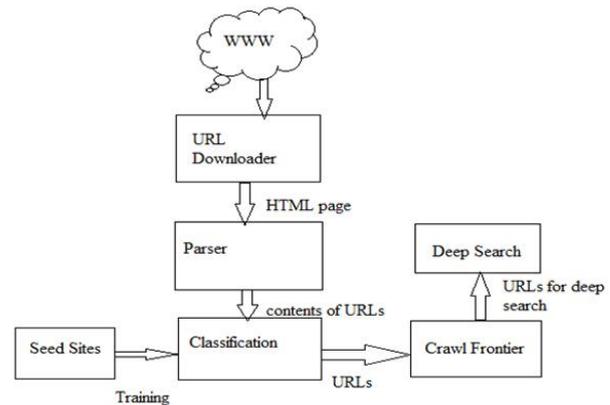


Fig. Web Crawler Architecture

Key Words: Deep Web, Page Ranking, Web Crawler.

1. INTRODUCTION

The deep web is also called invisible web. The deep web may have the valuable contents. 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]. Deep web makes up about 96% of all the content on the Internet, which is 500-550 times larger than the surface web [5], [6]. The opposite term to the deep web is surface web that can be easily seen by a search engine. The deep web is made up of Academic information, medical records, scientific reports, government resources and many more web contents. The deep web databases are not registered with any search engine because they change continuously and so cannot be easily indexed by a search engine. Thus, to locate the deep web or hidden web contents a need of web crawler arises. The size of deep web is increasing very fast these days. The use and the structure of the web is changing day by day. Old data is getting outdated and new information is being added to deep web. The existing approaches lack to efficiently locate the deep web which is hidden behind the surface web. Thus, the need of a dynamic focused crawler arises which can efficiently harvest the deep web contents. In this paper, we propose a focused semantic web crawler. The proposed crawler works in two stages, first to collect relevant sites and second stage for in-site exploring i.e. Deep search.

2. LITERATURE REVIEW

The very famous web which known as ALIWEB is brought up in 1993 as the web page alike to Archie and Veronica. In its place of categorization records, web care taker would submit a systematic data file including site information [3]. The following research in categorization is later in 1993 with spiders. spiders worn the web for web page information like robots. Before some days, we were looking on only the header information, and the uniform resource locator as a query keyword source. To query database techniques were very easy and primitive.

Excite, is first popular search engine which has its roots in these early days of web Categorizing. The undergrad students of Stanford university started this. It was proclamation for universal practice in 1994[3]. From the recent few years there are many procedures and practices are projected probing the hidden web or searching the hidden data from the web. The next level was development of meta search engine. The release of this meta search engine is around 1991-1994. It offers the contact to many search engines at a time by providing one query as input Graphical User Interface. It is developed in university of Washington. Later there are many work is going on this and directed. Google projected by cheek Hong dng and rajkumar bagga in which the use Google API for probing and supervisory exploration of Google. The inherent technique and purpose collection is guided [Google].

At the end of 2011 architecture of web service for Meta based search engine was proposed by K PV Shrinivas, A Govardhan conferring to their learning the Meta quest engine can be classified into two types. first is common purpose search engine and second superior purpose Meta search engine. There is shift from search in all domains to search data in specific domains [12]. Information retrieval is a technique of searching and retrieving the relevant information from the database. The efficacy and efficiency of searching is measure using performance measure matrix called precision and recall. Precision Specifies the document which are retrieved that are relevant and Recall Specifies that whether all the document that are retrieved are relevant or not. To retrieve the complex query information is still checking for search engine is known as deep web. Deep web is unseen web content of openly obtainable pages with information in database such as Catalogues and reference that are not indexed by search engine [12].

There are two ways to access the deep web. The first is to create vertical search engines for specific domains and the second way is surfacing. The prototype system for surfacing Deep-Web content is proposed. The proposed algorithm efficiently traverses the search space and identifies the URLs that can be indexed by the Google search engine [3].

The deep web contains a huge number of HTML forms with a variety of schemata like Google Base. Google Base allows storage of any kind of data as per the users need [22].

3. SYSTEM ARCHITECTURE

The proposed web crawler uses cosine similarity algorithm. The query is entered by a user through GUI. Then the crawler will fetch some relevant and irrelevant URLs from search engine. We apply stop word removal and stemming process on those URLs. After that the title and description matching of appropriate URLs is done with user query. If the page contains more relevant URLs, the same process is repeated for deep search. Now the crawler will apply Cosine Similarity algorithm and gives the values of precision and recall. Thus, our crawler gives good results efficiently.

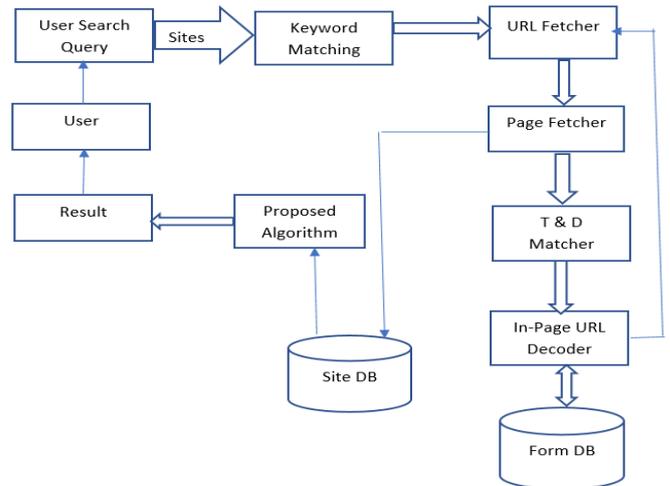


Fig: Proposed System Architecture

4. SYSTEM ANALYSIS

The user enters a search query the keyword matching is done with Google URLs. From these URLs, the relevant URLs will be selected for Deep web search one by one with the help of cosine score. Here the crawler uses adaptive learning strategy. The crawler also records the learned patterns and store it in the database. Thus, the crawler becomes smart and efficient.

5. SYSTEM ALGORITHM

- Step1: Retrieve URLs from Google search engine.
- Step2: Compare the keyword in the query with the Description and title of the URL.
- Step3: Classification is done of fetched URLs.
- Step4: Ranking is assigned to pages using Cosine similarity.
- Step5: The Relevant URLs will be displayed to the user according to their priority.
- Step6: The user can go for deep search through the links displayed.

5.1 Algorithm to calculate Cosine Score

```

    COSINESCORE(q)
    1 INITIALIZE(Scores[d ∈ D])
    2 INITIALIZE(Magnitude[d ∈ D])
    3 for each term(t ∈ q)
    4   do p ← FETCHPOSTINGSLIST(t)
    5     dft ← GETCORPUSWIDESTATS(p)
    6     αt,q ← WEIGHTINQUERY(t, q, dft)
    7     for each {d, tft,d} ∈ p
    8       do Scores[d] += αt,q · WEIGHTINDOCUMENT(t, q, dft)
    9 for d ∈ Scores
    10 do NORMALIZE(Scores[d], Magnitude[d])
    11 return top K ∈ Scores
  
```

6. MATHEMATICAL MODEL

There are two vectors to calculate cosine score.

1. User search query
2. URL

$$sim(q, d) = \frac{\vec{V}(q) \cdot \vec{V}(d)}{|\vec{V}(q)||\vec{V}(d)|}$$

Using this score relevant documents are fetched. Use the count to calculate precision and recall.

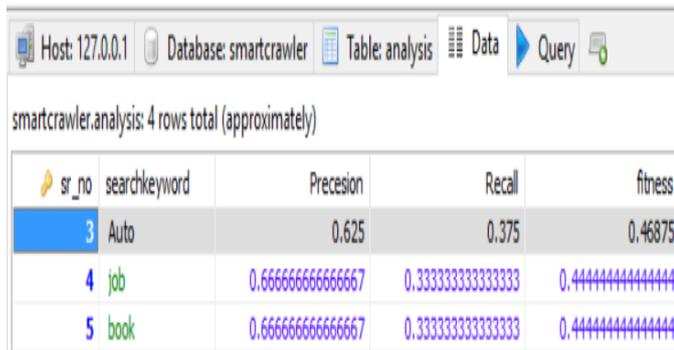
$$Recall = \frac{tp}{tp + fn}$$

$$Precision = \frac{tp}{tp + fp}$$

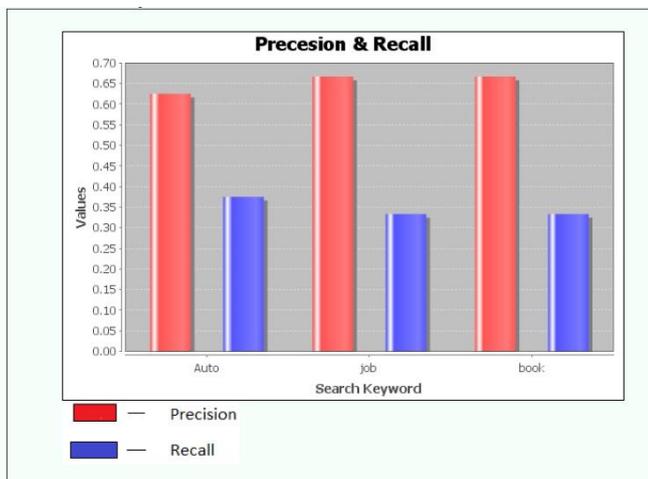
Above a predefined threshold, the relevant documents will be displayed to the user.

7. RESULTS AND DISCUSSION

The results are displayed on the basis of search keyword (the user search query).



sr_no	searchkeyword	Precesion	Recall	fitness
3	Auto	0.625	0.375	0.46875
4	job	0.66666666666667	0.33333333333333	0.44444444444444
5	book	0.66666666666667	0.33333333333333	0.44444444444444



8. CONCLUSION

In this paper, we assessed the different searching approaches for deep web. We developed a focused web crawler that harvests the deep web contents efficiently. The crawler works in two stages first locates the relevant sites and second stage for deep search. As the crawler is focused, it gives topic relevant result and use of cosine score helps to achieve more accurate results. Thus, the developed crawler overcomes the problem of accessing deep web contents and gives users the valuable contents that lie behind the surface web.

ACKNOWLEDGEMENT

I would like to thank my guide Prof. Lambhate P. D.

REFERENCES

[1] Feng Zheo, Jingyu Zhou, Chang Nie, Heqing Huang, Hai Jin. SmartCrawler: A Two-Stage Crawler for Efficiently Harvesting Deep-Web Interfaces. IEEE Transactions on Services Computing, vol 99, 2015

[2] Poonam P. Doshi, Dr. Emmanuel M : feature extraction techniques using semantic based crawler for search engine. proceedings of international conference on computing, communication and energy systems. 2016..

[3] Booksinprint. Books in print and global books in print access. <http://booksinprint.com/>, 2015.

[4] Idc worldwide predictions 2014: Battles for dominance and survival on the 3rd platform. <http://www.idc.com/research/Predictions14/index.jsp> 2014..

[5] Infomine. UC Riverside library. <http://lib->

[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, pages 355364. ACM, 2013.

[7] Martin Hilbert. How much information is there in the information society? Significance, 9(4):812, 2012.

[8] Balakrishnan Raju and KambhampatiSubbarao. Sourcerank: Relevance and trust assessment for deep web sources based on inter-source agreement. In Proceedings of the 20th international conference on World Wide Web, pages 227236, 2011.

[9] Denis Shestakov. Databases on the web: national web domain survey. In Proceedings of the 15th Symposium on International Database Engineering Applications, pages

179184. ACM, 2011. [10] Olston Christopher and Najork Marc. Web crawling. Foundations and Trends in Information Retrieval, 4(3):175246, 2010.

[11] Denis Shestakov and TapioSalakoski. Host-ip clustering technique for deep web characterization. In Proceedings of the 12th International Asia- Pacific Web Conference (APWEB), pages 378380. IEEE, 2010.

[12] Clustys searchable database dirctory. <http://www.clusty.com/>, 2009.

[13] Roger E. Bohn and James E. Short. How much information? 2009 report on american consumers. Technical report, University of California, San Diego, 2009.

[14] Jayant Madhavan, David Ko, ucjaKot, Vignesh Ganapathy, Alex Rasmussen, and Alon Halevy. Googles deep web crawl. Proceedings of the VLDB Endowment, 1(2):12411252, 2008. [15] Luciano Barbosa and Juliana Freire. An adaptive crawler for locating hidden-web entry points. In Proceedings of the 16th international conference on World Wide Web, pages 441450. ACM, 2007.

[16] Denis Shestakov and TapioSalakoski. On estimating the scale of national deep web. In Database and Expert Systems Applications, pages 780789. Springer, 2007.

[17] Luciano Barbosa and Juliana Freire. Searching for hidden-web databases. In WebDB, pages 16, 2005.

[18] Kevin Chen-Chuan Chang, Bin He, and Zhen Zhang. Toward large scale integration: Building a metaquerier over databases on the web. In CIDR, pages 4455, 2005.

[19] Peter Lyman and Hal R. Varian. How much information? 2003. Technical report, UC Berkeley, 2003.

[20] Michael K. Bergman. White paper: The deep web: Surfacing hidden value. Journal of electronic publishing, 7(1), 2001.

[21] SoumenChakrabarti, Martin Van den Berg, and Byron Dom. Focused crawling: a new approach to topic-specific web resource discovery. Computer Networks, 31(11):16231640, 1999.

[22] Jayant Madhavan, 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.

Author Profile



Patil Ashwini, is currently pursuing M.E (Computer) from Department of Computer Engineering, Jayawantrao Sawant College of Engineering,Pune, India. Savitribai Phule, Pune University, Pune, Maharashtra,India -411007.

She received her B.E.(Computer) Degree from BVCOEW Bharati Vidyapeeth's College of Engg. For Women, Savitribai Phule Pune University, Pune, Maharashtra, India - 411007. Her area of interest is Data Mining.



Prof. P.D.Lambhate, received her Degree from WIT, solapur, ME(Comp) from BVCOE Pune, Pursing PhD. in computer Engineering.

She is currently working as Professor at Department of Computer and IT, Jayawantrao Sawant College of Engineering, Hadapsar, Pune, India 411028, affiliated to Savitribai Phule Pune University, Pune, Maharashtra, India -411007. Her area of interest is Data mining, search engine.