ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 6, June 2017

User's Profession Based Web Searching by Using Multi-Word Smart Crawler

Pallavi Gangwal[1], D. B. Kshirsagar[2]

P.G. Student, Department of Computer Engineering, Sanjivani Engineering College, Kopargaon, Maharashtra, India[1]

Professor, Department of Computer Engineering, Sanjivani Engineering College, Kopargaon, Maharashtra, India[2]

ABSTRACT:The internet has very large collection of web pages which contains lots of information. To retrieve necessary and relevant information from this collection is hard task. This problem is addressed by Multi-word Smart Crawler using combination of reverse searching and site prioritizing.

This crawler is working in two types of search: Normal & Personalize search. In searching process, reverse searching and site prioritizing concept is used. Here system gives bookmark option to store the links. In normal search, system will search for a query word without considering user information and gives relevant results according to query word. In personalize search, system will consider user's profession and field information to search links relevant to query word. It results in retrieving more number of pages efficiently with higher harvest rate within less time than other crawlers by maintaining log file for each user and for each query. As well as, system gives the categorization of links with respect to source domains. Here, user can use the option to send mail of text file which contain bookmarked links on registered e-mail id.Experimental results shows that user gets more number of links which are relevant to his profession and query word. As relevancy increases, precision value also gets increased. Execution time for performing personalized search is very less as comparative to normal search. Resulting sites of personalized search are extracted from more number of different source domains such as google, facebook etc.

KEYWORDS:Crawler, Ranking, Site Prioritizing, Reverse searching, Multiword extraction.

I. INTRODUCTION

From large collection of web pages on internet, to retrieve most relevant pages according to user's need is very hard task. For example, if user enters a query word – "Apple" and expect in return the web pages with respect to "Apple Company". But most of the time user may get web pages which contain different meaning of apple such as fruit, company, phone, laptop, etc. To solve this problem, Multi-word smart crawler retrieves links which are relevant to user query word with respect to user's profession. While registration, user will enter his profession with relevant filed which is used further in searching process. Here, system gives two types of option of searching, i.e., normal and personalize search. In normal search, user gets the relevant links with respect to query word only. In personalize search, system will consider user's profession and filed value in addition with query word for searching relevant pages. In searching process, reverse searching and site prioritizing algorithm is used in different manner.

To get more number of pages, reverse searching algorithm is used. On these pages, site prioritising algorithm is used to prioritize these sites. For this purpose, system parses the homepages of links and finds query word frequency in that page. The page which has highest priority will consider as most relevant page by considering manual set threshold value of frequency.

In this system, user can save bookmarks of most required links in future and can send to his registered mail id. Also, system uses log file for same query word & same number of requested site from server within a day. Therefore it reduces search time of system for the same query word again & again in a same day. As day passes, system will start a new search for that query word.

Paper is organized as follows. Section II describes previous proposed crawlers. System overview is explained in section III. Section IV presents experimental results of query words tested. Finally, Section V presents conclusion.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 6, June 2017

II. RELATED WORK

There are many literatures in the area of web crawlers.In late 1994, The RBSE (Repository Based Software Engineering) project first launch the Web Crawler based on twoprograms: first was spider, it maintain a queue in a relationaldatabase, and second was mite, it is a modified www ASCIIbrowser that download the pages from web.But, first crawler did not focus on scalability. Later so many webcrawlers are made introduced. Some of them are: Lycos Infoseek,Exite, AltaVista and HotBot.

In 2001, S.Raghavan et.al. [5] gave new technique for crawlinghidden web.LITE technique is introduced.This technique giveseffective form processing label matching process. This technique has some drawbacks.-1)Task specific approach requires good quality input fromhuman.2) Unable to recognize respond to simple dependencies betweenform fields.3)Lack of support for partial information in form fields.

In 2002,P.G.Ipeirotis et.al. [8] proposed distributed searchover the hidden web by using hierarchical database samplingand selection[8]. This technique gives the facility of metasearch tools and content-summarization technique. It does notconsider query for flat classification. In 2007, Luciano Barbosa et.al.[7] hasaddressed the problem of identification of the domain of onlinedatabases. Here, new strategy is developed that automaticallyand accurately classifies online databases based on the featuresof Web forms.. Theuse of different classifiers leads to high accuracy, precisionand recall. In 2008, Denis Shestakov[9] has studied the problemof discovering hidden portion Web behind web searchinterfaces. Three classes of problems around the deep Webhas considered: characterization of hidden web, finding andclassifying deep web resources, and querying web databases.

The main technique used for hidden websearching is introduced by Christopher Olston et.al.[6], in2010. He introduced the steps in crawling of deep web-1) Locating sources of web content. 2) Selection of relevant sources.3).Extracting the underlying content of deep web pages.Here is the problem of retrieving unwanted pages whichneeds more time to crawl relevant results. In 2013, Yeye Heyet.al.[10] has a built a system that specializes in crawlingentity-oriented deep web sites. It handles important problemsof deep web crawling such as query generation, empty pagefiltering and URL duplication in specific context of entityoriented deep web sites. In 2015, Feng Zhao, et.al[1]has introduces a crawler which is more efficient to extractrelevant data. This crawler uses hierarchical tree generation toavoid bias nature of crawler. Reverse searching site prioritizingmechanisms are used This crawler needs to crawl large amount ofdata which may lead more time consumption. It also needs toclassify deep web forms to improve accuracy of form classifier.In 2016, Nandhini.G et.al.[2] has introduced crawler whichuses word extraction and text matching concept.

III. SYSTEM OVERVIEW

Fig1 represents detailed architecture of **proposed sysetm. Our aim is to retrieve mo**re relevant pages with respect to user's profession and reuse of log file to reduce search time. This is done by using the algorithm which has the combination of reverse searching and incremental prioritizing algorithm.

- Reverse Searching Algorithm: -

  In reverse searching algorithm, system will collect more number of link URL's , which contain query word. Here System classifies the links into relevant and irrelevant queues.

- Site Prioritizing Algorithm: -

  In this algorithm, system will parse the complete homepage of site in normal search. Whereas, in personalized search, system will parse < title, keyword, description> of site. After that, term frequency of query word in that will be counted. The link which shows more number of query term frequency with respect to predefined threshold value, will be considerd as highest ranked link.

According to that, result will be display.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 6, June 2017

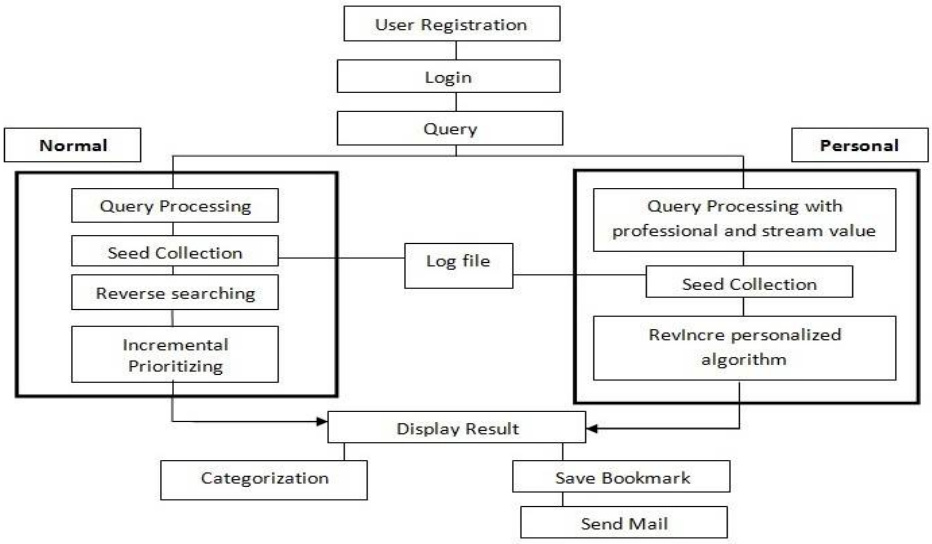


Fig. 1. System Architecture

- **Implementation Steps:**

The implementation steps of system are summarized as follows:

1. While registration, user will enter his email id and profession along with specialized field details.
   For example: profession- java developer & field- computer.
2. User will choose which kind of search he wants.
3. For normal search- at first word will pre-processed and remove stop word. A fined query word will give as input to extract links which contain that query word. Then reverse searching algorithm is applied on it to extract more number of links. After extraction, site prioritizing algorithm is used to parse the page and find the most relevant pages by considering query word frequency in homepage of that links. But it takes more time to parse complete homepage. Relevant sites will display to user.
4. For personalize search- at first query word will be pre-processed and removal of stop word. Here for further processing, combinations of words are considered by system.
   Input query word = query word + profession & query word + field.

5. In personalize search , all links which contain these combinations will extracted and parsed the url, title and description of link to find frequency of user query word by using prioritizing algorithm. Most relevant links are displayed to user.
6. In both searching process, user can check the categorization of links with respect to source domain. User can save bookmarks and send it in mail to himself.

IV. EXPERIMENTAL RESULTS

- **Experimental Setup:**
  We have implemented Multi-word Smart Crawlerin Java and evaluated system approach over 2 different Search type as described in fig1.To evaluate the performance of crawlingframework, we compare personalized search resultsto the normal search results.
  Our goals include:
  i. Evaluating the execution time efficiency
  ii. Evaluating the precision of collected sites.
  iii. Evaluating the count of relevant sites with respect to predefined threshold value of term frequency.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 6, June 2017

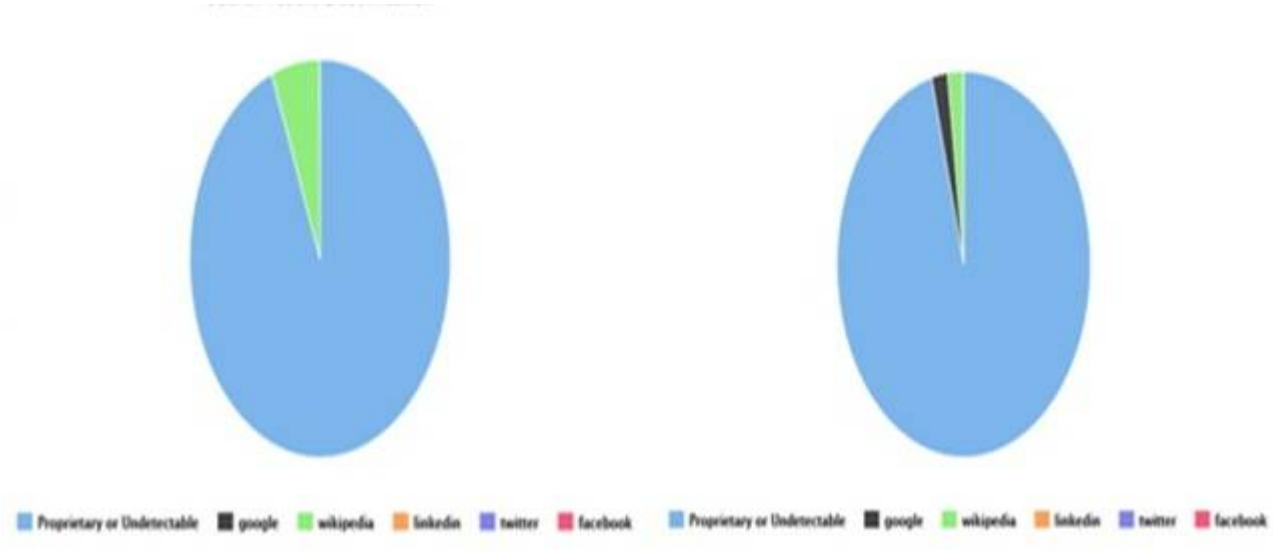
We have performed an extensive performance evaluationof our crawling framework over real web data in two search type with different number of sites extracted by the system as given in table1.

id	query	searchtype	threshold	freq	relevant	irrelevant	etime	pericison
1	class	normal	10	8	8	2	54278	80
2	tiger	normal	20	1	15	5	148398	75
3	tiger	personalize	21	1	21	0	56047	100
4	apple	normal	20	1	19	1	141227	95
5	object	normal	50	3	45	5	412598	90
6	object	personalize	50	3	48	2	60235	96

Table1- Result set on different queries.

In this table1, different number of links extracted by the crawler which is consider as “threshold” for various query words by using different search type. Here frequency is considered as predefined threshold value for term frequency which used to divide links into relevant and irrelevant categories.

In fig2, we compare the categorization of result. Here we can see, as compared to normal search, personalize search gives results from various source domain. Personalize search extract more number of links from Google as compared to normal search.



(a) (b)

Fig.2.Links Categorization With Respect To Source Domain (a) Normal Search (b) Personalize Search

In fig3, we compare the execution time for complete search. In normal search, time required to parse complete homepage in site prioritizing algorithm is more as compared to parsing in link description only. Therefore it requires more execution time. Though personalized search doesn’t parse complete page, still it gives almost same quality of sites in return in addition of more relevant site.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 6, June 2017

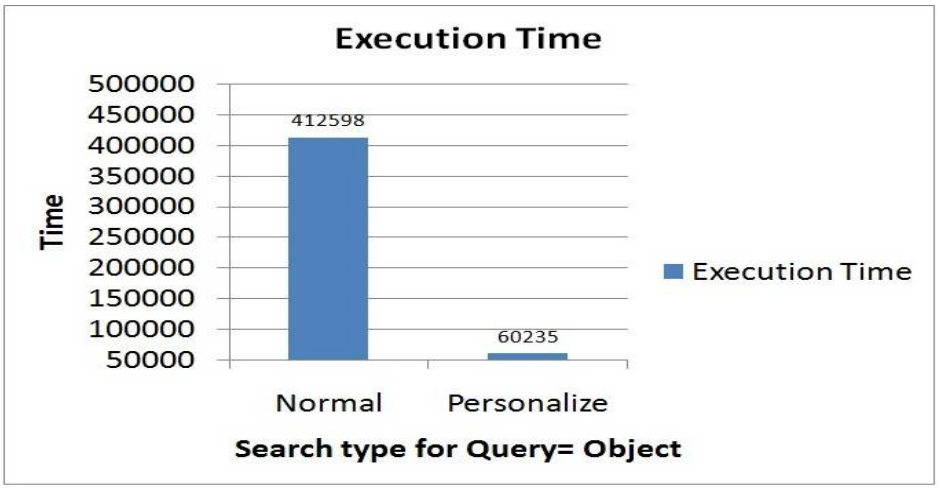


Fig. 3Execution Time for Query= Object

In fig4, we compare the number of relevant sites retrieved in searching process and the precision value for query="object" in different search type. As we can see, the more number of relevant links harvested in personalize search. In personalize search, we passed various combination of query word with profession and field, so that we get more number of relevant links. As we get more number of relevant links in personalized search, so that we get more precision value for it.

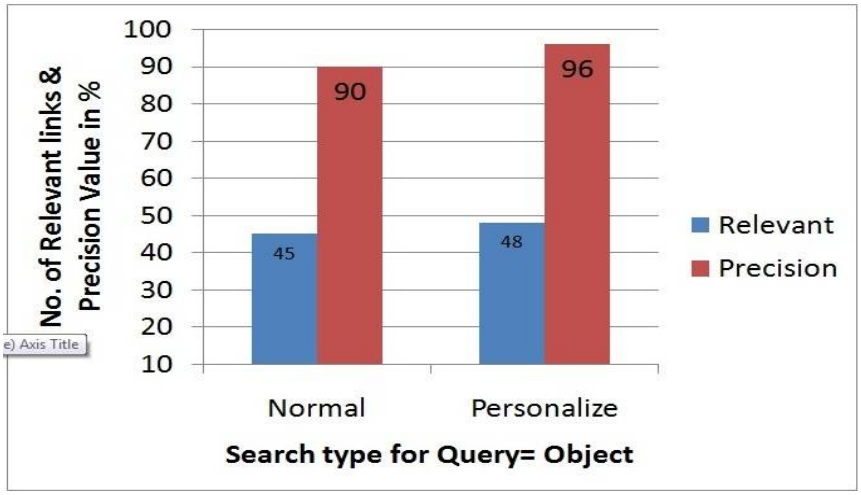


Fig. 4:- Number of Relevant Links and Precision Value for Query= Object

All these results show that, we get more number of relevant links within short execution time by using personalize search.

V. CONCLUSION

Here, System is able to retrieve more number of relevant links by using personalized search. In this system, reuse of log files decreases system time for searching within a day with same query word and same number of links extracted. By using user's information about his profession and stream, system will retrieve more number of pages relevant to it. User can save the bookmarks and can store it into a file which can send on email id of that user.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 6, June 2017

REFERENCES

[1] Feng Zhao, Jingyu Zhou, Chang Nie, Heqing Huang, Hai Jin ,"SmartCrawler: A Two Stage Crawler for Efficiently HarvestingDeep-Web Interfaces",IEEE , VOL. 99, 2015.
[2] Nandhini.G, Murali.D, Kumaresan, , "A Multi-Word CrawlerHarvesting For DWI" ,International Journal of AdvancedResearch in Biology Engineering Science and Technology(IJARBEST) Vol. 2, March 2016.
[3] Sonali Gupta, Komal Kumar Bhatia, "A Comparative Study of Hidden Web Crawlers", International Journal of Computer Trendsand Technology (IJCTT) Vol. 12, Jun 2014.
[4] Dr. Khanna, Samrat, Vivekanand, Omprakash, "Concept of SearchEngine Optimization in Web Search Engine", International Journalof Advanced Engineering Research and Studies(IJAERS) Vol. I,October-December, 2011.
[5] Sriram Raghavan, Hector Garcia-Molina, "Crawling the Hidden Web", in Proc. of the 27th VLDB Conf., Italy, 2001.
[6] Olston and M. Najork, "Web Crawling, Foundations and Trendsin Information Retrieval", in vol. 4, No. 3, pp. 175246, 2010.
[7] Luciano Barbosa, Juliana Freire, "Combining Classifiers to IdentifyOnline Databases" , In Proc. Int. Conf. Committee (IW3C2), May2007.
[8] Panagiotis G. Ipeirotis, Luis Gravano ,"Distributed Search Over TheHidden Web: Hierarchical Database Sampling and Selection",InProc. of the 28th VLDB Conf., Hong Kong, China, 2002.
[9] Denis Shestakov, Tapio Salakoski, "Search Interfaces on the Web:Querying and Characterizing", in TUCS Dissertations, May 2008.
[10] Yeye He, Dong Xin, Venkatesh Ganti, Sriram Rajaraman, andNirav Shah, "Crawling Deep Web Entity Pages", In Proc. Int. Conf.on Web search and data mining, ACM, 2013.