I J C T A, 9(11) 2016, pp. 5357-5368

EVOLUTION OF WEB CRAWLER ITS CHALLENGES

Gitika Sharma[*], Sumit Sharma [**] , and Heena Singla[***]

Abstract: Due to increase in the amount of information and dynamic nature of web pages, a searching mechanism is required to extract relevant information from the web. Web crawlers are the program to which is helpful to extract relevant hidden information that is of high quality. This paper discusses the brief history of web crawlers and various different available web crawlers. Further the concept of deep web is explained and challenges of web crawlers are discussed. The proposed solution for the improvement of deep web crawlers is also provided at the end of the paper.

Key Words: web crawler, types of crawling, crawling algorithm, deep web, Supervised Learning

1. INTRODUCTION

In present scenario, internet is very important part of our life. Now a day's use of internet is very fast growing .World Wide Web (www) provides different types of information. Every user depends on the search engine to complete his desire to get information. Every day information is updated and changed on search engine. Number of web page is increased day by day. Search engine depend upon the crawlers to get more relevant information. Web crawler is most important method to collecting the data and keeping up to data. It is automatically discovering web page or downloading documents.web mining also plays important role in extracting information.web crawler used only for search engine to collect information and index. It is a computer program that browses the Web pages in a methodical and automated manner. In 1994, Brian Pinkerton, a Computer Science and Engineering student at the University of Washington [20]. The Web Crawler aims at discovering the web pages of a web application by navigating through the application. It is also known as robots, web spiders, worm and ants.

1.1 Basic Crawling terminology

Before we discuss the Crawlers, it is worth to explain some of the basic terminology that is related with crawlers.

[*] Department of computer science, Chandigarh University, Gharuan-140413, Mohali, India
Email: gitikasharma41@gmail.com
[**] Department of computer science, Chandigarh University, Gharuan-140413, Mohali, India
Email: sumit_sharma@mailingaddress.org
[***] Department of computer science, Chandigarh University, Gharuan-140413, Mohali, India
Email: singlaheena77@gmail.com

Crawler frontier

The crawling process methods start with a given URL (seed) and extracting links from it. Crawler creates a list of unvisited URLs from frontier.

Seed Page

Crawling is the process to traverse the links recursively. It starts with a list of URL to visit called "Seed Page". The selection of a good seed is the most important factor in any crawling process

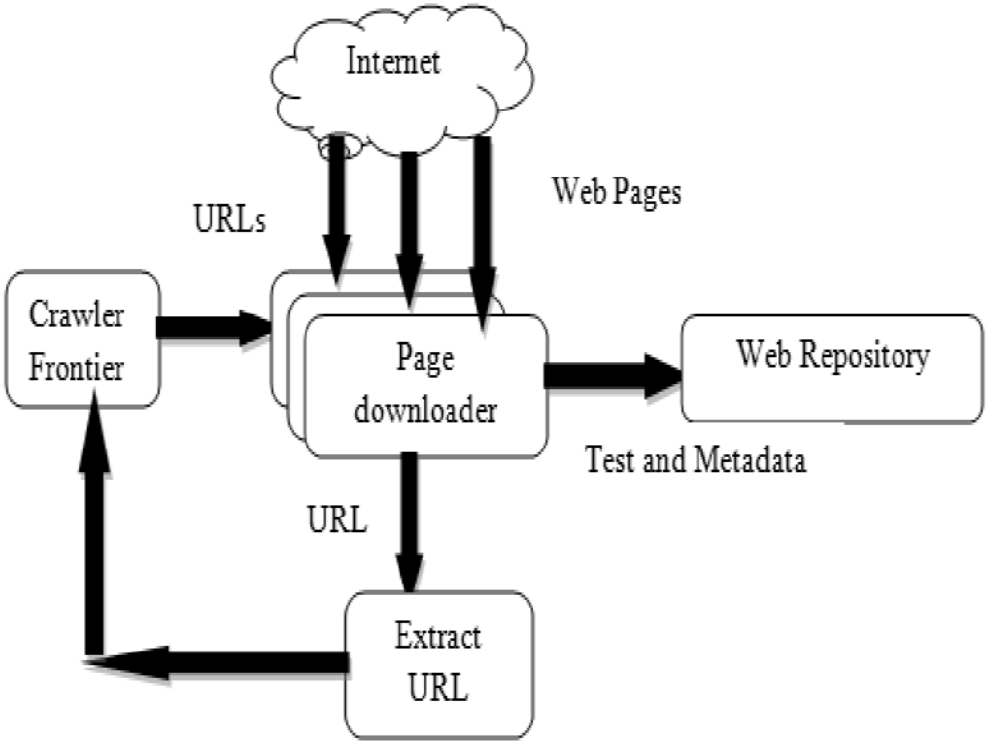

Figure 1. Architecture of web crawler [1]

Parser

Once a page has been fetched, we need to parse its content to extract information this information will guide the future path of the crawler. Parsing may imply simple hyperlink/URL extraction

Examples [3], [4] of web crawlers are:

- ***Yahoo! Slurp*** was the name of the Yahoo! Search crawler,
- ***Bingbot*** is the name of Microsoft's Bing WebCrawler
- ***FAST Crawler*** is a distributed crawler
- ***PolyBot*** is a distributed crawler
- ***RBSE*** was the first published web crawler
- ***WebCrawler*** was used to build the first publicly available full-text index of a subset of the Web
- ***Googlebot*** is the name of the Google search crawler etc

1.3 Various existing crawling algorithms [10]:

Breadth first search

The main motive of this algorithm is to search uniformly according to the neighboring URL present at the same level and Algorithm begins with the root URls and searching the entire neighboring URL at the same level. If the goal achieves then the result of reports success and search terminates. When all the URLs searched step by step or scanned, but objective is not complete then its report as failure. Breadth first search algorithm is used where the objective is complete on the depthless path in a deeper tree [13].

Andy yoo et al [14] proposed a distributed BFS for numerous branches using Poisson random graphs and achieved high scalability through a set of clever memory and communication optimizations

Depth first search algorithm

In this case of search algorithm it begins with the root Node and traverse deeper through the child node. If there are more than one child then priority is given to the left child then backtracking process is used to go further unvisited node and processes is repaid in similar manner [15]

Page rank algorithm

Page rank algorithm is used to understand the web pages by counting the back links to the given page. It is provided a web page to calculate the relationship (relatedness) between the web pages. The page rank of a given page is calculated as

PR (A) = (1-d) + d (PR (T1)/C (T1) + ... +PR (TN)/C (TN))

PR (A) - Page Rank of a Website,

D - Damping factor

T1….Tn – links, J.Kleinberg [10] proposed a dynamic page ranking algorithm.

Naive Bayes classification Algorithm

It represents as supervised learning method **.**It is based on Probabilistic learning. This algorithm proved to be efficient over many other approaches [18] although its simple assumption is not much applicable in realistic cases [17]. Peter Flach and Nicolas Lachiche [19] presented Naive Bayes classification of structured data on artificially generated data.

Focused Crawling algorithm-

It is a web crawler that tries to download the web pages which are belong to each other and it is based on the similarity of the page to a given query. It collects the document which is relevant or specific to topic [13].

1.4 Working of basic crawler

- Began the seed URLs
- Adding it to the frontier
- Select URls from the frontier

- Fetching the web page
- Parsing the retrieved page
- Add the all the unvisited links into the frontier

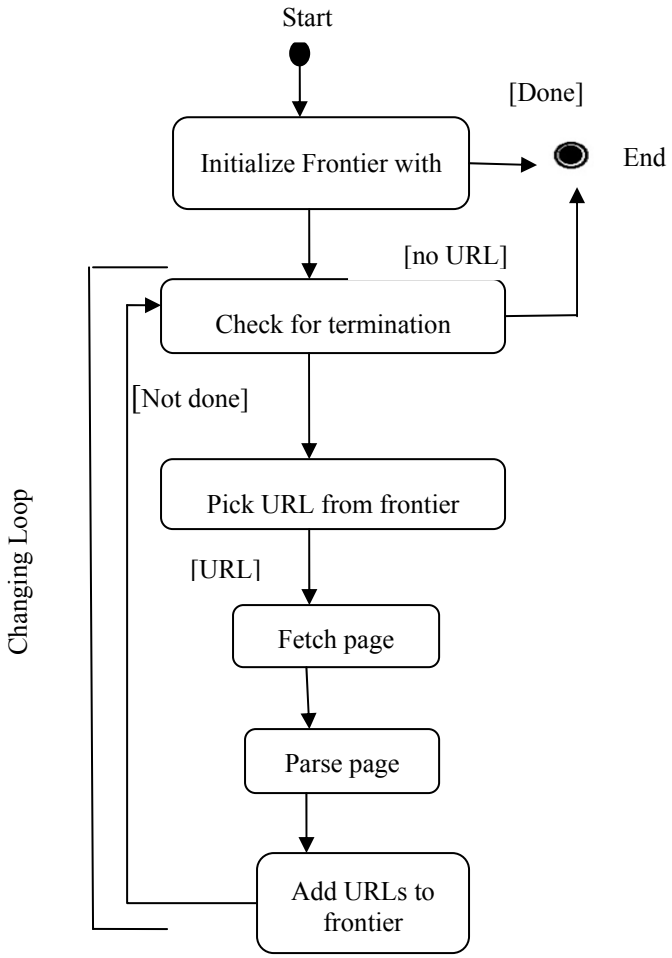

Figure 2. Working of basic crawler [1]

WebCrawler process is to traverse the links recursively, adding URLs into the frontier (database).

The Paper itself is structured as follows. Section II provides types of web crawlers. Section III discuss about history of search engines. Section IV discuss about concept of deep web. Section V related work in the field of web crawler. Section VI provides proposed solution for problem facing in deep web using supervised learning approach. Finally, Section VII summarizes the conclusion.

2. TYPES OF WEB CRAWLER

Here explained types of web crawler [11] [12] in table 1

Table 1
Types of web crawler

Name	*Description of web crawler*	*Advantage of web crawler*
Focused web crawler	In general it is used for searching information related to specific topics from the internet.	It only collects the documents which is specific or relevant to the given topic
Incremental web crawler	In this of crawler, refresh its collection, periodically replaces the old documents with the newly downloaded documents	It provides valuable data to the user.
Parallel web crawler	Parallel crawler work simultaneously to grab the web pages add to the central repository of the search engine	It is very vital from the point of view of downloading documents in a fair amount of time
Distributed web crawler	Due to increase the size of web and dynamic nature of web. With multithreading a single crawling process insufficient for the situation. In that case the process needs to be distributed to multiple process to make the process scalable.	It increases the overall download speed and reliability and reduces the hardware requirements

3. HISTORY OF SEARCH ENGINES

Today, the average web surfer can search for a number of differently keywords related to specific topic and still come up with a plethora of great information from a myriad of sources. However World Wide Web wasn't always so easy to navigate. Here is a complete history of the evolution of search engine. History of search engines is shown in table 2.

Table 2
History of Search Engines

Year	*Name of search engine*	*Description*
1990	Archie	Archie is the first search engine. Alan Emtage, a student at McGill University in Montreal. It helps to download the directory listings of all the files which is located on public anonymous FTP(File Transfer Protocol)sites and it creates files name with searchable data base
1991	Gopher	It creates two new search programs veronica (**V**ery **E**asy **R**odent-**O**riented **N**et-wide **I**ndex to **C**omputerized **A**rchives) gives a keyword search of most Gopher menu titles in the entire Gopher index system & jughead (**J**onzy's **U**niversal **G**opher **H**ierarchy **E**xcavation **A**nd **D**isplay) same as the Veronica menu information from gopher serves .it also obtaining the. it show plain text while Archie indexed computer file
1992	Virtual library of the web(VLib)	Timothy Berners-Lee creates the Virtual Library (VLib)& HTML & Web itself, in 1991 at CERN in Geneva.

Table 2 Contd…

1993 (Feb)	Excite	It involving six undergraduates students at Stanford to improve the relevancy of words relationships searches on the internet W3 (World *Wide Web Wanderer*) Catalog, written by Oscar Nierstrasz at the University of Geneva, it is the web search engine. it was used to obtain URLs, forming the first database of Web sites called *Wandex.*
1994 (January)	Infoseek	InfoSeek was a pay-for-use service started in January 1994
1994 (January)	AltaVista	It follows natural language queries & people add or delete their domains in 24 hours
1994 (April)	Web Crawler	Web crawler was the first search engine which explains the full text index surface of the web. It was created by Brian Pinkerton at the University of Washington.
1994	Yahoo search!!	David filo and jerry yang created yahoo directory in 1994
1995	Look smart	Complete with yahoo by increasing inclusion rates back and forth
1996 (January)	Google	Larry and Sergey began working on Backrub search engine which utilized back links for search It ranked pages using citation notation.
1996 (May)	Inktomi:Hotbot	It is a search engine Hotbot or it is listed on Hotwrite
1997 (April)	Ask.com/Ask Jeeves	It launch of a natural language search engine. It is powered by Direct Hit, which aimed to rank links by popularity.
1998	MSN	MSN search launches. Launched preview of new engine in july 2001. Dropped yahoo's search and program on may 2006
1998	dmoz	Open directory project –It is directory to download Largest internet directory run by volunteer editors
1999 (May)	AllTheWeb	interface with advanced features
2005 (October)	Snap	Show search volumes, revenues and advertisers
2006 (September)	Live Search	Live search launched by Microsoft
2008 (June)	Cuil	Managed and developed by former Google employee
2009 (June)	Bing	It is related to search directly in result set and rebranding of MSN/live search

4. DEEP WEB

Deep web is called invisible web. Deep web is not indexed in search engine easily. It is hard to access or it is big challenge in web community. It is opposite to the surface web. Surface web is a visible content web which is easily available in search engine. Size of the deep web is larger than the surface web [2]. Actual size of the deep web is impossible to measure. It is the way to find the information behind the HTML many resources are added or updated day by day. Currently over 85,000 Deep Web sources, grouped by source type. Examples of source type include-hospitals,

banks and social media. Some search engine is used to find the hidden web like Surf Wax, deep peep, and Nations line [3].Deep web as shown in figure 3.

4.1 Step of deep web crawling [2]

Locate deep web content sources

In this case crawler identify web sites contain from interfaces that lead to deep web.

Select relevant Sources

In second case we select only relevant content in available sources

Extract Underlying Content

In the final step content is extracted

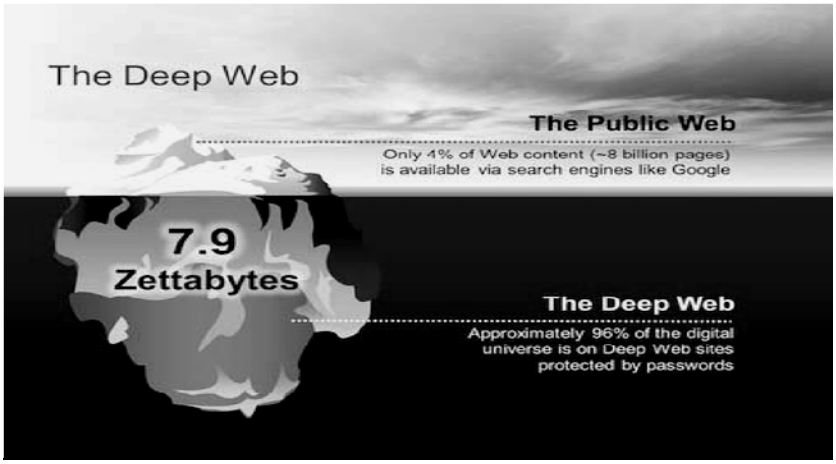

Figure 3. The deep v/s surface web [21]

5. RELATED WORK

With the help of in literature survey we gives an over view of web crawlers how it works and how it comes from.

WEB CRAWLER: In 1994, Brian Pinkerton, a computer Science & Egg. Student at the University of Washington creates web crawler. It is the first search engine which gives full text indexed. Brian generated 25 websites on March 1994 &one month later Brian announced the release of web crawler& 400 websites in database.Nov 14, 1994 it serves one millionth query till the end of the year web crawler signed two sponsers, Dealer Net & Starwave.

Basically it starts from the list of URL called seeds it visits the seed or identifies the hyperlinks and the list of URLs called crawler frontier**.**

Mejdl S. Safran, Abdullah Althagafi et.al. [7] In this paper author presents the new leaning approach that uses four attribute to find the unvisited URls .Focused crawler is used to specify which link is relevant or irrelevant. Or help to predict the relevancy. Soumen Chakraborty, Martin van den Berg et.al [9] this paper explains about the focused crawler .how it is work or extract topic specific web discovery. With the help of relevant link. The topics are specified using exemplary documents not the keywords. Luciano Barbosa, Juliana Freire et.al [8] studied about dynamic nature of web, high interest of information retrieval and integrates of deep web to extract the high quality data or in structure manner. Understanding the deep web [6] Deep web (invisible or hidden

web) it is the challenging problem how information is extracted in deep web .Deep web cannot be indexed in search engine easily. It is alternate to the surface web which is easily index able or accessible. Size of the deep web is large as comparison to surface web. Distribution of deep websites based on the material [3]Gang Liu, kai Liu [5] studied about deep web entries & firstly using the information of specific field DeepWeb entry form to establish domain ontology, and then web forms can be automatically judged by the process of the crawling.

Feng Zhao, Jingyu Zhou, Chang Nie [2] in this paper explains two stages: efficient site locating and balanced in-site exploring. Effective harvesting framework for deep web interfaces. This paper achieves both wide coverage for deep web interfaces and highly efficient crawling.

Conclusion of literature survey it is challenged to locate the deep web databases, to overcome the problem of deep web. To improve the deep web crawler using supervised learning approach. With the help of this approach .by using classifier like artificial neuron network, naïve Bayesian, decision tree. We find the relevancy classifier to classify the queryable (deep web) interfaces on the basis of relevant or irrelevant links. We use the TEL-group of UIUC repository. It includes 689 query forms covering 12 domains. To achieve high harvest rate and relevant accuracy we need to improve deep web crawler so in next section proposed solution is given.

6. PROPOSED SOLUTION

6.1 TEL Group Dataset

In this use the TEL group dataset of UIUC repository [16] is used in proposed methodologies which have 12 domains with 689 query forms to input.

6.2 Train the relevancy classifier

In this case start to search these links which is present in the dataset. Following steps helps to train the relevancy classifier.

Internet

The search Query is put user in the search box. The query is URL encoded & the sent to Google search The Google search page with result is received

Tokenization & stop word removal

In these cases remove symbols, punctuation marks**,** Numbers, lower case and then clear the words list.Futher remove the stop words and stemming. To achieve keyword list

Td-Idf (term frequency inverse document function)

This helps to calculate frequency TF-IDF is calculated by multiplying term frequency with inverse document frequency. In general Formula to calculate TF*IDF

$$\begin{aligned} TF &= \text{No. of times term occur/Total term} \\ IDF &= \log(\text{total document/document Contains Specific term}) \\ TF*IDF &= \text{No. of times term occur/Total term} * \log \\ &\quad (\text{total document/document Contains Specific term}) \end{aligned} \quad \ldots(1)$$

Decision tree, artificial neural network, Naive Bayes etc can be used to train the relevancy classifier and they widely used in literature, so these algorithms can be used and there comparison based on accuracy can be done to find out which classifier will give better result.

6.3 Select new page

Use back links to find the new page to crawl.

6.4 Search for queryable interface

Mark these pages are relevant or irrelevant. A relevant page helps to find the queryable (relevant) form. With the help of query to query interface. The response of the interface is parsed for relevance.

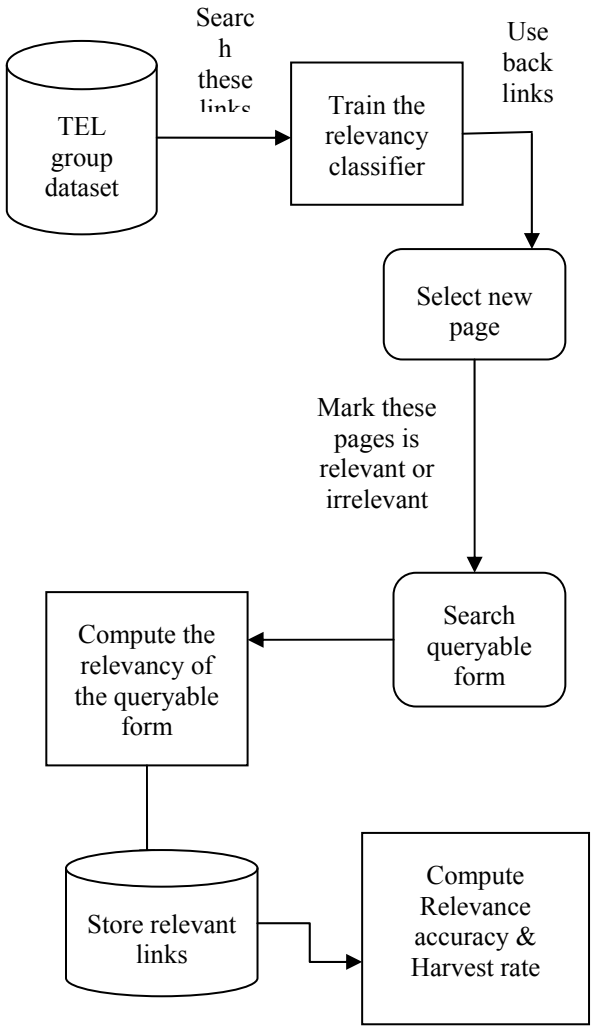

Fig.no.4 Proposed solution of deep web crawler

6.5 Compute the relevancy of the queryable form

Compute the relevancy of this queryable interface based on the classifier trained from the dataset if the response is found to be relevant, we will add this link to the database, else the link is discarded

6.6 Store relevant links

Store link of each relevant queryable interface in the database

6.7 Compute the accuracy relevance & Harvest Rate

Harvest rate is the number of links added to the database per unit time. (Unit time can be hour or minute)

6.8 Accuracy relevance

store relevant links queryable interface in the database

7. CONCLUSION

This paper explains about the evolution of the web crawlers, different algorithm and types of web crawlers. In summary, web crawler is a Meta search engine which blends the top most results, basically crawling is a program to crawl the web pages and to create the entries for a search engine index. It is also called web spider or robot. The deep web is called hidden web and it is not index able easily. In this paper a propose solution is given to improve the deep web using supervised learning approach.

References

[1] Udapure, Trupti V., Ravindra D. Kale, and Rajesh C. Dharmik. "Study of web crawler and its different types." *IOSR Journal of Computer Engineering (IOSR-JCE)* 16 (2014): 01-05.

[2] Zhao, Feng, Jingyu Zhou, Chang Nie, Heqing Huang, and Hai Jin. "SmartCrawler Two-stage Crawler for Efficiently Harvesting Deep-Web Interfaces." (2015).

[3] Agrawal, Smita, and Kriti Agrawal. "Deep Web Crawler: A Review."

[4] Singh, Mini Singh Ahuja Dr Jatinder, and Bal Varnica. "Web Crawler: Extracting the Web Data."

[5] Liu, Gang, Kai Liu, and Yuan-yuan Dang. "Research on discovering Deep Web entries based ontopic crawling and ontology." In *Electrical andControl Engineering (ICECE), 2011 International Conference on*, pp. 2488-2490. IEEE, 2011

[6] Iffat, Rabia, and Lalitha K. Sami. "Understanding the Deep Web." *Library Philosophy and Practice (e-journal)* (2010): 364.

[7] Safran, Mejdl S., Abdullah Althagafi, and Dunren Che. "Improving relevance Prediction for Focused Web crawlers." In *Computer and Information Science (ICIS), 2012 IEEE/ACIS 11th International Conference on*, pp. 161-166. IEEE, 2012

[8] Barbosa, Luciano, and Juliana Freire. "Searching for Hidden-Web Databases." In *WebDB*, pp. 1-6. 2005

[9] Chakrabarti, Soumen, Martin Van den Berg, and Byron Dom. "Focused crawling: a new approach to topic-specific Web resource discovery."*Computer Networks* 31, no. 11 (1999): 1623-1640

[10] Pavalam, S. M., SV Kashmir Raja, Felix K. Akorli, and M. Jawahar. "A survey of Web crawler algorithms." (2011).

[11] Kausar, Md Abu, V. S. Dhaka, and Sanjeev Kumar Singh. "Web Crawler: Review." *International Journal of Computer Applications* 63, no. 2 (2013)

[12] Singh, Mini Singh Ahuja Dr Jatinder, and Bal Varnica. "Web Crawler: Extracting the Web Data."

[13] Singh, Apoorv Vikram, and Achyut Mishra Vikas. "A review of web crawler Algorithms." *International Journal of Computer Science & Information Technologies* 5, no. 5 (2014): 6689-6691.

[14] Andy Yoo, Edmond Chow, Keith Henderson, William McLendon, Bruce Hendrickson, ÄUmit CatalyÄurek "A Scalable Distributed Parallel Breadth-First Search Algorithm on BlueGene/L" ACM 2005

[15] Ben Coppin ―Artificial Intelligence illuminated‖ Jones and Barlett Publishers, 2004, Pg 77.

[16] The UIUC web integration repository [Online].Available: http://metaquerier.cs.uiuc.edu/repository/, 2003.

[17] Harry Zhang "The Optimality of Naive Bayes" American Association for Artificial Intelligence 2004.

[18] Rich Caruana, Alexandru Niculescu-Mizil "An Empirical Comparison of Supervised Learning Algorithms" Proc 23 rd International Conference on Machine Learning, Pittsburgh, PA, 2006.

[19] Peter A. Flach and Nicolas Lachiche "Naive Bayesian Classification of Structured Data" Machine Learning, Kluwer Academic Publishers

[20] The History of Search Engines-An Infographic [Online].Available:www.wordstream.com/Articles/internet-search-engines-history

[21] Seven things you didn't know about the DeepWeb [Online].Available:www.cso.com.au/Slideshow/579375/seven-things-didn-t-know-about-deep-web/