NOVATEUR PUBLICATIONS
INTERNATIONAL JOURNAL OF INNOVATIONS IN ENGINEERING RESEARCH AND TECHNOLOGY
[IJIERT] ISSN: 2394-3696 Website: ijiert.org
VOLUME 8, ISSUE 9, Sep. -2021

SEED SELECTION BASED WEB CRAWLER FOR WEB PAGE CLASSIFICATION: A SURVEY

Vikash Kumar,
M. Tech Scholar, Computer Science & Engineering,
Technocrats Institute of Technology, Bhopal (M.P.), India
vikashkumar17dec2013@gmail.com

Dr Yogadhar Pandey
Professor, Computer Science & Engineering, Technocrats Institute of Technology,
Bhopal (M.P.), India

ABSTRACT
A web search engine is a three-phase method of which the first phase is Web crawling. Web crawler works to collect data for search engines. Web crawler Collects pages through hyperlinks exist on web pages; these pages are elements of Seed URLs set that defined in the initial stage of the web crawling process. The search engine performs the ranking algorithm on a set of web pages that are covered through crawlers. Therefore the result of the search engine is typically based on web crawler coverage. Most research is already done in the web crawler area. Different web crawling strategies, resulting from different ways of ordering the URLs in the frontier, can explore the web in different ways. Researchers have also studied the different issues and challenges that web crawlers faced. As per literature web crawlers performance depends on the Seed URLs. This paper gives a bird eye over Web crawling methods, recent research over it and possible research gap.

Keywords: WWW, Machine Learning, Deep Learning, Seed URLs, Web Crawling, Search Engines

INTRODUCTION
In the last few decades, the World Wide Web has developed from a little research venture into an immense store of data and another medium of correspondence. The Web is a humankind created data repository on the largest scalable for the public. It is required for users that they must be knowledgeable to extract quality data from this enormous data ocean. One important issue related to quality is partial data, which exists in different forms. The web is a system of substance and hyperlinks; with over a billion interlinked "pages" those are resultant of ungraceful activities of a huge number of users. These are 3 types of web data: Structured data, Semi-structured data, and Unstructured data.

Structured data – Data in which all components are accessible for effective analysis is known as structured data. The organization of the structured data is a formatted repository that is a regular database. Web structured data play the role of information provider and classifier that provide information about a page in a standardized format and classify the content of the page. Google uses structured data that it finds on the web for recognizing the web page content, as well as for collecting the information about the web and the world in general. Google Search also uses structured data for enhancing search and for enabling special search result features.

Semi-structured data – Semi-structured data is information that does not reside in a rational database but that have some organizational properties that make it easier to analyze. Semi-structured data lies somewhere between Structured and Unstructured data. It contains components in which some are structured and others are not. It can be formatted in relation database after operating some process (it can be very tough for some type of semi-structured data), but Semi-structured data is present to make space easier. Example: XML data**.**

Unstructured data – Data that is not organized in a pre-defined manner or does not have a pre-defined data model is known as unstructured data, thus it is not well-suited for a mainstream relational database.

Unstructured data also may be considered as loosely structured data, wherein the data sources exist in a structured form, but the structure of all data may vary.

In view of the decentralized idea of its development, the web has been broadly acceptedto need structure and association. In this way, a considerable measure of examinations of the Web graph depends on hyperlinks which have uncovered a many-sided structure that is turned out to be important and essential for arranging data, this has enhanced the comprehension of inquiry techniques and social setting. The Web contains a substantial, emphatically associated center in which each page can achieve each other by a way of hyperlinks. This center contains the vast majority of the unmistakable destinations on the Web. The rest of the pages can be portrayed by their connection deeply.

A Web search engine is worked as a document retrieval system designed to help in extracting information that has been stored in a computer system, like in the WWW or inside a business or proprietary network or on a personal computer. Several experimental approaches are utilized over a few years to enhance the effectiveness of the search engine. The search engine has become the most practical tool to find WWW.All search engines procedure is a 3 phase approach these phases are: crawling, ranking and returning search results.

Crawling is the process of retrieving pages from the web, extracting the hyperlinks from those pages and following the new links on them. Link analysis algorithms are incredibly useful to guide the crawler to crawl websites that you have no prior knowledge.

There are billions of pages and they modify periodically. New valid pages will be added and old invalid pages can be removed. Meanwhile, the content of these web pages may also show discrepancies every day. In this way, the analysis of hyperlinks will give an excellent leapto the study of search engines [1, 2]. The link analysis plan has competed for an integral role within the ranking functions of this generation of web search engines in the ranking functions of the current generation of Web search engines, as well as Google, Yahoo!, Microsoft's program Bing, and Ask. There are some algorithms on the thought nowadays and being applied to several search engines. The foremost initial ones are Google's Page Rank algorithm [3] and Kleinberg's HITS algorithm[4]. When they have been developed, there are some derivatives and enhancements created supported these two techniques. One in all the algorithmic rules that have an important impact is that the SALSA[5]. There are plenty of variations, however their core plan returns from those three original web search algorithms.

The objective of web crawlers is to check the whole or part of the nodes, which are web pages, on the web graph. Therefore the traversal technique is basically a graph-based search algorithm. A web crawler follows the traversing method based on their specific purposes. For accessing the web graph a general web crawler normally implements three searching methodologies: a breadth-first search (BFS), a depth-first search (DFS) or a focused search. BFS is used as an exploring algorithm for the graph that starts with a selected node then discovers all the connected nodes of it. Then apply BFS for all these discovered nodes to explore their unexplored connecting nodes until it achieves the goal. DFS used as an exploring algorithm forthe graph that starts with a selected node then discovers as far as possible along each branch before backtracking. BFS crawlers required more space to store all traversed pages at every node level than DFS crawlers, which only need to store traversed pages in a single branch of the web graph. DFS crawlers may be at risk of trapping by infinite link loops. BFS crawlers also yield higher quality pages [47].

Web Mining: Overview

Web mining is the new mining techniques to automatically retrieve extract and evaluate the information for knowledge discovery from web data. This is more powerful and efficient than existing techniques. It consists of the following tasks[6]:

1. Resource finding: The procedure of extracting the data by users from either offline or online accessible text on the web is known as resource finding process in web mining.
2. Information selection and pre-processing: This process is used to transform the structureof the original extracted data as per the requirement of the web mining algorithm.
3. Generalisation: This process is used to perform extracting the task of the generic pattern from the variant resources such as either from the distinct websites or across the group of websites automatically.
4. Analysis: This process performs a significant role in pattern mining procedures. It is used to accredit and

interpret the discovered patterns.

In the last few decades, a number of papers [7, 8]have considered the use of web mining to automatically discover and extract information from web documents and services. In general, web mining is divided into three categories:

Web Content Mining: Web Content Mining (WCM) focuses on the discovery of useful information from the contents or data or services available on the web.

Web Usage Mining: Web Usage Mining (WUM) tries to discover useful information from the interactions of the users while surfing on the web. The web contains billions of pages together with links between them. These links create paths from one page to another and allow people to go from one site to almost every other.

Although the web is very rich, it creates a huge challenge for people to retrieve useful information. For people who tried to design a search engine for such information retrieval, the challenge is even higher. The web does not have a designed structure to follow; it consists of text, hyperlinks, and network. Due to the growth rate of information, web structure becomes more and more intricate. How to organize, understand and utilize the structure in order to improve the searching technique is crucial.

Web Structure Mining

The goal of web structure mining is to generate a structural summary of the website and web page. Web structure mining is based on hyperlink analysis to analyze the dependencies between pages to find out the relationship between the references pages to discovered interesting patterns, improve search engine optimization. web structure mining uses the hyperlink structure of the web as an information source. It helps the users to retrieve the relevant documents by analyzing the link structure [9]. This can be divided into two kinds based on the type of structure information used:

Hyperlink Structure: It is related to extract information from a hyperlink in the web and designstructure through analyzing hyperlinks that connect the web pages.

Document Level Structure: It is related to mining the document structure and design a tree-likestructure of the page to describe the use of HTML or XML tag.

The web contains different kinds of objects as web pages, and links are in-out- and co-citation.

Web Crawling and Web Information Extraction

Crawling and information extraction are two fundamental components for almost all web-scale search engines. They are usually the first two steps of a search engine's system pipeline. First, a web crawling system (a.k.a \crawler", \robot", or \spider") traverses the web in a certain manner and provides the raw content (crawled web pages) for search engines. Then, an extraction system is used to understand those crawled pages correctly before they can be indexed and presented tothe end-user.

This proposal includes several ongoing works that attempt to address some fundamental issues that are often encountered in web crawling and web information extraction.

Web Crawler

A web crawler [10-12]is an automated program that starts from an initial set of URLs i.e. referred to as Seed set and downloads all the web pages associated with these URLs. After fetching a web page associated with a URL, the URL is removed from the working queue. The web crawler then parses the downloaded page, extracts the linked URLs from it, and adds new URLs to the list of seed URLs. This process continues iteratively until all of the contents reachable from seed URLs are reached.

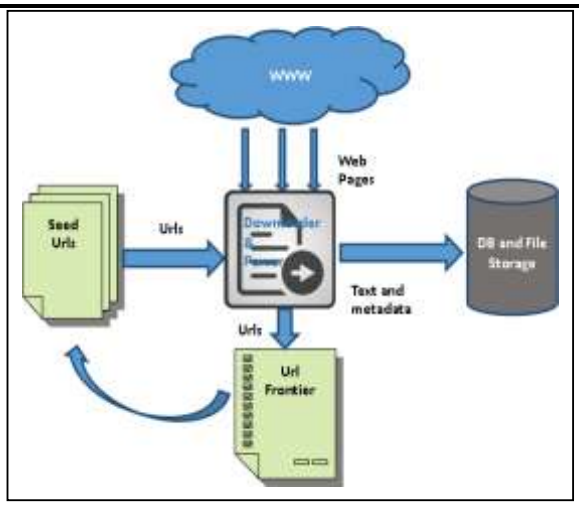

Figure 1 Web Crawler Procedure

Web Crawler Classification

A. **Traditional Crawlers:** Set of seed URLs Nodes are pages with distinct URL and a direct edge exist from page p1 to page p2 if there is a hyperlink in page p1 that points to page p2.

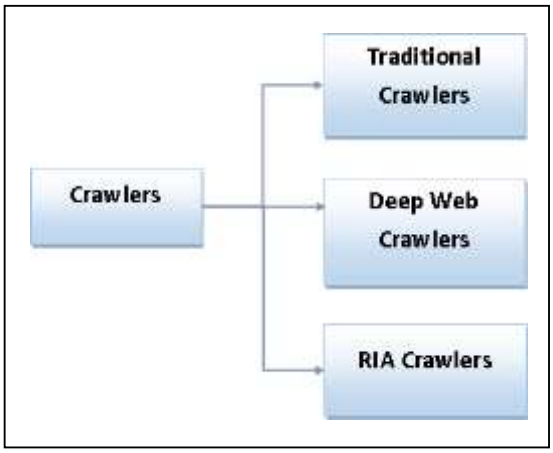

Figure 2 Crawling Taxonomy

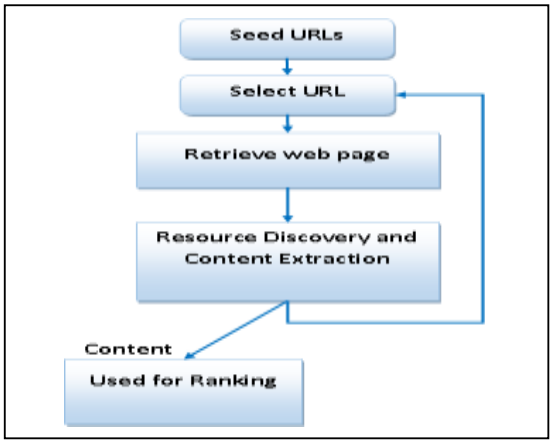

Figure 3 Traditional Web Crawler Process

B. **Deep Web Crawlers:** Deep Web Crawlers: Set of Seed URLs, user context-specific data, domain taxonomy Nodes are pages and a directed edge exists between page p1 to page p2 if submitting a form in page p1 gets the user to page p2.

C. **RIA Crawlers:** A starting page Nodes are DOM states of the application and a directed edge exists from DOM d1 to DOM d2 if there is a client-side JavaScript event, detectable by the web crawler that if triggered on d1 changes the DOM state to d2.

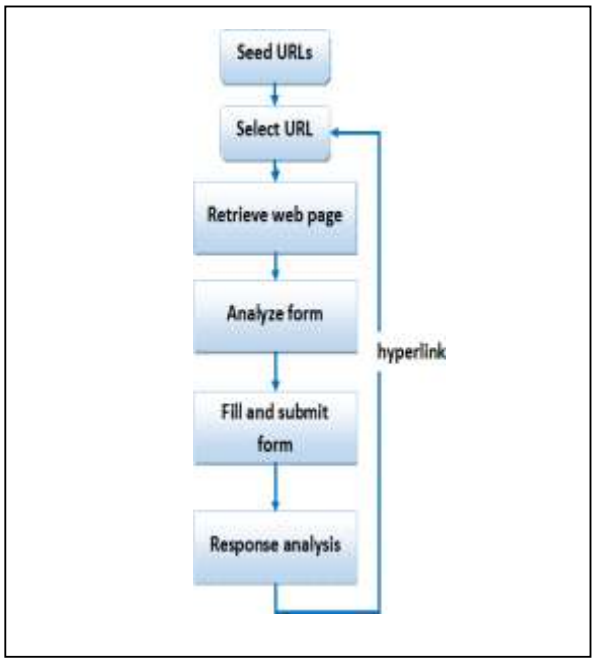

Figure.5 Deep Web Crawler Process

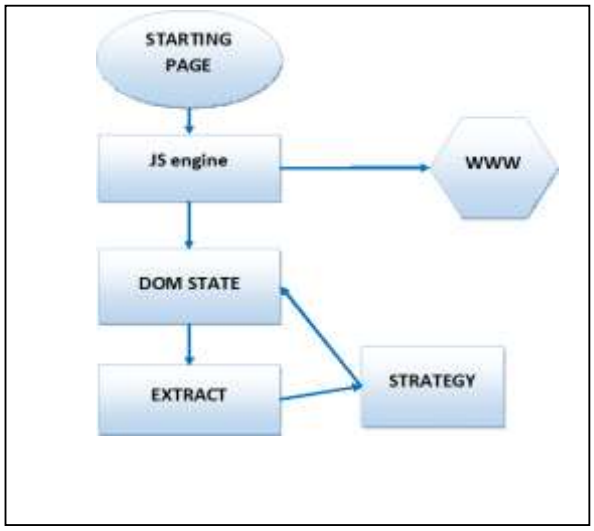

Figure 6 RIA Web Crawler Process

Different crawling strategies resulting from different ways of ordering the URLs in the frontier can explore the web in different ways. However, all of them start from the seed pages and proceed by exploring the neighborhoods of the seed pages in one way or another. Thus, to a large extent, selecting good quality seed determines the quality of the crawl.

One may think that simply starting from root pages of several well-known sites and crawling very deep will allow the crawler to reach all useful pages on the web. Unfortunately, this is not so. As Brodart. al [13] showed, even 9 years ago, close to half of all web pages could not be reached from the "central" strongly connected portion of the web. Moreover, the situation is likely to be even worse nowadays. Recently, many websites that contain millions of pages have emerged. Our study shows that many large

websites are not strongly connected.

Applications of Web Crawler

a. Web Search Engine: Web Search engines process is divided into three components:
- Crawler: It collects web pages using hyperlinks on the web page.
- Indexing: It generates a ranking value for all visited web pages.
- Query Result: It gives search results as per the user's query.

b. Web Achieving: Web archiving is the way of gathering parts of the World Wide Web and certifying that the collection is protected as an archive, such as an archive site, for further research, historians, and the public. Because of the huge size of the Web, ordinarily, web crawlers are used by web archivists for automated collection.

c. Vertical Search Engine: The vertical search engine is referred to as a topical search. It aggregates data from many sources on specific topics.

d. **Web Data Mining :** Web mining used web crawlers, to analyzed web pages for different assets like statistics, and data analytics are then performed on them.

e. Web Monitoring: In web monitoring, a web crawler is used to observed sites/pages for changes and updates.

f. Detection of Malicious web sites : A web crawler is used to collect web pages for further tasks so it is also able to identify maliciously. Therefore Web Crawler is also worked as an identification algorithm to detect malicious Web pages based on the content**.**

g. Web site/application testing: A web crawler is used to understand web applications through automated crawling analysis. It was also used to analyze any issue spotted during the website's liability testing. Crawlers canalso provide services to automate Web site maintenance tasks, for example, check links or validate HTML code.

h. Fighting crime: A web crawler is also working for webspam and deceitful web site detection as well as it is used to find unaccredited use of copyrighted content.

i. Web scrapping: The determination of scratching is to gather data from diverse web pages and after that stock it to the local database. Therefore web scraping is done in the manner of extricating required information from web pages.

j. **Web Mirroring:** In general, replicated collections constitute several hundred or thousands of pages and are mirrored in several tens or hundreds of web sites. The web crawler can be used to create a site mirror.

ISSUES OF WEB CRAWLER

- **Duplicate Pages:** The problem with a web crawler arises when a web page exists with multiple times under different URLs. Because the web crawler methodology is following the strategy that a unique URL (Unique resource locator) corresponds to a unique webpage.

- **Mirror Sites:** Web crawler facing the problem of identifying cloned documents and hyperlinked documents collection efficiently. Because it affects the improvement in the performance of functions, used in search engines and perform on web data, such as web crawlers, archivers, and ranking.

- **Identifying Similar Pages:** Possibility of downloading the same page multiple times by different crawling nodes download is high. It is undesirable to downloads the same page multiples. Thus the

challenge is to improve techniques so reduce or eliminate these overlapping of pages.

- **Deep Web:** A large part of the Web is "hidden" behind search forms and it is required user interaction to transfer the next stage, by filling these forms. Such pages are often referred to as Hidden or Deep Web. So it is difficult for web crawlers to reach this portion of the web content by following the links.

- **When to stop:** It is facing the major issue that is to restrict the overall size and depth of a crawl either before crawler start or during the crawling process.

- **Incremental Crawler:** The web collection is frequently changed because of creation and removal of pages constantly and a portion of the new pages can be more significant to existing pages, the crawler should build the nature of the local web collection by supplanting less significant pages with increasingly significant ones.

- **Refresh Policies**: The crawler must also decide the revisiting policy of the already traversed pages, in order to keep its client informed of changes on the Web.

- **Evolution of the Web**: Web mining behavior has been temporarily classified into three types of web data: web content mining, web structure mining, and web usage mining. The researchers have to be work on discovering the growth of web content, web structures, and web communities, authorities, etc. in extracting temporal models.

- **Crawling the "good" pages first:** It is beneficial for the crawling process to trace "valuable" pages first with the goal that visited (and stayed up with the latest) part of the Web that is more important.

- **Focused Crawling**: Focused crawlers are expected to crawl relevant pages that give an immediate benefit but the issue is those might be missed during crawling.

- **Distributed Crawlers:** Distributed crawler face the problem of URLs assignment efficiently and dynamically to download among the crawling agents, as the major challenge.

- **Crawler-Friendly Web servers**: A web crawler is treated in the same manner, as a web server, by a web server. This offers to ascend to few issues like freshness, incomplete data, and performance impact on web site, etc. [14].

Related work

Web crawlers-also known as robots, spiders, worms, walkers, and wanderers- are almost as old as the web itself. The first crawler, Matthew Gray's Wandered, was written in the spring of 1993, roughly coinciding with the first release of NCSA mosaic . Web crawler has become an increasingly important application in recent years, because of its unique ability to search and store web page over the internet.

Wang et.al. [1] present probabilistic model for relevant web page selection. Crawler employed the TF-IDF algorithm to quote the feature of page content and Bayes classifier to evaluate page rank.Whereas, Saleh et al. [2] present a probabilistic domain distiller for a focused crawler. Domain distiller combines SVM, naïve Bayes, and genetic algorithm (GA) and present optimized instance of probabilistic classifier, i.e., optimized Naïve Bayes (ONB) classifier. Where initially, GA optimized the marginal distance of different class labels for vector space, and SVM alienates the outlier's keywords. NB scrutinized the ambiguous nature of the outlier domain keyword and finally proposed a disambiguation model for classification.

Yajun Du et al. [3]\cite{19} present a semantic focused crawler that combines document frequency, semantic similarities, and vectors. This semantic vector mapped double-term set and evaluate the cosine similarities between anchor text to extract similar documents and anchor texts of unvisited hyperlinks. Manish kumar et al. [4] presents a keyword query based focused crawler by using URLs, DOM tree, K level, and Max Ancestor based feature. Wei Yan et al. [5] present evolutionary focused crawler by encapsulating

page rank, vector space model, genetic algorithm, and similarity function. Jianghui Zhou et al. [6] present theme relevancy crawling for multi-mode agricultural market analysis through Aho Corasick Algorithm.
Zhao et al. [7] presents deep web interface for harvesting Smart Crawler. Smart Crawler use search engine for extra cting center page and prioritize highly relevant URLs. Simultaneously, SmartCrawler employed adaptive link-ranking algorithm for extracting most relevant links. Whereas, Xiaojun Liu et al. [8] presents Sina Weibo based web crawler for extracting Chinese citizen sentiment about green building. This approach applied text mining, ontology, and keywords search for dictionary-based sentiment crawler.
Apart from search domain focused crawler also being employed for vulnerability and security analysis, Kim and Pant [9]-focused crawler for Detection of Malicious Web Page with the help of machine learning technique and audience Demography. Malicious web page focused crawler employed Naive Bayes, SVM, Logistic Regression with Statistical Analysis of web content. Khalil et al. [10] present a multithreaded duplicate content detection web crawler as RCrawler that useful for web crawling, scraping, and link analysis. RCrawler extract URL, Page Content, and depth level feature and employed similarity hash function algorithm for parallel web crawling and scraping. Janis Dalins et al.[11] presents focused crawler for extracting and blocking dark web for child pornography. Focused crawler for dark web used labeled and page content as suspicious text feature. Pedro Ivo et al. [12] present a focus crawler for analyzing business threats and opportunities analysis by integrating pattern recognition, ontology, and weak signal monitoring. Anchor text for Business threats is extracted by using Part of Speech tagging and SVO Typology. Rong Wang et al. [13] present a focused crawler for blacklisting malicious web page. The focused crawler identifies malicious URLs by applying decision tree-based machine learning techniques over correlation and DOM tree-based feature. Whereas Harry T Yani et al.[14] present distributed Focused crawler for tracking cyber attack, which is a multi-thread web crawler that use the optimal number of threads to maximize the download speed.
Fayyad et al. [15] have presented an analysis on KDD process, Knowledge discovery in databases (KDD) has been characterized as the non-trivial procedure of distinguishing substantial, novel, possibly valuable, and at last justifiable information from the data.
Apte [16] described Data Mining, Data Mining is a process by which precise and beforehand obscure information can be released from a large amount of data in a form that can be accepted, influenced, and employed for enhancing decision-making procedures. Therefore by Q. Luo [17] Data mining or knowledge discovery is the way toward dissecting data from different perceptions and briefing it into valuable information that can be utilized to recognize risks relevant to a particular project.
Bennouas et al. [18] proposed a random Web crawl model. That deals with the hyperlink structure, whose vertices are the pages and whose edges are the hypertextual links. It models the simpler web graph crawling process instead of the page writing procedure. Kosala et al. [19] have presented a review on Web mining. Web Mining is the application of data mining techniques to discover patterns and knowledge from the Web. There are three different types of web mining: web content mining, web structure mining, and web usage mining. M. Ghosh et al. [20] developed a better opinion mining algorithm.

Research Gap

There are many challenges while dealing with web crawlers. Some of the important challenges are mentioned below.

- There are no universal norms available for building websites; this causes data collec- tion difficult due to the unstructured format of web pages.
- Due to the dynamic nature of the web pages, it isn't very easy to maintain the freshness of the database in search engines. A good Focused Crawler should collect all relevant contents while ignoring the irrelevant contents from the web pages.
- Currently, many web crawlers work efficiently with text data but crawling the multi- media content is still a challenging task.
- There is a significant portion of the web which is hidden called deep web. It is difficult to reach the hidden web directly because it is only accessible through querying the database and selection of query is also a challenge.

- While evaluating the performance of a focused crawler, deciding whether or not a page is relevant to a topic or genre is one of the major challenges in this field.
- While a Focused Crawler saves a lot of resources by not crawling irrelevant, it might also miss out on certain relevant pages. This is a severe problem for languages which have less presence on the web.
- Selection of seed URLs for special purpose crawling is another challenging task be- cause the performance of focused crawling depends on the seed selection.

CONCLUSION

The use of the web has increased a lot in the last few decades. The World Wide Web became the best source of information in all areas. Therefore, it is necessary for the user to make the web simple and easy to use. As the web is an enormous store of links and web pages. The field of web mining has been already researched in so many perspectives and many aspects are still untouched. In this thesis research work focused on various problems related to information extraction from web structure data like analysis of page ranking, Data preprocessing for web structure mining, Algorithm for Mining Top-N High Utility URL sets, automated seed URLs selection and combination selection of seed URLs.

Data mining is the process of analyzing data from different perspectives and summarizing it into useful information. With the growth of the Web, a large amount of data is now available for users on the web. Web Preprocessed data improves the efficiency and scalability of later stages of Web structure mining. This can be done in several phases: Data fusion, Data Extraction, Data cleaning, links and metadata extraction, Path completion, etc. Data fusion includes collecting of pages collected from various Web servers. Data cleaning refers to the cleaning of irrelevant links which is not useful for the purpose of structure analysis i.e. multimedia les html style sheet etc

REFERENCES

1) W. Wang, X. Chen, Y. Zou, H. Wang, and Z. Dai, A focused crawler based on naive bayes classifier", in 2010 Third International Symposium on Intelligent Information Technology and Security Informatics, pp. 517-521, April 2010. 2.
2) A. I. Saleh, A. E. Abulwafa, and M. F. A. Rahmawy, "A web page distillation strategy for e-cient focused crawling based on optimized naïve bayes (onb) classifier," Applied Soft Computing, vol. 53, pp. 181- 204, 2017. 3.
3) Y. Du, W. Liu, X. Lv, and G. Peng, "An improved focused crawler based on semantic similarity vector space model," Applied Soft Computing, vol. 36, pp. 392 - 407, 2015.
4) M. Kumar, A. Bindal, R. Gautam, and R. Bhatia, "Keyword query based focused web crawler," Procedia Computer Science, vol. 125, pp. 584 - 590, 2018. The 6th International Conference on Smart Computing and Communications.
5) W. Yan and L. Pan, "Designing focused crawler based on improved genetic algorithm," in 2018 Tenth International Conference on Advanced Computational Intelligence (ICACI), pp. 319-323, March 2018.
6) J. Zhou, C. Cheng, L. Kang, and R. Sun, "Integration and analysis of agricultural market information based on web mining," IFAC-PapersOnLine, vol. 51, no. 17, pp. 778- 783, 2018. 6th IFAC Conference on Bio-Robotics BIOROBOTICS 2018.
7) F. Zhao, J. Zhou, C. Nie, H. Huang, and H. Jin,"Smartcrawler: A two-stage crawler for e-ciently harvesting deep-web interfaces," IEEE Transactions on Services Computing, vol. 9, pp. 608-620, July 2016.
8) X. Liu and W. Hu, "Attention and sentiment of chinese public toward green buildings based on sina weibo," Sustainable Cities and Society, vol. 44, pp. 550 - 558, 2019.
9) I. Kim and G. Pant, "Predicting web site audience demographics using content and design cues," Information and Management, 2018.
10) S. Khalil and M. Fakir, "Rcrawler: An r package for parallel web crawling and scraping" SoftwareX, vol. 6, pp. 98 -106, 2017.
11) J. Dalins, C. Wilson, and M. Carman, "Criminal motivation on the dark web: A categorisation model for law enforcement," Digital Investigation, vol. 24, pp. 62 - 71, 2018.

12) P. I. Garcia-Nunes and A. E. A. da Silva, "Using a conceptual system for weak signals classication to detect threats and opportunities from web," Futures, vol. 107, pp. 1 - 16, 2019.
13) R. Wang, Y. Zhu, J. Tan, and B. Zhou, "Detection of malicious web pages based on hybrid analysis," Journal of Information Security and Applications, vol. 35, pp. 68 - 74, 2017.
14) H. T. Y. Achsan and W. C. Wibowo, "A fast distributed focused-web crawling," Procedia Engineering, vol. 69, pp. 492 - 499, 2014. 24th DAAAM International Symposium on Intelligent Manufacturing and Automation, 2013.
15) C. Olston and M. Najork, "Web crawling" Foundations and Trends® in Information Retrieval, vol. 4, no. 3, pp. 175-246, 2010.
16) X. Qi and B. Davison, "Web page classication: Features and algorithms," ACM Comput. Surv., vol. 41, 01 2009.
17) J.-H. Lee, W.-C. Yeh, and M.-C. Chuang, "Web page classi- cation based on a simplied swarm optimization" Applied Mathematics and Computation, vol. 270, pp. 13 - 24, 2015.
18) S. Batsakis, E. G. Petrakis, and E. Milios, "Improving the performance of focused web crawlers," Data and Knowledge Engineering, vol. 68, no. 10, pp. 1001 - 1013, 2009.
19) H. Zhang and J. Lu, "Sctwc: An online semi-supervised clustering approach to topical web crawlers," Applied Soft Computing, vol. 10, no. 2, pp. 490 - 495, 2010.
20) H. Dong and F. K. Hussain, "Self-adaptive semantic focused crawler for mining services information discovery," IEEE Transactions on Industrial Informatics, vol. 10, pp. 1616- 1626, May 2014.