Web Crawlers for Searching Hidden Pages: A Survey

K.F.Bharati
Asst. Professor
CSE Dept, JNTUACEA
Ananthapur

P.Premchand
Dean, Faculty of Engineering
CSE Dept, UCEOU
Osmania University
Hyderabad

A.Govardhan
Director of Evaluation
JNTUH
Hyderabad

ABSTRACT

Many researchers have addressed the need of a dynamic proven model of web crawler that will address the need of several dynamic commerce, research and ecommerce establishments over the web that majorly runs with the help of a search engine. The entire web architecture is changing from a traditional to a semantic. And on the other hand the web crawlers. The web crawler of today is vulnerable to omit several tons of pages without searching and also is incapable of capturing the hidden pages. There are several research problems of information retrieval, far from optimization such as supporting user to analyze the problem to determine information needs. The paper makes an analytical survey of several proven web crawlers capable of searching hidden pages. It also addresses the prospects and constraints of the methods and the ways to further enhance.

Keywords: Web crawler, Hidden pages search, search optimization.

1. INTRODUCTION

The World Wide Web consortium has reported the growth of web from few thousand pages in 1990's to more than two billion pages at this stage. Nowadays, the information is available in several forms like Websites, databases, images, sound, videos, etc. Due to the vastness of the available information, web search engines have become a primary tool on the web to order, organize and retrieve the information. Searching for information is a primary activity on the Web and about 80% of Web users use search engines to retrieve information from the Web [2]. Searching tools like Google forms the primary tool for information retrieval, but they are limited to certain restriction and are not eligible in finding hidden pages of the web that needs certain authorization or certificate or a prior registration or querying interface to retrieve the information.

1.1. Motivation

The present day search engines are capable of querying web content to a certain extent and almost all the web engines act on a same method. The classification of services wrap only a part of the web called the openly index able Web referring to the set of web pages available simply by next hypertext links, ignoring search forms and pages that need authorization or prior registration.

Whenever a search engine or a crawler is designed, a mathematical model for the same is designed and the implementation of the algorithm is done using a platform and a programming languages. The mathematical models of information retrieval are the channel for the implementation of Information retrieval systems. The conventional search engines, which are usually operated by professional searchers, only the matching process are automated, indexing and query formulation are manual processes. For these systems, mathematical models of information retrieval are used to model the matching process alone.

There are several research problems of information retrieval, far from optimization such as guiding user in order to determine one's needs, the analysis of people's way of using and processing information, accumulating a package of information that facilitates the user to come closer to a solution, representing Knowledge, the ways of processing knowledge/information, the human computer interface for better information retrieval, a better user-enhanced information systems design and a optimal method to evaluate a information retrieval system.

There are several other improvements to be made to the crawler architecture, compression of data and information, crawling algorithms for hidden pages and scaling of algorithms. The concentration of the algorithms have to swift to attributes like number of documents indexed, queries per second, index freshness and update rate, query latency information of each document[27].

2. RELATED WORK

The entire internet lies on the search engine and more than 85% of the users use search engines to find their information [1]. Internet search engines runs on the classical interactive information retrieval method of entering a query, retrieving references to the documents, examining some documents and accordingly reformulating the query. Usually search engine was used by professionals for medical research, indexing libraries, and for archiving. This decade they observed the latest in search engine and it is used by browsers for reasons like shopping, for information retrieval and for almost all purposes.

Professional search engines acts as a search middleware for end users or customers and try to figure out in an interactive dialogue with the system and the customer, what the customer needs, and how this information should be used in a successful search. There are several proven mathematical models that guide the implementation of information retrieval systems. The extension of this search engine is a specialized crawler used to find and retrieve hidden pages. This paper will analyze the techniques and methodologies used by the web crawlers which are used to retrieve web pages.

2.1. The Survey

The main intention of the web mining is to crawl important pages that have been on the rise by using a separate set of data mining algorithms that are on the rise. These factors necessitate the creation of server-side and client side

intelligent systems capable of mining knowledge equally across the Internet and in particular several Web localities. All the firms are forced to provide information services on the web like Customer support, online trading and several web services for electronic commerce, collaboration, news and broadcasting [28].

2.2. Segmentation

The way of setting apart noisy and unimportant blocks from the web pages can facilitate search and to improve the web crawler. This way can facilitate even to search hidden web pages. However, still there is no uniform approach to divide the pages into blocks and measure it. In order to distinguish and establish different information in a web page, the need is to segment a web page into a set of blocks. Several methods exist for web page segmentation. The most popular ones are DOM-based segmentation [5], location-based segmentation [10] and Vision-based Page Segmentation [4]. The paper deals with capability of differentiating features of the web page as blocks and modeling is done on the same to find some insights to get the knowledge of the page using two methods based on Neural Network and SVM facilitating the page to be found.

2.3. Data Extraction Techniques

The availability of robust, flexible Information Extraction (IE) systems for transforming the Web pages into algorithm and program readable structures like relational database that will help the search engine to search easily. Several approaches for data extraction from web pages have been always there, but they were limited to certain extent. The paper analyses major web data extraction techniques and approaches, tabulating them and finds the prospects and constraints of the technique used and also surveys the major Web data extraction approaches and compares them in several magnitude like, the task domain, the automation degree, and the techniques used. It also explains the reason why the IE system fails to handle some Web sites of particular structures. The second dimension classifies IE systems based on the techniques used. The third dimension criteria measure the degree of automation for IE systems [6].

The lists of available web crawler architectures are Yahoo! Slurp, Bingbot, FAST Crawler, Googlebot, PolyBot, RBSE, Web Crawler, Web Fountain and there are also open source crawler like Abot, Aspseek, DataparkSearch and GNU Wget that can be used to update and test newer algorithms as the crawlers are open to change.

2.4. Skeleton of Web sites

Extracting the underlying hyperlink structure used to organize the content pages in a concerned websites. They have proposed an automated BOT like algorithm that has the functionality of discovering the skeleton of a given website. The SEW algorithm [7], its examines hyperlinks in groups and identifies the navigation links that point to pages in the next level in the website structure. Here the entire skeleton is then constructed by recursively fetching pages pointed by the discovered links and analyzing these pages using the same process. The paper experiments real time websites for the same algorithm.

2.5. Scalability

The issue of extraction of search term for over millions and billions of information and have touched upon the issue of scalability and how approaches can be made for a very large databases. The key algorithms discussed for scaled up information extraction include the usage of general-purpose search engines plus certain proven indexing techniques specialized for information extraction applications. Scalable information extraction is one untouched area and the papers actively emphasize the challenges in the area. The discussion of the paper continues to the introduction of several new approaches, like scanning approach, that is done using template based efficient rules. In the case, every document is processed using the help of patterns and template rules highly optimized for speed. The next approach is to exploit general-purpose search engines to avoid scanning all documents in a group. The next approach is using specialized indexes and custom search engines: A special-purpose search engine capable of indexing and make query annotations useful for extraction. The paper discusses about a final distributed processing approach defining distributed data mining solutions that can be used for scalable text mining and also for information extraction. These approaches have been tested for its extraction, completeness, accuracy and scalability [8].

3. CRAWLERS

The current day crawlers and their inefficiencies in pulling the correct data. Their analysis covers the concept of Current-day crawlers retrieving content only from the publicly index able Web, the pages reachable only by following hypertext links and ignoring the pages that require certain authorization or prior registration for viewing them. There are different types of crawlers as shown in Table 1.

The paper says that the crawlers ignore completely a huge amount of highly qualified and quality content, as they were hidden to the crawlers. The ways and techniques of collecting the hidden pages are also discussed. The design of one such crawler capable of extracting information from this hidden Web is modeled by using a generic operational model. The realization of the model is made using Hidden Web Exposer, a prototype crawler.

A Layout-based Information Extraction Technique (LITE) demonstrates the way it automatically extract semantic content from search forms and response pages. The whole concepts presented in the paper is proved by experimentation have provided with a generic high-level operational model of a hidden Web crawler and metrics for calculating the performance of such crawlers. At last identification of the key design issues for coming out with such a crawler is done. The design issues in the paper answers several questions like type of information about each form element from which the crawler should collect and the meta-information about each form that is likely to be useful in designing better matching functions. It also describes how the task-specific database has to be organized, updated, and accessed [9].

3.1. Techniques for Creating Crawlers

The different characteristics of web data, the basic mechanism of web mining and its several types. The reason for the usage of web mining for the crawler functionality is well explained in the paper. Even the limitations of some of the algorithms are listed. The paper talks about the usage of fields like soft computing, fuzzy logic, artificial networks and genetic algorithms for the creation of crawler. The paper gives the reader the future design that can be done with the help of the alternate technologies available.

The later part of the paper deals with describing the characteristics of web data, different components, types of web mining and the limitations of existing web mining methods[11]. The applications that can be done with the help of these alternative techniques are also described. The survey involved in the paper is in-depth and surveys all systems which aim to dynamically extract information from unfamiliar resources. Intelligent web agents are available to search for related content using characteristics of an exact domain got from the user profile to put in order and read the discovered information. There are several available agents such as Harvest [15], FAQ-Finder [16], Information Manifold [17], OCCAM [[18], and Parasite [19], that rely on the predefined domain specific template information and are experts in finding and retrieving specific information.

The Harvest system depends upon the semi-structured documents to extract information and it has the capability to exercise a search in a latex file and a post-script file. That is mostly used well in bibliography search and reference search, is a great tool for researchers as it searches with key terms like authors and conference information. In the same way FAQ-Finder [16], is a great tool to answer Frequently Asked Questions (FAQs) [15], by collecting answers from the web. The other systems described are ShopBot [20] and Internet Learning Agent [21] retrieves product information from numerous vendor website using generic information of the product domain. A search about "laptop" gives a search results have pages taken from different vendor web pages and also results certain hidden pages. Internet Learning Agent learns to extract information from unfamiliar by search with querying objects of interest.

3.2.Semantic Web

The evolving web architecture and the behavior of web search engines have to be altered in order to get the desired results. The next-generation Web architecture popularly known as semantic web needs accurate search crawler to overcome the limitation of the traditional web searcher. The ranking system among the result has also been made an impact. Relevance is measured as the probability that a retrieved resource actually contains those relations whose existence was assumed by the user at the time of query definition.

3.3. Ranking

Ranking based search tools like Pubmed that allows users to submit highly expressive Boolean keyword queries, but ranks the query results by date only. A proposed approach is to submit a disjunctive query with all query keywords, recover all the returned identical documents, and then re-rank them. But the expensiveness of such an operation leads to the finding of a newer approach that returns the top results for a query, ranked according to the proposed ranking function [13]. This approach can also be applied to several other setting when the ranking is monatomic [12].

Table 1:Types of Crawlers

SNO	TYPES OF CRAWLERS	DECRIPTION	PROs	CONs
1	BREADTH FIRST CRAWLER	Starts with a small set of pages and then explores additional pages by subsequent links in the breadth-first fashion.	It crawls most significant pages first.	The Internet Archive crawler does not carry out a breadth first search of the whole web.
2	INCREMENTAL WEB CRAWLER	It updates an existing set of downloaded pages as a replacement for of restarting the crawl from scratch every time.	It can develop the freshness and the quality of its index/collection radically and saves time.	It is costly in terms of hardware.
3	FOCUSED CRAWLER	The focused crawler aims at providing a simpler substitute for overcoming the issue that immediate pages that are lowly ranked associated to the topic at hand.	Spends less time and effort for processing web pages .	The problem of zero probability and find out the relevancy of unvisited URLs.
4	PARALLEL CRAWLER	A parallel crawler is a crawler that runs several processes in parallel. The goal is to maximize the download rate while minimizing the overhead from parallelization and to avoid frequent downloads	Scalability, Network-load dispersion.	Redundancy storage is required.

		of the similar page.		
5	DISTRIBUTED WEB CRAWLER	In circulated web crawler a URL server distributes individual URLs to multiple crawlers, which download web pages in parallel, the crawlers then send the downloaded pages to a central indexer on which links are extracted and sent via the URL server to the crawlers.	It reduces the hardware necessities and increases the overall download Speed and reliability.	Web partitioning/repartitioning and data center placement are required.

3.4.Form Filling

An approach by which a user fills up a form in order to get a set of relevant data. The process is tedious for a long run and when the number of data to be retrieved is heavy. For the same an alternate method is discussed, by which an agent fills the forms automatically as it has the ability to learn. This approach also helps to retrieve hidden pages systematically [14].

In the thesis by Tina Eliassi-Rad, several works that retrieve hidden pages are discussed. There many proposed hidden pages techniques which are unique to web crawler algorithms, to search the hidden pages. [22] Automatically detects the domain specific search interfaces by looking at the urls name and also at the title of the html attributes. It's done using a set of categories using domain ontology.

An architectural model for extracting hidden web data. The main focus of this work is to learn Hidden-Web query interfaces, not to generate queries automatically. Their approach is not automatic and requires human input [24].

The scheduling algorithms for web crawling is discussed, the paper proposes methods for Web page ordering happening through a web crawl and compare them using a simulation by considering a competitive and efficient scenario. Real Web crawler is used for the approach [29].

Several scheduling strategies whose design is based on a heap priority queue with nodes representing sites are considered. For each site-node they have another heap representing the pages in the Web site, thereby simulating the real time scenario.

A novel technique of modeling the web crawler's traversal path and their by modifying the behavior of the web crawler to the required form. For the same they have used a symbolic model checking tool called nuSMV. The correctness of the crawler path and the entire possible states the crawler can acquire is detected in their work. The paper provides a modeling technique to analysis the design of any crawler model and their by optimizing the existing feature of a crawler in terms of politeness, robustness, quality and distribution. The authors have adequately used symbolic model checking to verify the constraints placed on the system by analyzing the entire state space of the system. This paper provides with an example with the trace path highlighting the location of error, if the constraints is not met by the crawler [12].

4. CONCLUSION

The paper surveys several search algorithms that are used for extracting hidden pages for the web. Each of the paper follows a specific for extracting hidden pages with the advent of several newer techniques like genetic algorithms, artificial neural networks, expert system, machine learning and fuzzy logic. The survey also portrays the need of newer methods of web crawler as the internet is never a same and it changes its architecture dynamically. A proven model of web crawler, which is capable of pulling the prominent required information from several hidden part of web.

5. ACKNOWLEDGMENTS

My sincere thanks to Prof. P.Premchand and Prof. A.Govardhan who has contributed me to develop this paper.

6. REFERENCES

[1] S.Lawrence, C. L. Giles, "Accessibility of Information on the Web," Nature, 400, 107-109, 1999.

[2] Djoerd Hiemstra : Using language models for Information Retrieval . Univ. Twente 2001: I-VIII, 1-163.

[3] Ruihua Song, Haifeng Liu, Ji- Rong Wen, Wei-Ying Machine: Learning Important Models for Web Page Blocks Based On Layout and Content Analysis. SIGKDD Explorations 6(2): 14-23 (2004).

[4] Cai, D., Yu, S., Wen, J.-R. and Ma, W.-Y., VIPS: A Vision Based Page Segmentation Algorithm, Microsoft Technical Report, MSR-TR-2003-79, (2003).

[5] Chen, J., Zhou, B., Shi , J., Zhang, H.-J. and Qiu, F Function - Based Object Model Towards Website Adaptation, in the proceedings Of the 10th World Wide Web conference (WWW10), Budapest, Hungary, May (2001).

[6] Chia-Hui Chang, Mohammed Kayed, Moheb R. Girgis, Khaled F. Shaalan : A Survey of Web Information Extraction Systems IEEE Trans. Knowl. Data Eng. 18(10): 1411-1428 (2006).

[7] Zehua Liu, Wee Keong Ng, Ee-Peng Lim: An Automated Algorithm for Extracting Website Skeleton. DASFAA 2004: 799-811.

[8] EugeneAgichtein:Scaling Information Extraction to Large Document Collections. IEEE Data Eng. Bull. 28(4): 3-10 (2005).

[9] Sriram Raghavan, Hector Garcia Molina: Crawling the Hidden Web. VLDB 2001: 129-138.

[10] Kovacevic, M., Diligenti, M., Gori, M. and Milutinovic, V.,Recognition of Common Areas in a Web Page Using Visual Information: A Possible Application In A Page Classification,in the proceedings of 2002 IEEE International Conference on Data Mining (ICDM'02), Maebashi City, Japan, December,(2002).

[11] Sankar K. Pal, Varun Talwar, Pabitra Mitra: Web Mining In Soft Computing Framework:relevance, state of the art and future directions. IEEE Transactions on Neural Networks 13(5): 1163-1177 (2002).

[12] Fabrizio Lamberti, Andrea Sanna, Claudio Demartini: A Relation - Based Page RankAlgorithm for Semantic Web Search Engines. IEEE Trans. Knowl. Data Eng. 21(1): 23-136.

[13] Vagelis Hristidis, Yuheng Hu, Panagiotis G. Ipeirotis Relevance - Based Retrieval on Hidden Web Text Databases Without Ranking Support. IEEE Trans. Knowl. Data Eng. 23(10): 1555-1568 (2011).

[14] Stephen W. Liddle, Sai Ho Yau, David W. Embley: On the Automatic Extraction of Data From the Hidden Web. ER (Workshops) 2001: 212-226.

[15] K. Hammond, R. Burke, C. Martin, and S.Lytinen, "Faq-finder: A case based approach to knowledge Navigation," presented at theWorking Notes of AAAI Spring Symposium on Information Gathering From Heterogeneous Distributed Environments, Stanford, CA, (1995).

[16] A. Y. Levy, T. Kirk, and Y. Sagiv, "The gll information manifold," presented at the AAAI Spring Symposium on Information Gathering From Heterogeneous Distributed Environments, (1995).

[17] C. Kwok and D. Weld, "Planning to gather information," in Proc. 14th Nat. Conf. AI, (1996).

[18] E. Spertus, "Parasite: Mining Structural Information on the web," presented at the Proc. 6th WWW Conf., (1997).

[19] O. Etzioni, D. S. Weld, and R. B. Doorenbos, "A Scalable Comparison Shopping Agent for The World Wide Web," Univ. Washington, Dept. Comput. Sci., Seattle, Tech. Rep. TR 96- 01-03, (1996).

[20] O. Etzioni and M. Perkowitz, "Category translation: Learning to Understand Information on the internet," in Proc. 15th Int. Joint Conf. Artificial Intell, Montreal, QC, Canada, (1995). pp. 930–936.

[21] M. Craven, D. Freitag, A. McCallum, T. tchell, K. Nigam, S. Slattery , and D. DiPasquo,"Learning to extract Symbolic Knowledge from the World Wide Web," in Proc. 15th Nat. Conf. AI (AAAI98), 1998, pp.509–516.

[22] Anuradha, A.K.Sharma, "A Novel Approach for Automatic Detection and Unification of Web Search Query Interfaces using Domain Ontology" selected in International Journal of Information Technology and knowledge management(IJITKM), August (2009).

[23] S. Raghavan and H. Garcia - Molina. Crawling The Hidden Web. In Proceedings of VLDB, pages 129–138, 2001.

[24] Shetty, K.S.; Bhat, S.; Singh, S.; , "Symbolic verification of web crawler functionality and its properties," Computer Communication and Informatics (ICCCI), 2012 International Conference on , vol., no., pp.1-6, 10-12 Jan. (2012).

[25] Weicheng Ma ; Xiuxia Chen; Wenqian Shang; "Advanced Deep Web Crawler Based on Dom," Computational Sciences and Optimization (CSO),2012 Fifth International JointConference on , vol., no., pp.605-609, 23-26 June (2012).

[26] Jeff Dean, Google Fellow," Google Challenges in Building Large-Scale Information Retrieval Systems" Research.google.com.

[27] Subhendu kumar pani et. al.," Integration of Web Mining and web Crawler : Relevance and State of Art," International Journal on Computer Science and Engineering, Vol. 02, No. 03, 2010, 772-776.

[28] Carlos Castillo, Mauricio Marín, Andrea Rodríguez, Ricardo A. Baeza-Yates: Scheduling Algorithms for Web Crawling. 10-17.

[29] Birrell, A.D., Levin, R., Needham, R.M. and Schroeder, M.D. Grapevine : an exercise In distributed computing. Communications Of the ACM, 25 (4) 260-274.(1992).

[30] A. K. Sharma, J.P. Gupta, D. P. Agarwal, "Augmented Hypertext Documents Suitable For Parallel Crawlers", Communicated to 21st IASTED International Multi-conference_Applied Informatics AI-2003, Feb 10- 13,2003, Austria.

[31] Dhiraj Khurana, Satish Kumar " Web Crawler: A Review", IJCSMS, Vol 12, Issue 01 Jan, 2012, ISSN (online): 2231-5268