RESEARCH ARTICLE OPEN ACCESS

Key Issues And Challenges with Web Crawlers

[1]P.Srikanth[*,] [2]K.V.Sai phani[**,][3]G.Venu Gopal[***]
[1,2,3,]Assistant professor,Dept of CSE Nalla narasimha reddy educational society's
Group of institution's Korremulla, Narapally,
Ghatkesar Hyderabad, Telangana
Corresponding Authoe: [1]P.Srikanth[,]*

ABSTRACT

Due to the current size of the Web and its dynamic nature, building an efficient search mechanism is very important. A vast number of web pages are continually being added every day, and information is constantly changing. Search engines are used to extract valuable Information from the internet. Web crawlers are the principal part of search engine, it is program or computer software that browses the internet in an automated manner or in an orderly fashion. It is an essential method for collecting data on, and keeping in touch with the rapidly increasing Internet.

Keyword: Crawling techniques, Web Crawler, Search engine, WWW

Date of Submission: 02-11-2017 Date of acceptance: 11-12-2017

I. INTRODUCTION

A web crawler (also called web spider, web robot) is typically a script or computer program that browses the targeted website in an orderly and automated manner. It is an important method for collecting information on the Internet and is a critical component of search engine technology. Most popular search engines, such as GoogleBot and BaiduSpider, use underlying web crawlers to get the latest data on the internet. All web crawlers take up internet bandwidth. But not all web crawlers are benign. A well behaved web crawler usually identifies itself and balances the crawling frequencies and contents and thus the bandwidth consumption.

On the other hand, an ill-behaved or malicious web crawler can consume large amounts of bandwidth and cause disruptions, especially to companies that rely on web traffic or content for their business. For companies that rely on their website and online content to conduct business, if a web crawler is created by a hacker or unauthorized users and used on bots, it can be used to steal data and information from businesses with the possibility of staging DDOS attacks towards targeted websites. How to effectively detect malicious web crawlers has become a critical topic in today's cyber threat defense sector. In modern life use of internet is growing in rapid way.

The World Wide Web provides a vast source of information of almost all type. Now a day's people use search engines every now and then, large volumes of data can be explored easily through search engines, to extract valuable information from web. However, large size of the Web, searching all the Web Servers and the pages, is not realistic. Every day number of web pages is added and nature of information gets changed [1]. Due to the extremely large number of pages present on Web, the search engine depends upon crawlers for the collection of required pages [6]. Web crawling is an important method for collecting data and keeping up to date with the rapidly expanding Internet. A web crawler is a program, which automatically traverses the web by downloading documents and following links from page to page [3]. It is a tool for the search engines and other information seekers to gather data for indexing and to enable them to keep their databases up to date [1].

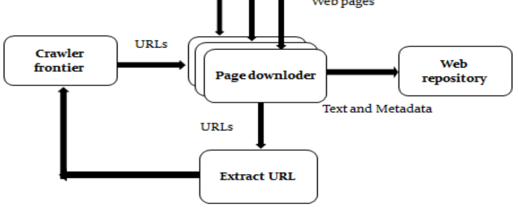

Fig 1: Web crawler

II. LITERATURE REVIEW

WWW contains millions of information beneficial for the users, many information seekers usage search engine to initiate their Web activity. Every search engine rely on a crawler module to provide the grist for its operation [18]. Matthew Gray wrote the first Crawler, the World Wide Web Wanderer, which was used from 1993 to 1996 [10]. J. Cho. describes various search techniques and how the search engines works by using crawler and he has described how the search engines should cope with the evolving Web, in an attempt to provide users with upto- date results. He has made the various studies on crawler policies. And Proposes how one can maintain local copies of remote data sources fresh, when the source data is updated autonomously and independently.Gautam Pant and Filippo Menczer examined the us e of focused crawler in [16]. S.S. Dhenakaran1 and K. Thirugnana Sambanthan [3] give an overview about Different types of Web crawler and the policies being used in the web crawlers and their evolution. Ms. Swati Mali and Dr. B.B. Meshram in [4] implements effective multiuser personal web crawler where one user can manage multiple topics of interest.

This type of web crawler can be configured to target precisely what user needs. It offers a high degree of control over the information that is returned for a particular search, vastly increasing the likelihood that it will be relevant. A crawler is a program that downloads and stores web pages often for a web search engine. The rapid growth of World Wide Web poses challenges to search for the most appropriate link. Author Pooja gupta and Mrs. Kalpana Johari [5] has developed a Focused crawler using breadth-first search to extract only the relevant web pages of interested topic from the Internet. In [6] author Keerthi S. Shetty, Swaraj Bhat and Sanjay Singh, used symbolic model checking approach to model the basic operation of crawler and verify its properties by using The tool NuSMV. It helps to verify the constraints placed on the system by exploring the entire state space of the system. Hiroshi Takeno, Makoto Muto, Noriyuki Fujimoto introduced a new Web crawler that collects Web content suitable for viewing on mobile terminals such as PDA or cell phones. They have described Mobile Search Service that provides content suitable for mobile terminals.

Crawler

A web crawler is a software or programmed script that browses the World Wide Web in a systematic, automated manner. The structure of the WWW is a graphical structure, i.e., the links presented in a web page may be used to open other web pages. Internet is a directed graph where webpage as a node and hyperlink as an edge, thus the search operation may be summarized as a process of traversing directed graph. By following the linked structure of the Web, web crawler may traverse several new web pages starting from a webpage. A web crawler move from page to page by the using of graphical structure of the web pages. Such programs are also known as robots, spiders, and worms. Web crawlers are resigned to retrieve Web pages and insert them to local repository. Crawlers are basically used to create a replica of all the visited pages that are later processed by a search engine that will index the downloaded pages that help in quick searches. Search engines job is to storing information about several webs pages, which they retrieve from WWW.

The working of a web crawler is as follows:

1. Initializing the seed URL or URLs
2. Adding it to the frontier
3. Selecting the URL from the frontier
4. Fetching the web-page corresponding to that URLs
5. Parsing the retrieved page to extract the URLs
6. Adding all the unvisited links to the list of URL i.e. into the frontier
7. Again start with step 2 and repeat till the frontier is empty.

Types Of Web Crawler

Different strategies are being employed in web crawling. These are as follows.

- **Focused Web Crawler**

Focused Crawler is the Web crawler that tries to download pages that are related to each other [4]. It collects documents which are specific and relevant to the given topic [7]. It is also known as a Topic Crawler because of its way of working [4]. The focused crawler determines the following – Relevancy, Way forward. It determines how far the given page is relevant to the particular topic and how to proceed forward. The benefits of focused web crawler is that it is economically feasible in terms of hardware and network resources, it can reduce the amount of network traffic and downloads [11]. The search exposure of focused web crawler is also huge [2][9].

Issues and Challenges with Focused Crawler

i. Missing Relevant Pages
ii. Maintaining Freshness of Database
iii. Network Bandwidth and Impact on Web Servers

- **Incremental Crawler**

A traditional crawler, in order to refresh its collection, periodically replaces the old documents with the newly downloaded documents. On the contrary, an incremental crawler incrementally refreshes the existing collection of pages by visiting them frequently; based upon the estimate as to how often pages change. It also exchanges less important pages by new and more important pages. It resolves the problem of the freshness of the pages. The benefit of incremental crawler is that only the valuable data is provided to the user, thus network bandwidth is saved and data enrichment is achieved

Issues and Challenges with Incremental Crawler

i. Keep the local collection fresh
ii. Improve quality of the local collection

- **Distributed Crawler**

Distributed web crawling is a distributed computing technique. Many crawlers are working to distribute in the process of web crawling, in order to have the most coverage of the web. A central server manages the communication and synchronization of the nodes, as it is geographically distributed [2]. It basically uses Page rank algorithm for its increased efficiency and quality search. The benefit of distributed web crawler is that it is robust against system crashes and other events, and can be adapted to various crawling applications .

Issues and Challenges with Distributed Crawler

i. Assignment of URL's among different agents
ii. Priority in Crawling
iii. Effective way of partitioning the collection
iv. Load Balancing
v. Network bandwidth consumption

- **Parallel Crawler**

Multiple crawlers are often run in parallel, which are referred as Parallel crawlers . A parallel crawler consists of multiple crawling Processes called as C-procs which can run on network of workstations. The Parallel crawlers depend on Page freshness and Page Selection [20]. A Parallel crawler can be on local network or be distributed at geographically distant locations [2].Parallelization of crawling system is very vital from the point of view of downloading documents in a reasonable amount of time

Issues and Challenges with Parallel Crawler

i. Multiple downloading of pages:
ii. Quality of pages
iii. Increased bandwidth Consumption

- **Mobile Crawler**

The mobile crawlers are constructed as mobile agents. Crawler mobility provides sophisticated crawling algorithms and avoids some of the inefficiencies associated with the strategies used by current crawlers. The mobile crawling is an efficient, scalable solution to establish a specialized search index in the highly distributed, decentralized and dynamic environment of the Web.

Issues and Challenges with the Mobile Crawlers

i. Security
ii. Integration of the mobile crawler virtual machine into the Web
iii. Less research in mobile crawling algorithms

Future work:

Many of the issues and challenges in these architectures are common i.e. reducing the network bandwidth consumption, maintaining the freshness of the database and maintaining the quality of pages etc. The mobile crawler was constructed as mobile agent. The major challenges in designing the mobile crawler were to maintain the security, non availability of required environment on most of the machines and less research in mobile crawling algorithms. Further, mobile crawlers are found to be the new paradigm and needs to be explored to get its benefits.

REFERENCES

[1] Bharat Bhushan1, Narender Kumar2, Intelligent Crawling On Open Web for Business Prospects", IJCSNS International Journal of Computer Science and Network Security, VOL.12 No.6, June 2012

[2] Pavalam S. M., S. V. Kasmir Raja, Jawahar M., and Felix K. Akorli, Web Crawler in Mobile Systems, International Journal of Machine Learning and Computing, Vol. 2, No. 4, August 2012

[3] S.S. Dhenakaran1 and K. Thirugnana Sambanthan2, WEB CRAWLER - AN OVERVIEW, International Journal of Computer Science and Communication Vol. 2, No. 1, January-June 2011, pp. 265-267.

[4] Ms. Swati Mali, Dr. B.B. Meshram, Implementation of Multiuser Personal Web Crawler, CSI Sixth International Conference on Software Engineering (CONSEG), IEEE Conference Publications, 2012.

[5] Pooja Gupta and Mrs. Kalpana Johari, Implementation of Web Crawler, Second International Conference On Emerging Trends In Engineering and Technology, ICETET-09, IEEE Conference Publications,2009

[6] Keerthi S. Shetty, Swaraj Bhat and Sanjay Singh, Symbolic Verification of Web Crawler Functionality and Its Properties , International Conference on Computer Communication and Informatics (ICCCI -2012), Coimbatore, INDIA, IEEE Conference Publications ,2012

[7] Md. Abu Kausar, V. S. Dhaka, Sanjeev Kumar Singh, Web Crawler: A Review, International Journal of Computer Applications (0975 – 8887) ,Volume 63– No.2, February 2013

[8] Wenxian Wang, Xingshu Chen, Yongbin Zou, Haizhou Wang, Zongkun Dai, A Focused Crawler Based on Naive Bayes Classifier, Third International Symposium on Intelligent Information Technology and Security Informatics, IEEE Conference Publications,2010 [9] Manas Kanti Dey, Debakar Shamanta, Hasan Md Suhag Chowdhury, Khandakar Entenam Unayes Ahmed, Focused Web Crawling: A Framework for Crawling of Country Based Financial Data, Information and Financial Engineering (ICIFE), IEEE Conference Publications, 2010

[9] Dr Rajender Nath, Khyati Chopra, Web Crawlers: Taxonomy, Issues & Challenges, International Journal of Advanced Research Computer Science and Software Engineer ing, Volume 3, Issue 4, April 2013

[10] Debashis Hati, Biswajit Sahoo, Amritesh Kumar, Adaptive Focused Crawling Based on Link Analysis , 2nd International Conference on Education Technology and Computer (ICETC),2010

[11] Jun Hirai Sriram Raghavan Hector Garcia-Molina Andreas Paepcke, WebBase : A repository of web pages, available: http://ilpubs.stanford.edu:8090/380/1/1999-26.pdf

[12] Frank McCown, Michael L. Nelson, Evaluation of Crawling Policies for a Web Repository Crawler , Copyright 2006 ACM 1595934170/06/0008, HT’06, August 22–25, 2006

[13] Shashi Shekhar, Rohit Agrawal and Karm Veer Arya, An Architectural Framework of a Crawler for Retrieving Highly Relevant Web Documents by Filtering Replicated We Collections, 2010 International Conference on Advances in Computer Engineering, IEEE Conference Publications 2010.

[14] Ioannis Avraam, Ioannis Anagnostopoulos, A Comparison over Focused Web Crawling Strategies 2011 Panhellenic Conference on Informatics, IEEE Conference Publications, 2011.

International Journal of Engineering Research and Applications (IJERA) is **UGC approved** Journal with Sl. No. 4525, Journal no. 47088. Indexed in Cross Ref, Index Copernicus (ICV 80.82), NASA, Ads, Researcher Id Thomson Reuters, DOAJ.

P.Srikanth*. “Key Issues And Challenges with Web Crawlers.” International Journal of Engineering Research and Applications (IJERA) , vol. 07, no. 12, 2017, pp. 58-61.