IJIRCCE

e-ISSN: 2320-9801 | p-ISSN: 2320-9798

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

Volume 9, Issue 10, October 2021

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 7.542

International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 7.542

|| Volume 9, Issue 10, October 2021 ||

| DOI: 10.15680/IJIRCCE.2021.0910013 |

Web Scraping Using Different Types of Web Bots

Rohit Kolape, Rama Bansode

PG Student, Dept. of M.C.A., Savitribai Phule Pune University, Pune, India

Assistant Professor, Dept. of M.C.A., Savitribai Phule Pune University, Pune, India

ABSTRACT: A web bot, which we generally call a "spider," is a man-made intelligence that browses the web to index and look for content by following links and exploring, like a person with an excessive amount of time on their hands. The internet is crawling with bots. A bot is a software program that runs automated tasks over the internet, usually performing simple, reiterative tasks at great speeds unattainable, or undesirable by humans. They're responsible for multiple small jobs that we take for granted like as search engine crawling, website health monitoring, acquiring web content, measuring website speed and powering APIs. They can also be used to automate security auditing by examining your network and websites to find vulnerabilities and help remediate them.These bots are nearly always operated by search engines. By applying an enquiry algorithm to the info collected by web crawlers, search engines can give related links in response to user seeking queries, generating the list of webpages that show up after a user type an enquiry into Google or Bing (or another search engine).A web scraping bot is like someone who goes through all the books during a disorganized library and puts together a card checklist in order that anyone who visits the library can swiftly and effortlessly find the information they need. To help classify and sort the library's books by subject, the organizer will read the title, summary, and some of the internal text of each book to figure out what it's about.

KEYWORDS: Web bot, World Wide Web, Search Engine

I. INTRODUCTION

In today's life use of internet is growing in fast way. The World Wide Web provides an abundance 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 travel easily through search engines, to collect valuable information from web. However, large size of the web, searching all the web Servers and therefore the pages, isn't practical. 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 bots for the collection of required pages. [6] These bots helps in searching the particular content on web page. Bots are small programs that 'browse' the web on the search engine's behalf, similarly to how a human user would follow links to reach different pages. [3] Google crawlers run on a distributed network of thousands of low-cost computers and can therefore carry out fast parallel processing. This is why and how Google returns results within fraction of seconds. [4]

Web bots also known as web crawlers, spiders, worms, walkers, and wanderers are almost as old as the web itself. The first bot, Matthew Gray's Wandered, was written in the spring of 1993, roughly coinciding with the first release of NCSA mosaic.[5]

II. LITERATURE REVIEW

WWW contains many information beneficial for the users, many information seekers usage search engine to initiate their Web task. Every search engine depends on a bot module to provide the support for its operation Matthew Gray wrote the first bot, the World Wide Web Wanderer, which was used from 1993 to 1996.J. Cho.[10] it describes numerous search approaches and how the search engines works by using bots.he has described how the search engines should deal with the evolving Web, in an attempt to supply users with up-to- date results. He has made the varied studies on bot policies. Proposes how one can maintain local clones of remote data sources " fresh,"when the source data is changing autonomously and solely. gives an idea about different types of bots. Gautam Pant and Filippo Menczer examined the use of focused bots inMs. Swati Mali andDr.B.B. Meshram in [4] tools effective multiuser privy web bot where one user can manage multiple contents of interest. Ms. Swati Mali and Dr.B.B.Meshram in [4] devices effective multiuser personal web bot where one user can manage multiple contents of interest. This type of web bot can

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 7.542

|| Volume 9, Issue 10, October 2021 ||

| DOI: 10.15680/IJIRCCE.2021.0910013 |

be configured to target precisely what user needs. It offers a high degree of control over the information that's returned for a particular search, vastly augmenting the likelihood that it'll be relevant. A bot is a program that downloads and stores web pages frequently for a web search engine. The quick 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 centered bot using breadth-first search to acquire only the relevant web pages of interested content 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. In author 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.[14] They have described "Mobile Search Service" that provides content suitable for mobile terminals.

III. WEB BOT

A web bot 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 bot may traverse several new web pages starting from a webpage. A web bot moves from page to page by the using of graphical structure of the web pages. Such scripts are also known as crawler, spiders, and worms. Web bots are designed to retrieve Web pages and take them to local repository. bots 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 scanning. Search engines job is to storing information about several web pages, which they retrieve from WWW. These pages are retrieved by a Web bot that is an automated Web browser that follows each link it sees [7].

A Search Engine Bot (also known as a crawler, Robot) is a program that most search engines use to find what's new on the websites. Google's web bot is known as GoogleBot. There are many types of web crawlers in use, but for now, we're only interested in the Bots that actually "crawls" the web and collects documents to build a searchable index for the different search engines. The script starts at a website and follows every hyperlink on every page. So they can say that everything on the web will ultimately be found and connected, as the so called "spider" crawls from one website to another. Search engines may run thousands of instances of their web crawling programs simultaneously, on multiple servers. A web bot visits one of your pages, it fetch the site's content into a database. Once a page has been fetched, the text of your page is loaded into the search engine's index, which is a huge database of words, and where they occur on different web pages. All of this might sound too technical for many people, but it's important to know the fundamentals of how a web bot works. So, there are basically three steps that are involved in the web scraping procedure. First, the search bot starts by crawling pages of your site.[12] Then it continues indexing the words and content of the site, and finally it visit links that are found in your site. When the crawler doesn't find a page, it will ultimately be deleted from the index. However, some of the crawlers will check again for a second time to verify that the page really is offline. The first thing a crawler is supposed to do when it visits your website is look for a file called "robots.txt". This file contains instructions for the crawler on which parts of the website to index, and which parts to ignore. The only way to control what a crawler sees on your site is by using a robots.txt file. All crawlers are supposed to follow some rules, and the major search engines do follow these rules for the most part. Fortunately, the major search engines like Google or Bing are finally working together on standards.

The general process that bot takes is as follows:

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

The working of web crawler shows that it is recursively keep on adding newer URLs to the database repository of the search engine. This shows that the major function of a web crawler is to add new links into the frontier and to choose a

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 7.542

|| Volume 9, Issue 10, October 2021 ||

| DOI: 10.15680/IJIRCCE.2021.0910013 |

recent URL from it for further processing after every recursive step [7]. Flow of basic crawler is shown in figure 1.0 [12]

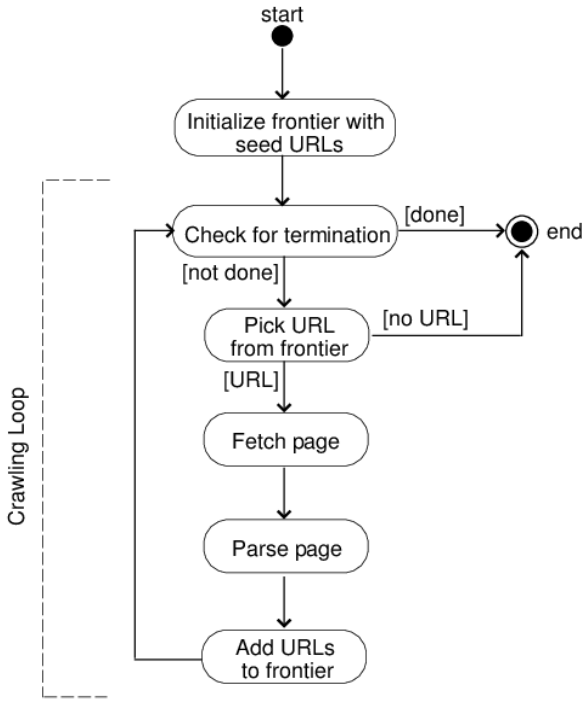

Fig 1.0 flow of web Bot

IV. TYPES OF WEB SCRAPING BOTS

INCREMENTAL WEB BOT

A traditional bot, in order to refresh its collection, periodically replaces the old documents with the newly downloaded documents. On the contrary, an incremental bot incrementally refreshes the existing collection of pages by visiting them usually; based upon the estimate as to how often pages change. It also exchanges trivial pages by new and more important pages. It resolves the matter of the freshness of the pages. The advantage of incremental crawler is that only the precious data is provided to the user, thus network bandwidth is saved and data enrichment is achieved.[12]

GENERAL PURPOSE WEB BOT

General-purpose web bot collects and process the entire contents of the Web in a centralized location, so that it can be indexed in advance to be able to respond to many user queries. In the early stage when the Web is still not very large, simple or random spider method was enough to index the whole web. However, after the Web has grown very large, a bot can have large coverage but rarely refresh its crawls, or a bot can have good coverage and fast refresh rates but not have good ranking functions or support advanced query capabilities that need more processing power. Therefore, more advancecrawling methodologies are needed due to the limited resources like time and network and width.[11]

FOCUSED WEB BOT

Focused bot is the Web bot that tries to download pages that are associated with one another. It collects documents which are specific and associated with the given topic. it's also referred to as a subject bot due to its way of working. The focused bot determines the subsequent – Relevancy, Wayforward. It determines how far the given page has relevancy to the actual topic and the way to proceed forward. The benefits of focused web bot is that it's economically feasible in terms of hardware and network resources, it can reduce the amount of network traffic and downloads. The search exposure of focused web bot is also huge.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 7.542

|| Volume 9, Issue 10, October 2021 ||

| DOI: 10.15680/IJIRCCE.2021.0910013 |

BREADTH FIRST WEB BOT

It starts with a little set of pages then explores other pages by following links within the breadth-first fashion. Actually, web pages aren't traversed strictly in width first fashion but may use a spread of policies. For example, it may crawl most important pages first. [12]

ADAPTIVE WEB BOT

Adaptive crawler is classified as an incremental type of crawler which will continually crawl the entire web, based on some set of crawling cycles. The adaptive model used would use data from previous cycles to make a decision which pages should be checked for updates. Adaptive Crawling also can be viewed as an extension of focused crawling technology. It has the basic concept of doing focus crawling with additional adaptive crawling ability. Since the web is changing dynamically, adaptive crawler is designed to crawl the net more dynamically, by additionally taking into consideration more important parameters like freshness or up to date-ness, whether pages are obsolete, the way pages change, when pages will change, how often pages change and etc. These parameters will be added into the optimization model for controlling the crawling strategy, and contribute to defining the discrete period of time and crawling cycle. Therefore, it is expected that more cycles the adaptive crawler goes in operation, more reliable and refined will the output results.[11]

V. CONCLUSION

Web Bot is the imperative source of information retrieval which traverses the Web and downloads web documents that suit the user's need. Web bot is used by the search engine and other users to regularly ensure that their database is up-to-date. The overview of different bot technologies has been presented in this paper. When only information about a predefined topic set is required, “focused bot” technology is being used. Compared to other bot technology the Focused bot technology is designed for advanced web users focuses on particular topic and it does not waste resources on irrelevant material.

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
10. Dr Rajender Nath, Khyati Chopra, “Web Crawlers: Taxonomy, Issues & Challenges”, International Journal of Advanced Research in Computer Science and Software Engineering, Volume 3, Issue 4, April 2013
11. S.S. Dhenakaran and K. Thirugnana Sambanthan,” WEB CRAWLER - AN OVERVIEW”, International Journal of Computer Science and Communication Vol. 2, No. 1, January-June 2011, pp. 265-267.
12. Mini Singh Ahuja, Dr Jatinder Singh, Bal Varnica, “Web Crawler: Extracting the Web Data”, International Journal of Computer Trends and Technology (IJCTT) – volume 13 number 3 – Jul 2014.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 7.542

|| Volume 9, Issue 10, October 2021 ||

| DOI: 10.15680/IJIRCCE.2021.0910013 |

13. Mridul B. Sahu, Prof. Samiksha Bharne, "A Survey On Various Kinds Of Web Crawlers And Intelligent Crawler", International Journal of Scientific Engineering and Applied Science (IJSEAS) – Volume-2, Issue-3, March 2016 ISSN: 2395-3470.
14. J. Cho, Hector Garcia-Molina "Effective Page Refresh Policies for Web Crawlers", ACM Transactions on Database Systems, Vol. 28, No. 4, December 2003,

INNO SPACE
SJIF Scientific Journal Impact Factor
Impact Factor: 7.542

doi cross ref

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

9940 572 462 6381 907 438 ijircce@gmail.com

www.ijircce.com

Scan to save the contact details