ABSTRACT

Despite the fact that their citation counts may be low, Researchers frequently examine Google Scholar (GS), Web of Science & Scopus, which are all used in research reviews.. From 252 Google Scholar subject categories there are 2,299 documents which have 2,448,055 citations in 2006. When GS, WoS and Scopus are compared then it is found that Among all areas, GS has the highest percentage of citations, approximately 67%-96% whereas Scopus has 35%-77% and WoS has 27%-73%.
In GS, About 48% to 65% of citations came from non-journal sources such as books, conference papers, and unpublished papers. A further 19%-38% are non-English, and they tend to be less frequently cited than those in Scopus and WoS . So to extract the data from these publication sources we use web scrapping and web crawling. In this paper we will discuss The crawler shoots queries on the search interface by using keywords related to a user's topic of interest. This allows crawlers to gather links relevant to a particular domain without actually having to dig into the root domain of the domain. The proposed work provides the most pertinent information on a specific domain based on keywords without crawling through a large number of irrelevant links between them. The information from web is found by search engine. These search engines are of two types - human powered and crawler based. This user-generated search engine brings together a collection of high-quality websites that have been hand-picked by users.
In a crawler-based search engine, three components are involved: A search-ranking algorithm, a crawler, and an indexer.

TABLE OF CONTENTS

LIST OF FIGURES

CHAPTER-1

INTRODUCTION

Google Scholar, which was launched in November 2004, transformed the way researchers and the general public searched for academic content by bringing the convenience of Google search to academic surroundings. Google Scholar was created as a search engine with the goal of determining the most pertinent publications for a particular query based on citation data . As of yet, the data of Google Scholar cannot be accessed in bulk. The free software Publish or Perish makes availability free citation of Google scholar data without database subscription. At present, Using third-party web scraping software is the only way to get more data from GS than Publish or Perish allows. Moreover The web repositories of scientific journals and agencies often provide data to funders and research institutions. Using Google Scholar data, researchers can analyze the performance of individual publications. Using these tools, stakeholders like search committees can find the most suitable candidate for a job. We provide a tool for verifying the publications of a group of researchers working in the same field. Online scrapers provide the needed data for the tool, The data for the tool is provided using a web-based user interface. We gathered the challenges we encountered scraping scientific web repositories throughout our research and summarized them here. Even when permission has been granted for data use, scraping is still necessary in some cases. During web crawling, A search engine is a piece of computer software that looks for information on the Internet. Using indexed databases, this program Looks for a result according to the user's query. Search engines use regularly updated indexes in order to cope with fast and efficient search results. These search engines keep their database index up to date by scouring a huge chunk of the internet. These search engines are different from web directories because directories are maintained by human editors whereas search engines use crawlers. Web spiders and web robots are terms used to describe web crawlers. These are computer viruses that recursively scan the internet using hyperlinks. Web crawling, often known as spidering, is the process of a crawler extracting data from the internet.

1.1 DOMAIN DESCRIPTION

1.1.1 WEB CRAWLER

A web crawler is a program/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 designed 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. These pages are retrieved by a Web crawler that is an automated Web browser that follows each link it sees.

1.1.2 The History of Web Crawler

The first Internet “search engine”, a tool called “Archie” — shortened from “Archives”, was developed in 1990 and downloaded the directory listings from specified public anonymous FTP (File Transfer Protocol) sites into local files, around once a month [5], [6]. In 1991, “Gopher” was created, that indexed plain text documents. “Jughead” and “Veronica” programs are helpful to explore the said Gopher indexes [7], [8], [9], [10]. With the introduction of the World Wide Web in 1991 [11], [12] numerous of these Gopher sites changed to web sites that were properly linked by HTML links. In the year 1993, the “World WideWebWanderer” was formed the first crawler [13].

Although this crawler was initially used to measure the size of the Web, it was later used to retrieve URLs that were then stored in a database called Wandex, the first web search engine [14]. Another early search engine, “Aliweb” (Archie-Like Indexing for the Web) [15] allowed users to submit the URL of a manually constructed index of their site. The index contained a list of URLs and a list of user wrote keywords and descriptions. The network overhead of crawlers initially caused much controversy, but this issue was resolved in 1994 with the introduction of the Robots Exclusion Standard [16] which allowed web site administrators to block crawlers from retrieving part or all of their sites. Also, in the year 1994, “WebCrawler” was launched [17] the first “full text” crawler and search engine. The “WebCrawler” permitted the users to explore the web content of documents rather than the keywords and descriptors written by the web administrators, reducing the possibility of confusing results and allowing better search capabilities. Around this time, commercial search engines began to appear with [18], [19], [20], [21], [22], [23], [24] and [25] being launched from 1994 to 1997 [26]. Also introduced in 1994 was Yahoo! , a directory of web sites that was manually maintained, though later incorporating a search engine. During these early years Yahoo! and Altavista maintained the largest market share [26]. In 1998 Google was launched, quickly capturing the market [26]. Unlike many of the search engines at the time, Google had a simple uncluttered interface, unbiased search results that were reasonably relevant, and a lower number of spam results [27]. These last two qualities were due to Google’s use of the PageRank [28] algorithm and the use of anchor term weighting [29]. While early crawlers dealt with relatively small amounts of data, modern crawlers, such as the one used by Google, need to handle a substantially larger volume of data due to the dramatic enhance in the Web

1.1.3 Working of Web Crawler

The working of Web crawler is beginning with initial set of URLs known as seed URLs. They download web pages for the seed URLs and extract new links present in the downloaded pages. The retrieved web pages are stored and well indexed on the storage area so that by the help of these indexes they can later be retrieved as and when required. The extracted URLs from the downloaded page are confirmed to know whether their related documents have already been downloaded or not. If they are not downloaded, the URLs are again assigned to web crawlers for further downloading. This process is repeated till no more URLs are missing for downloading. Millions of pages are downloaded per day by a crawler to complete the target. Figure 2 illustrates the crawling processes.

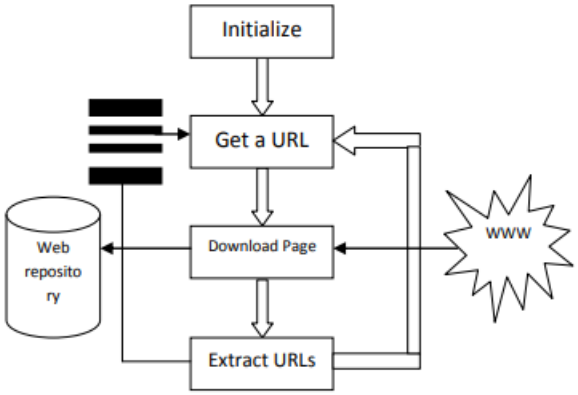

Figure 2: Flow of a crawling process

The working of a web crawler may be discussed as follows:

- Selecting a starting seed URL or URLs
- Adding it to the frontier ¬ Now picking the URL from the frontier
- Fetching the web-page corresponding to that URL
- Parsing that web-page to find new URL links
- Adding all the newly found URLs into the frontier
- Go to step 2 and reiterate till the frontier is empty

Thus a web crawler will recursively keep on inserting newer URLs to the database repository of the search engine. So we can see that the major function of a web crawler is to insert new links into the frontier and to choose a fresh URL from the frontier for further processing after every recursive

step. To classify whether a transaction is fraud or legal, a large collection of transaction is required. The dataset is downloaded from Kaggle database. In this section the methodology followed is discussed in detail.

1.1.4 CRAWLING TECHNIQUES

There are a few crawling techniques used by Web Crawlers, mainly used are:

A. General Purpose Crawling A general purpose Web Crawler collects as many pages as it can from a particular set of URL's and their links. In this, the crawler is able to fetch a large number of pages from different locations. General purpose crawling can slow down the speed and network bandwidth because it is fetching all the pages.
B. Focused Crawling A focused crawler is designed to collect documents only on a specific topic which can reduce the amount of network traffic and downloads. The purpose of the focused crawler is to selectively look for pages that are appropriate to a pre-defined set of matters. It crawl only the relevant regions of the web and leads to significant savings in hardware and network resources.
C. Distributed Crawling In distributed crawling, multiple processes are used to crawl and download pages from the Web.

1.1.5 PARALLEL CRAWLER

Now search engines do not depend on a single but on multiple web crawlers that run in parallel to complete the target. While functioning in parallel, crawlers still face many challenging difficulties such as overlapping, quality and network. Given below Figure illustrates the flow of multiple crawling processes.

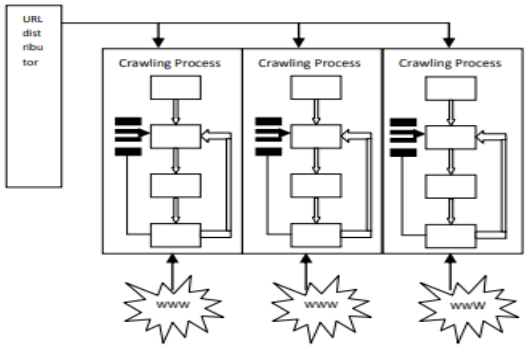

Figure 3: Flow of multiple crawling processes.

Scraping Scientific Web Repositories: Challenges and Solutions for Automated Content Extraction Authors: Philipp Meshenmoser ; Bela Gipp ; Norman Meuschke ; 2015 10th International Conference on Computer Science & Education (ICCSE) Aside from improving the visibility and accessibility of scientific publications, many scientific Web repositories also assess researchers'quantitative and qualitative publication performance, e.g., by displaying metrics such as the h-index. These metrics have become important for research institutions and other stakeholders to support impactful decision making processes such as hiring or funding decisions. However, scientific Web repositories typically offer only simple performance metrics and limited analysis options. Moreover, the data and algorithms to compute performance metrics are usually not published. Hence, it is not transparent or verifiable which publications the systems include in the computation and how the systems rank the results. Many researchers are interested in accessing the underlying scientometric raw data to increase the transparency of these systems. In this paper, we discuss the challenges and present strategies to programmatically access such data in scientific Web repositories. We demonstrate the strategies as part of an open source tool (MIT license) that allows research performance comparisons based on Google Scholar data. We would like to emphasize that the scraper included in the tool should only be used if consent was given by the operator of a repository. In our experience, consent is often given if the research goals are clearly explained and the project is of a non-commercial nature

Data Capturing: Methods, Issues and Concern

Authors: Afiqah Amirah Hamzah, Saiful Farik Mat Yatin, Nurul Athirah Ismail, Siti Faridah Ghazali , Dr. C. S. Mala International Journal ofEngineering Research & Technology NCESC – 2018 Conference Proceedings Data capturing is the method of putting a document into an electronic format. Many organizations implement to automatically identify and classify information and make the available within particular systems. It takes documents content, in any format, and converts it into