Retrieval of Data with Web Crawling to Benefit Data Science

[1]Mr. K Srinivas, Assistant Professor, Department of Computer science and engineering

Geethanjali college of engineering and Technology, Hyderabad, TS, India 501301

[2]Mr. M Vijay Bhaskar Reddy , Assistant Professor, Department of Computer science and engineering

Geethanjali college of engineering and Technology, Hyderabad, TS, India 501301

[3]Mr. M Santhosh Kumar , Assistant Professor, Department of Computer science and engineering

Geethanjali college of engineering and Technology, Hyderabad, TS, India 501301

Abstract:

As the digitalization is increasing and Internet is flooded with bulk data, the difficulty in handling the same is also increasing. Several techniques are coming in scenario now every day to tackle the challenges to handle data efficiently. The Information retrieval from data provides the users with important information required for analysis. Information retrieval helps to have useful and enough information gathered at fast pace. In this paper, the different approaches to collect data utilizing web crawling has been discussed and an approach has implemented and discussed. The comparison of different techniques has also been done in the paper.

Keywords: Web crawling, Information Retrieval, data, spider.

I. INTRODUCTION

The technique of Information Retrieval handles retrieving, evaluation, storage and organization of information from different repositories of texual information. The tool behaves like an assistant to users for helping them in finding information they need but it is not capable to give answers to the questions. It will help to locate the documents and data where the related information can be found out i.e. helps to locate relevant documents. A good Information retrieval system will provide just those documents which are relevant to the user. Figure 1 represents process and technique of Information retrieval.

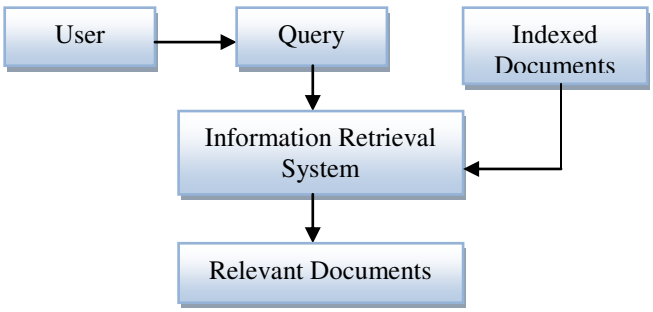

Figure 1: Information Retrieval

The above figure clearly shows that the user who needs information will have to put a query based request in Natural Language. Later, the Information Retrieval system will give related output for the query in form of necessary information required by user or documents.

One of the main problems attached with Information Retrieval Systems is of Ad Hoc Removal problem in which it is required by the user to put a query in natural language which elaborates the necessary information required. Then only Information Retrieval system will give output in the form of related documents.

For example, If a user search something on search engine with some specific requirement but along with the related documents and information, they get irrelevant data also. This happens because of Ad Hoc Removal problem.

Web Crawling

One of the techniques for Information Retrieval is Web crawling which is a method to collect Web pages. The aim of Web crawler is to gather the important web pages fast and quick as soon as possible along with efficiency.

Some of the features that Web Crawler should carry are:

- *Robustness:* The internet comprises of different servers which has spider traps and web page generators which can make crawler stuck in getting numerous number of webpages. The crawler developed should be capable to tackle such traps.
- *Distributed:* The web crawler should be able to run on different machines efficiently.
- *Extensible:* The designed crawler should be modular in nature so that it can easily adapt new features, new environment and protocols.

- *Quality:* The crawler developed should be focused on catching important and relevant pages on priority.

Applications of Web Crawling in Data Science

The web crawling provides several applications in area of Data Science. Some of them include:

- *Real time analytics:* The projects based on Data science required some real time data for analysis which can be collected by web crawling utilizing data extraction at high frequency and somewhat near real time data can be collected for analytics.

- *Predictive Modeling:* This is about data analysis and utilizing probability for outcome prediction for future. There are several predictors present in model which can affect the results. The necessary information required for developing predictor can be collected from website by using web crawler.

- *Natural Language Processing:* This is utilized for equipment of machines with the capability of interpretation. The data on internet is highly diverse and same can be utilized in Natural Language processing. For example, e-commerce websites data, reviews, tweets, blogs text are very useful in NLP.

- *Training machine Learning Models:* The training datasets can help models of machine learning to perform clustering, classification, etc.

The working of a web crawler is shown in figure 2.

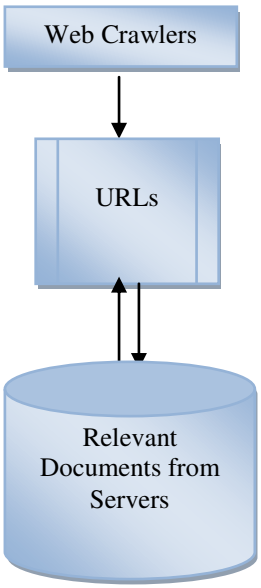

Figure 2: Web Crawling process

II. RELATED WORK

In 1980, along with the information retrieval and Web crawlers, the Genetic Algorithms were being utilized. Researchers (V. V. Raghavan, 1987) became the first one to try out GA with web crawlers and information retrieval. The first search engine named "Archie" came out from the name Archives has been developed in 1990 and utilized FTP for downloading files from various sites (P. J. Deutsch, 1991). The authors (M. Gordon, 1987) improved an approach of using Genetic Algorithms with Web Crawler. The fashion was then followed by Yang et al. who improved the weights of keywords related to the specific document topic. Petry et al. also used genetic algorithms to boost the process of retrieving content from a collection of weighted indexed documents, by changing the weights of queries. Genetic algorithm (GA) has been used by Chen, et al. to develop a global and personal spider. A set of standard experiments are performed to compare the performance and efficiency of the Best first search (BFS) and GA algorithms based spiders.

Using metagenetic algorithm is an easy to learn and adaptive approach. Judy Johnson et al. used genetic algorithm in emerging scenarios by evaluating the previous search queries made by the user. The resultant queries are criticized on the basis of rank function that combines text contents and hyperlinks of the query. This results in better performance compared to earlier used best first strategies. An Intelligent crawler with online processing facility is proposed by Milad shokouhi et al.. He introduced an intelligent agent that uses genetic algorithm named Gcrawler. Gcrawler finds the best path for crawling and expands the keyword by using genetic algorithm. Superiority of this algorithm is there is no need to interact with users. Ibrahim Kushchu et al. presented applications of Evolutionary and Adaptive Systems (EAS) in Webbased (IR). Research is done in two phases: In first phase, researchers often use GA or GP as an optimizer in the contexts of clustering, query improvement, or keyword selection. In second phase, adaptive intelligent agents are employed and evolutionary methods are used as learning mechanisms. EAS approach is more domain independent and flexible to create dynamic web. Jialun Qin et al. used a genetic algorithm with focused crawling for improving its crawling performance. An advanced crawling technique is proposed to build domain-specific collections for web search engines that consolidate a global search algorithm. The proposed work apply a genetic algorithm to the focused crawling process to find more relevant Web pages in order to overcome the drawbacks of traditional focused crawlers. Web site clustering consists in finding meaningful groups of related web sites. Esteban Meneses et al. analyzed two models and four associated algorithms: vector models are

analyzed with k-means and self-organizing maps (SOM), graphs are analyzed with simulated annealing and genetic algorithms and these algorithms are tested by clustering some web sites.

Real time analytics: Many data science projects require real time or near real time data for analytics. This can be facilitated by crawling websites using a low latency crawl. Low latency crawls work by extracting data at a very high frequency that matches with update speed of target site. This gives near real time data for analytics.

Predictive modeling: Predictive modeling is all about analyzing data and using probability to predict outcomes for future scenarios. Every model includes a number of predictors, which are variables that can influence the future results. The data required for making relevant predictors can be acquired from different websites by using web scraping. A statistical model is formulated once the processing is done.

Natural language processing: Natural language processing is used to equip machines with the ability to interpret and process natural languages used by humans like English as opposed to a computer language like Java or Python. As it's difficult to determine a definite meaning for words or even sentences in natural languages, natural language processing is a vast and complicated field. Since the data available on the web is of diverse nature, it happens to be highly useful in NLP. Web data can be extracted to form a large text corpora which can be used in Natural language processing. Forums, blogs and websites with customer reviews are great sources for Natural language processing.

Training machine learning models: Machine learning is all about equipping machines to learn on their own by providing them training data. Training data could differ according to individual cases. However, data from the web is ideal for training machine learning models for a wide range of use cases. With training data sets, machine learning models can be taught to do correlational tasks like classification, clustering, attribution etc. Since the performance of a machine learning model will depend on the quality of training data, it is important to crawl only high quality sources.

Provided with the training data, machine learning programs learn to do correlational tasks like classification, clustering, attribution etc. Here, the efficiency and power of the machine learning program will hugely depend on the quality of training data.

An improved search engine has been proposed by M. Koorangi et al. to increase the efficiency of the web searches. Stationary and mobile agents have been used collectively to enhance the performance.

III. METHODOLOGY

The web Crawler developed to collect data and the methodology and flow of work for the same is shown in figure 3.

The crawler is being developed to collect data from e commerce website. This crawler is able to call 2000API in a minute.

The developed crawler first requires all URLs to visit as an input. Then it goes on those URLs and collect data as requested by the user.

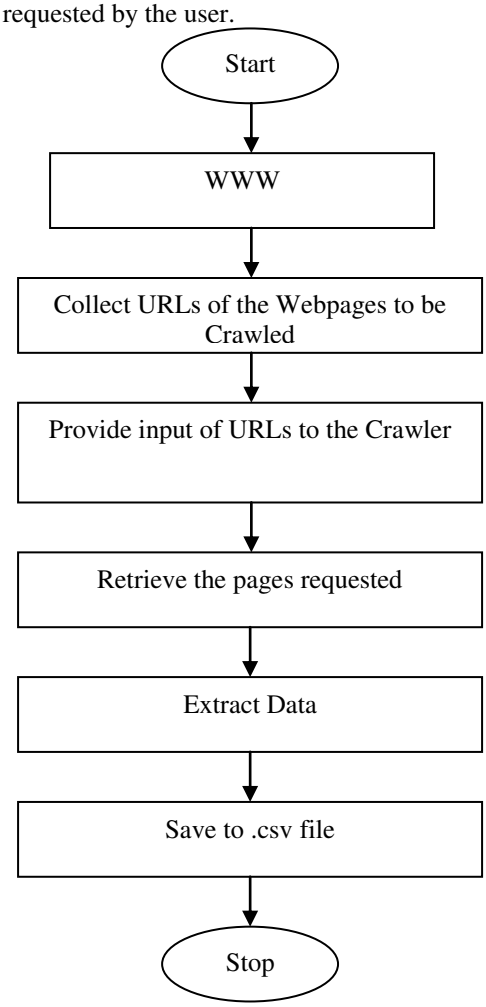

Figure 3 : Flow of Crawler

IV. RESULTS

The presented work is applied on an e-commerce website for collection of data. The properties of the dataset is shown in table 2.

Table 2: Dataset Properties

Parameter	Value
Number of products	7
Data Type	Customer Reviews
Product type	Any product on E-commerce site.
Image Size	20,000 to 50,000 reviews
Technique	Web crawling using Spider

The collection process is here defined. The collection of data is done utilizing different URLs and the maximum 40,000 reviews have been collected on products. Figure 4 shows different types of products taken into consideration for collection of data. Figure 5 shows sample of data collected using Web crawler with Spider for mobile reviews.

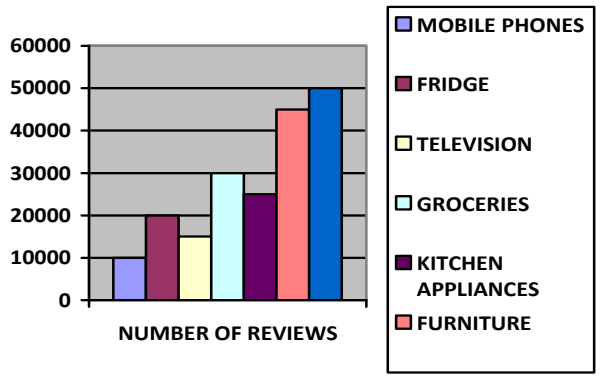

Figure 4: Number of Reviews from different products

user	date	rating	opinion
Ravindra	on 4 June 2018	4	Good working condition super
Munna	on 25 October 2017	4	Good mobile good price
vipin	on 4 January 2017	5	Nice product...fast delivery
Amazon Customer	on 5 October 2017	5	Good product.
Amazon Customer	on 18 August 2017	3	Delivery was good. On this price best budget phone

Figure 5: Data Collected (Sample)

V. CONCLUSION

This paper is about very important topic of Information Retrieval using Web Crawling in the research field. The collection of data is one of the crucial part in research field. While working with data science, the information retrieval becomes very important. The concept of web crawling has brought a drastic change in collection of data. This paper has discussed about different techniques and methodologies available for crawling web to collect data. The applications of Web crawling in area of data science is also discussed. In this work, the crawler developed to collect the data and for the same, the results achieved are discussed here.

With Data Science being a growing topic in the software industry, and machine-learning being at the forefront of the technological sphere, there are new applications to make the job easier and faster being developed every day. And with that exciting growth, we are constantly at an influx of new creators, scientists, and analysts alike joining the ranks of lifetime learning.

REFERENCES

[1] V. V. Raghavan, B. Agarwal, "Optimal determination of user-oriented clusters: an application for the reproductive plan. In: Genetic Algorithms and Their Applications", Proceedings of the Second International Conference on Genetic Algorithms, 28–31 July 1987, at the Massachusetts Institute of Technology, Cambridge, MA. L. Erlhaum Associates, Hillsdale,1987

[2] P. J. Deutsch, "Original Archie Announcement", URL http://groups.google.com/group/comp.archives/msg/a773 43f9175b24c3?output=gplain, 1990.

[3] M. Gordon, "Probabilistic and genetic algorithms in document retrieval" Commun. ACM 31(10), 1208–1218, 1988.

[4] Johnson, J., Tsioutsiouliklis, K., Lee Giles, C.: Evolving strategies for focused web crawling. In: ICML, pp. 298–305 (2003)

[5] Shokouhi, M., Chubak, P., Raeesy, Z.: Enhancing focused crawling with genetic algorithms. In: International Conference on Information Technology: Coding and Computing, ITCC 2005, vol. 2, pp. 503–508. IEEE (2005)

[6] Kushchu, I.: Web-based evolutionary and adaptive information retrieval. IEEE Trans. Evol. Comput. 9(2), 117–125 (2005)

[7] Qin, J., Chen, H.: Using genetic algorithm in building domain-specific collections: an experiment in the nanotechnology domain. In: Proceedings of the 38th Annual Hawaii International Conference on System Sciences, HICSS 2005, p. 102b. IEEE (2005)

[8] Meneses, E.: Vectors and graphs: two representations to cluster web sites using hyperstructure. In: Fourth Latin American Web Congress, LA-Web 2006, pp. 172– 178. IEEE (2006)

[9] Koorangi, M., Zamanifar, K.: A distributed agent based web search using a genetic algorithm. Int. J. Comput. Sci. Netw. Secur. 7(1), 65–76 (2007)

[10] https://towardsdatascience.com/how-web-crawling-benefit-data-science-a6ff0bd4cd1