International Journal of Technical Innovation in Modern Engineering & Science (IJTIMES)
Impact Factor: 3.45 (SJIF-2015), e-ISSN: 2455-2585
Volume 3, Issue 11, November-2017

Review on Data Mining based Spider Bot

[1]Mr. Tushar Gonawala, [2]Ms. Hima Khimani, [3]Prof. Vatsal Shah

[1]IT Department, Birla Vishvakarma Mahavidyalaya,
[2]IT Department, Birla Vishvakarma Mahavidyalaya,
[3]IT Department, Birla Vishvakarma Mahavidyalaya,

Abstract-World Wide Web is a huge pool of data and getting accurate and relevant data as per the query is essential. Crawling algorithms are used in traversing the pages and on the derived data text mining and text classification is performed that would narrow down the search results and obtain crucial information that can be further utilized. This paper reviews the researches on web crawling algorithms used on searching, text mining techniques and text classifiers.

Keywords- Spider Bot, Web Crawler, Web Bot, Text Mining, Text Classification, Crawling Algorithms

I. INTRODUCTION

Recent statistics state that, web search generates more than 13% of the traffic to websites [5]. On a particular search, hundreds and thousands of search result appear in form of links and the search engine has to deal with the data that is growing exponentially. It is necessary for the search engine to fulfill the interest of the user so that optimum outputs can be drawn. But yet it becomes difficult for the end user to derive proper conclusions on the query.
In such cases, web mining play a very important role. For mining relevant information or opinions, it is necessary that large amount of data should be traversed with great speed and accuracy. Once the data is collected come the analysis part that is taken care using a lot of machine learning and data mining algorithms. The key to find proper output and draw conclusions lies in tag based classification. For proper mining the crawler not only has to go through the content of the web pages but also have to create a hierarchical structure of the links for enhanced results.

II. SYSTEM ARCHITECTURE

2.1 Web Crawler

A Web Crawler generally known as Spider Bot or Web Spider or Web Bot is an auto scripted program that browses the World Wide Web in a methodical manner in order to find relevant information and narrow down the number of results or the outcome.There are few questions that come in mind while the Web Crawler does its work.
How the targets are selected? The size of web is huge and is increasing day by day and only 60 percent of it is indexed [6]. Not all webpages can be traversed. Only a fraction of segment of web is traversed by the crawler, so it becomes important to prioritize the web pages so that we get relevant pages in the first few downloads.
Where to start? Crawling can be started from any seed URL, but it should be kept in mind that the page references in the seed URL should not refer back to the same page. Doing so would restart the crawl. Good seed URL can be obtained from Google or Yahoo by entering the keyword as they are the most popular search engines whose results are prominent [7].
Various strategies are used for Web Crawler

A. Breadth First Search Algorithm
In this algorithm, all the neighboring nodes or URLs present at the same level are traversed. Starting from the root node and then searching all the URLs a level below the root node i.e. the root URL. If the desired result is obtained and is optimum the search is terminated and if not then it proceeds towards the next lower level and so on. Breadth First Search is best suited for the results that are available at the initial levels of a deeper tree. If the results are located at the deeper level then it would create problem as more time would be required for traversing.

B. Depth First Search
This algorithm has a tendency to do deeper inside the tree. The child URLs are traversed first and not the neighboring URLs. Once the leaf node is reached, the algorithm backtracks and the next unvisited URL in the stack is visited. This continued till the required output is obtained. Depth First Search Algorithm is

feasible if there are less number of branches. If the number of branches increase it might end up in an infinite loop [8].

C. Page rank Algorithm

A certain value is assigned to each and every page based on the citation and back links on that page. This value is known as Page Rank. Page Ranks are calculates as:

PR(A)=(1-d)+d(PR(T1)/C(T1)+…+ PR(Tn)/C(Tn))

PR(A): Page Rank of a given Page

d: Dumping factor

Ti: Links

In order to find the Page Rank for a page, called PR(A), we need to find all the pages that link to page A and Out Link from A. We find a page T1, which has link from A then page C(T1) will give no. of Outbound links to page A. We do the same for T2, T3 and all other pages linking to Main page A – and Sum of the values will provide Rank of the web page [1].

Tian Chong [9] proposed a new type of algorithm for generating Page Rank by combining classified tree and the original Page Rank Algorithm. This enabled to construct a tree using a large number of users' similar search results which helped reduce the problem of outdated pages and also increase efficiency and effectiveness of the search.

D. Focused Crawling algorithm

A crawler considers those pages with higher significance that are categorized as a function of similarity. In this approach we can intend web crawler to download pages that are similar to each other, thus it would be called focused crawler or topical crawler [10].

2.2 Text Mining

Text mining that is often referred as text data mining, extracts patterns, information and knowledge from the unstructured data which is obtained from a huge pool of resources. There are a number of techniques for text mining like Document classification (text classification, document standardization), information retrieval (keyword search / querying and indexing), document clustering (phrase clustering), natural language processing (spelling correction, lemmatization, grammatical parsing, and word sense disambiguation), information extraction (relationship extraction / link analysis), and web mining (web link analysis) [11].

A. Information Retrieval

Information Retrieval (IR) is the process of gathering knowledge and necessary information from words and phrases. Different algorithms are used mainly for the functioning of Search Engines such as Google and Yahoo, which are based on query. These algorithms helps in providing results that more relevant and satisfy user's needs.

B. Natural Language Processing

The basic concept of Natural Language Processing (NLP) is to automatically analyze and process unstructured textual data and draw meaning out of it. The entities and instances recognized from the document are matched with a dictionary list that are mapped to the corresponding values. Natural Language (NL) have lot of complexity as the identical words extracted from various sources have different meaning and abbreviations that are difficult to resolve.

2.3 Text Classifier

Supervision techniques such as Text Classifier is based on set of input output examples that are used to train a model on which the classification is based. Previously known examples are the building block and later on unknown examples are categorized automatically.

A. Naïve Bayes Classification Algorithm

Probabilistic learning and classification are the basis on which Naïve Bayes Classification. It assumes that one feature is dependent on another, and this approach is proved right in many of the cases. Let's consider this classifier as a tree, then the parent node is connected only to its child node and no other node.

2.4 Analysis

Gaining the data from all the above sources and going through various classification processes the final result obtained is the most accurate. Lot of preprocessing is to be done also irrelevant and unnecessary information is to be ignored and the focus is to be constrained only on the data that best servers our purpose.

III. CONCLUSION

The main objective of the review paper was tothrow some light on the amount of data being generated and the importance of getting relevant data through search. We also discussed the various we crawling, text mining and text classifier algorithms that helps forbetter, accurate and robust results.

REFERENCES

[1] Apoorv Vikram Singh et al, / (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 5 (5) , 2014, 6689-6691
[2] A Survey of Web Crawler Algorithms: IJCSI International Journal of Computer Science Issues, Vol. 8, Issue 6, No 1, November 2011-
[3]Opinion Mining based Spider Bot: Proceedings of National Conference on New Horizons in IT - NCNHIT 2013
[4] Text Mining: Techniques, Applications and Issues: (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 7 No. 11, 2016
[5] StatMarket. Search engine referrals nearly double worldwide.
[6] Maurice de kunder, "Size of the world wide web", retrieved from http://www.worldwidewebsize.com /8/8/11
[7] Rashmi Janbandhu, Prashant Dahiwale, M.M.Raghuwanshi "Analysis Of Web Crawling algorithms" International Journal on Recent and Innovation Trends in Computing and Communication, Vol. 2
[8] Ben Coppin "Artificial Intelligence illuminated" Jones and Barlett Publishers, 2004, Pg 77.
[9] TIAN Chong "A Kind of Algorithm For Page Ranking Based on Classified Tree In Search Engine" Proc International Conference on Computer Application and System Modeling (ICCASM 2010)
[10] Kim, S. J. and Lee, S. H. "An improved computation of the PageRank algorithm‖ in Proc. Of the European Conference on Information Retrieval
[11] W. He, "Examining students online interaction in a live video streaming environment using data mining and text mining," Computers in Human Behavior, vol. 29, no. 1, pp. 90–102, 2013.