IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661, p- ISSN: 2278-8727Volume 15, Issue 3 (Nov. - Dec. 2013), PP 08-12
www.iosrjournals.org

WSO-LINK: Algorithm to Eliminate Web Structure Outliers in Web Pages

Rachna Singh Bhullar, Dr. Praveen Dhyani
(Computer Science Department, Guru Nanak Dev University, Amritsar, Punjab, India)
(Banasthali University – Jaipur Campus, Jaipur, Rajasthan, India)

Abstract: *Web Mining is specialized field of Data Mining which deals with the methods and techniques of data mining to extract useful patterns from the web data that is available in web server logs/databases. Web content mining is one of the classifications of web mining which extracts information from the web documents containing texts, links, videos and multimedia data available in World Wide Web databases. Further, web structure mining is a kind of web content mining which extracts patterns and meaningful information from the structure of hyperlinks contained in web documents having the same domain. The hyperlinks which are not related to content or the invalid ones are called web structure outliers. In this paper the basic aim is to find out these web structure outliers.*

Keywords- *Outliers, web outlier mining, web structure mining, Web mining, web structure documents.*

I. Introduction

Millions and millions of users are uploading and downloading web data into/from the web databases in World Wide Web. That's why, data in web server logs and databases are increasing exponentially. Updating and retrieving efficient and relevant data from web databases is a major concern. The aim of our research is to develop a new methodology for efficiently and effectively mine useful and relevant data from the web documents having the same domain. Web mining tasks can be divided into three main categories, namely, Web Structure Mining, Web Usage Mining and Web Content Mining. Web Structure Mining mines relevant knowledge and meaningful patterns from the structure of hyperlinks contained in web pages. Web Usage Mining is the application of web mining techniques to mine information from web usage logs. Web Content Mining extracts efficient and relevant information from web pages having text, image, video and hyperlinks as their content [3], [4], [5] and [6]. Web structure mining is a kind of Web content mining as it mines relevant data from the hyperlinks of web documents to be mined by the algorithms of web content mining [1]. Existing Web Content Mining algorithms focus on web documents of same domain; these algorithms do not consider web pages with varying contents of the same domain called the Web Content Outliers. In general, Outliers are the data that are irrelevant in terms of meaning and behavior of the existing data.

II. Outline of work

Section II provides the brief review of related work in web content mining. Section III explains the proposed algorithm. Section IV provides the results while in section V conclusions and future work is summarized.

III. Related Work

Outliers are those data objects which behave differently on the basis of their properties and valuable information that they contain. Outlier Mining is mainly studied in statistics because standard distribution techniques are applied on data objects to find out the outliers. A prior knowledge of data distribution like Poisson, Normal, etc. is mainly required to apply the statistical techniques which are the major setback. Outlier Detection techniques can be of following categories:

1. **Statistical techniques**: The statistical techniques like depth, distance, derivation and density based techniques can be applied on numeric data objects/sets.
2. **Web Text Outlier Mining Algorithm:** This computes the difference in web texts within a certain domain.
3. **WCO-ND algorithm:** This algorithm is designed to determine the similarity between different but related words in text processing.

All the above discussed algorithms make use of web texts present in the various web documents. For example, consider the following scenario for a web search engine.

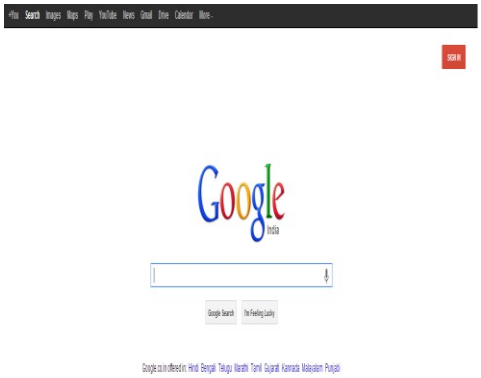


Figure1

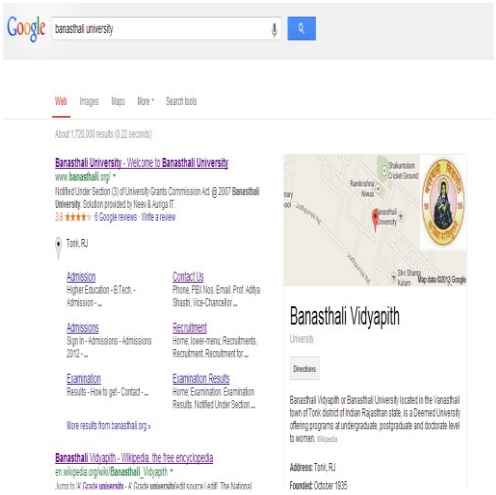
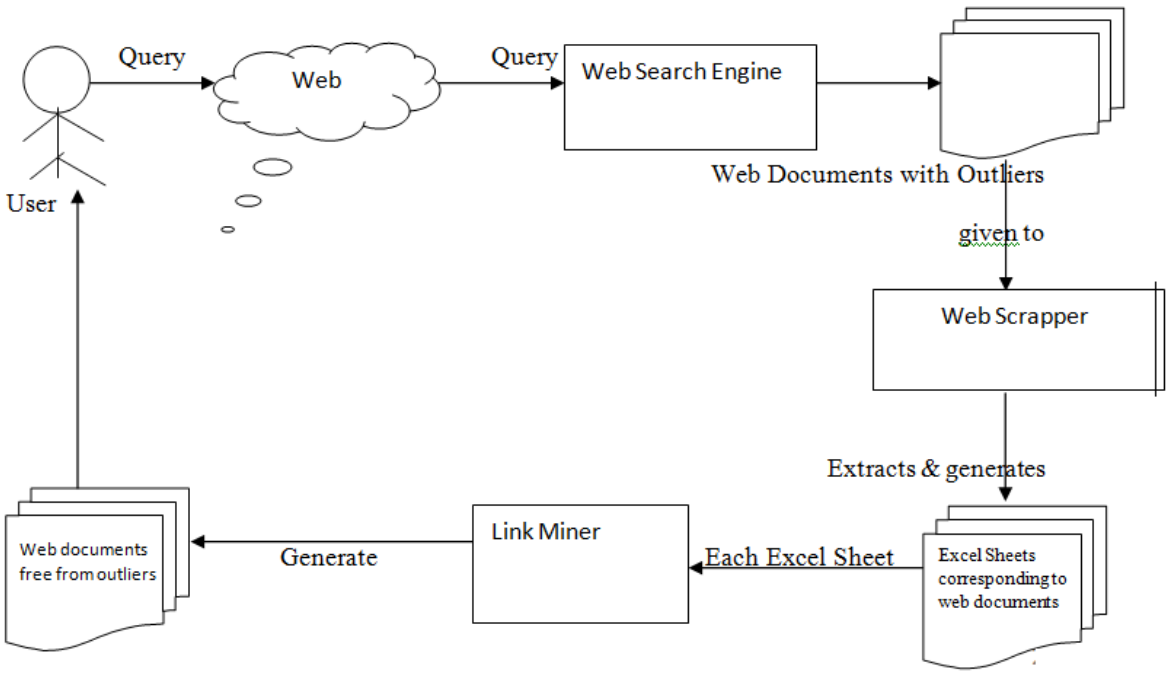


figure2

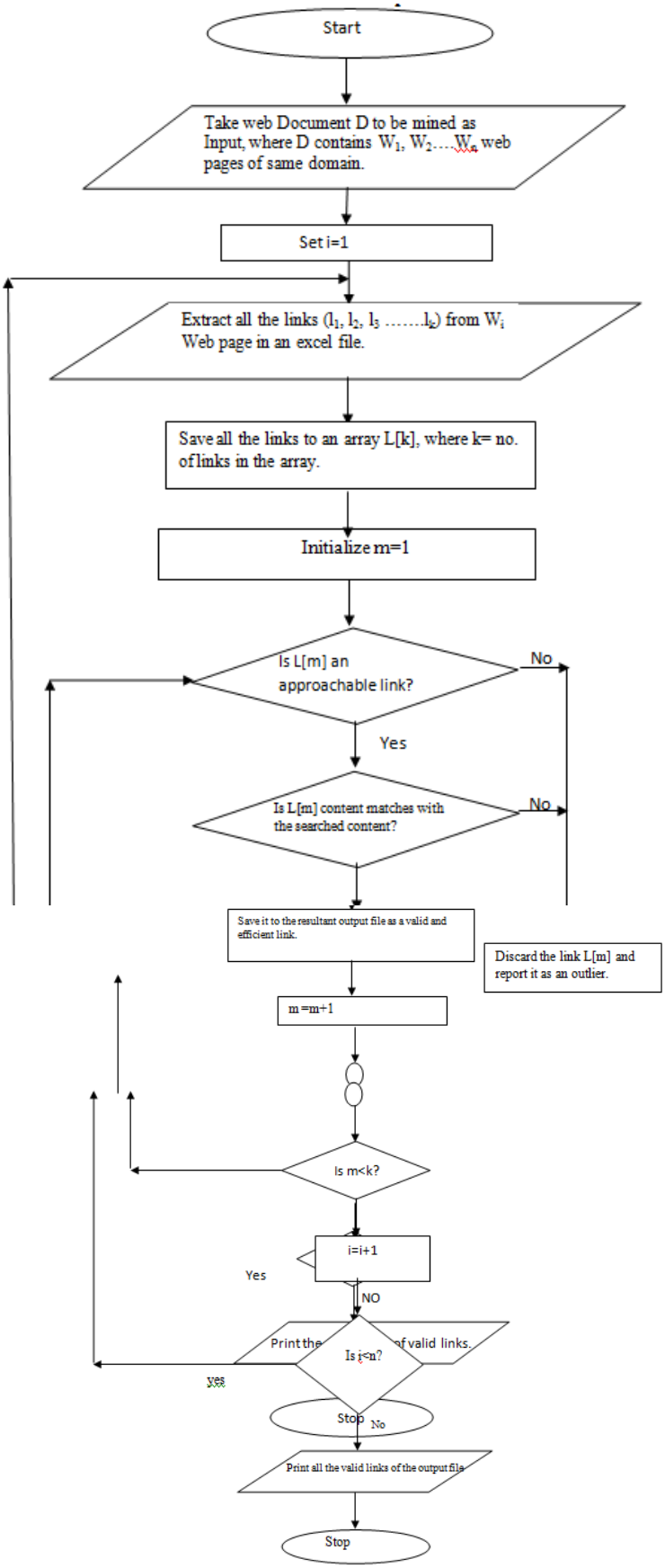
In figure1 Google search engine is there and in figure2, after searching **Banasthali** by Google, list of web documents containing hyperlinks (not the simple text related to Banasthali) is listed. That is why, the above simple web discussed algorithms are not sufficient to yield the desired and efficient output.

IV. Architecture of the proposed system

In the proposed system, query given by the user is searched using a web scrapper. Web search engine, opened in web scrapper, then generates a list of related web pages. Each web page is preprocessed by extracting all the links in an excel file. Now corresponding to each page, a separate excel file on the disk is placed. After this, each excel file is processed by a programming code to eliminate the web structure outliers.



V. Proposed Flowchart



VI. Proposed Algorithm

Step1:- Enter the query on the web search engine opened in the web scrapper.

Step2:- Take the input as document D to be mined.

Step3:- Document D is consisted of:

$$D=\bigcup_{i=1}^{n} W_i$$

Where i=1, 2, 3……n web pages.

Step4:- Initialize i=1

Step5:- Assign $L[K]=\bigcup_{t=1}^{k} L_t$

Where L=name of array whose elements are links from web page W_i

L=name of array whose elements links from webpage W_i

t=1, 2, 3……K where k= total no. of links in W_i

Step6:- Our aim is to find the web structure outlier from W_i which are the links not related to searched content as well as not reachable that means the fake hyperlinks.

Step7:- We can say $D=\bigcup_{I=1}^{n} W_I = \bigcup_{i=1}^{n} \left(\bigcup_{t=1}^{k} L_t\right)_i$

Step8:- for m=1 to k, check whether L_m of L[k](which is equivalent to $L[k]=\bigcup_{m=1}^{k} L_m$) where m=1, 2..k

At first instance, m=1 L [1] is

(i) A valid hyperlink or not by checking through a java code.

(ii) A hyperlink related to the searched content or not.

Step9:- If (i) and (ii) are true then repeat the above step8 for k times.

Step10:- Repeat both the step8 & 9 for n times so that we can remove outliers from all the web pages contained in a document 'D'.

VII. Observations

Elimination of outliers results in the reduction of space and time complexity. Quality of search engine gets increased as web content is efficient and relevant to the searched content. In statistics, we have a measurement to find the quality of refined pages which is known as Precision. It can be defined as the ratio between the number of relevant pages and the total number of relevant documents returned after the elimination of outliers [9].

$$\text{Precision} = \frac{\text{Relevant documents retrieved originally}}{\text{Refined documents retrieved}}$$

VIII. Future Work

Web mining is a growing research area in data mining research. This paper proposes an algorithm to find the outliers to improve the efficiency of web search engine. Future work aims at experimental evaluation and comparative study of our algorithm with results of existing web content mining algorithms.

References

Journals

[1]. Signed approach for mining web content outliers by G.Poonkuzhali, K. Thaiagrajan, K. Sarukesi and G.V.Uma, World Academy of Science Engineering & Technology, 32, 2009.

[2]. Bing Liu, Kevin chen-chuan chang, Editorial special issue on web content mining, SIGKDD Explorations Volume 6, issue 2.

[3]. Hongqili, Zhuang Wu, Xia.ogang Ji research on the techniques for effectively searching and retrieving information from Internet Symposium on Electronic Commerce & Security, IEEE 2008.

[4]. G.Poonkuzhali, K.Thaigarajan, K.Sarukesi, set theoretical approach for mining web content through outlier detection, International Journal on Research & Industrial Applications, Volume 2, January 2009.

[5]. "Chinese web text outlier mining Based on domain knowledge ", by Xia Huosang , Fan Xhaoyan, Pang Liuyan in 2010 Second WRI Global Congress on Intelligent Systems.

Proceedings

[6]. WCO ND-Mine: Algorithm for detecting web content outliers from web documents by Malik Agyemang, Ken Barker, Raja S.Anthajj, proceedings of the 10th IEEE symposium Computer & Communications (ISCC2005).

[7]. G.Poonkuzhali, K.Thaigarajan, K.Sarukesi elimination of redundant links in web pages –Mathematical approach, Proceedings of World Academy of Science, Engineering & Technology, Volume 40,April 2009, PP 555-562.

[8]. Jhonshon T, Kwok I, Ng R. "Fast computation of 2-D Depth Contours" , In Proceedings of KDD 98, PP 224-228.

[9]. Knorr E.M., Ng R.T. "Algorithm for Mining distant based outliers in large datasets" in Proceedings of the 24th VLDB conference, New York, 1998, PP 392-403.

Books:

[10]. Data mining Concepts and Techniques by Jiawei Hen and Micheline Kamber.