Data Analytics For Web Structure Mining In Business Website

Sathish Kumar G, Ramya R, Dinesh P S, Prabha Devi D, Janani A P

Abstract: The huge amount of information is available in the World Wide Web (WWW) which creates the interesting factor for web mining. The access pattern of the webpages or websites, properties of the documents, behavior of the distinct customers can be analyzed by the web mining methods. Web miner software is employed to discover the similar patterns in the websites and is used to certify the information which is extracted from the pages. Our focus is to enhance the profit by web services in business domains. We have performed the algorithmic strategies for the effective implementation of our proposed technique. Our objective is to find the webpages as a widespread or popular webpages by using the Hit counts and the number of advertisements in that particular webpage.

Index Terms: World Wide Web, Web Mining, Patterns, Hits, Nodes, Hyperlink, Document Object Model.

1. INTRODUCTION

The web structure can be retrieved as nodes and edges (webpages are treated like nodes and hyperlinks are treated like edges). It will be easy to identify the popularity of webpages. Structure information can be discovered from the web by the process called as Web structure [1]. Web mining is more efficient when compared with the traditional data mining task. We have to consider the 3 key dissimilarities between the data mining and web mining. They are Access, Scaling and Structure [2].

ACCESS

To access the corporate information for data mining we are in need of access rights. The same is applicable to the web mining, but it is a rare case. In web structure mining it is not necessary to obtain the access rights to access the corporate information. But it needs the implicit agreement with web masters.

SCALE

In Web data mining it is easy for us to process more than 10 million pages, but in the case of outmoded data mining it is a hectic job to process 1 million of records. Web data mining will provide the good scaling factor for even very large number of records.

STRUCTURE

The outmoded data mining task will get data from the database. The database will provide explicit structure up to some levels.

- *Sathish Kumar G , Ramya R , Dinesh P S , Prabha Devi D currently working as Assistant Professor in the department of Computer Science and Engineering in Bannari Amman Institute of Technology, Erode, Tamil Nadu, India . E-mail: saathhish@gmail.com, rvr.ramya@gmail.com, dineshpudhu@gmail.com, dprabhadevi@gmail.com*
- *Janani A P is currently working as Associate Professor in the department of Information Technology in Dr. Mahalingam College of Engineering and Technology, Coimbatore, Tamil Nadu, India . E-mail: janani97@gmail.com*

The task of web mining is to process the semi structured data or unstructured data from the web pages.The presence of strategic analysis sector in a company will mine the client archives with the software related to data mining to determine the offers that can be given for the clients who has the maximum conversions [3].By this it is easy for us to determine that the organization is streaming towards the correct path in their marketing. This pays a way for the organization to make sure that their money is not spent unnecessarily [4]. Fraudulent Payments can be tracked by the companies with the help of data mining. The detailed research and study of data mining will help to achieve this [5]. The Buying fashion of the customers can be shown by web data mining and it will help us to make the projection on our investors [6]. The way we perform this analysis will permit the industry to load their stocks correctly for every month based on the predictions they laid by the analytics of buying trends[7].The large amount of data in the internet leads to the process of data mining and many new techniques are available to mine the data. The web data mining is raising the issues related to data security. Web data mining helps to keep the personal information in a secured way [8].

2 PROPOSED METHODOLOGY

In the proposed methodology the webpages are considered as nodes and the hyperlinks are considered as edges. The hyperlinks (edges) will connect a webpage to the related other webpages. The web structure mining will use the Graph theory for analyzing the connection and node design of the webpage. The Hits on a webpage is the major concern to predict the ranking of the webpage. The most liked and the famous webpages will get the more hits than the others. Each module will be extracted from the website by breaking the website structure. By default, web address is assigned for every module in the website. We have to save the web address of all the modules in the database. The update function in the database query is used to get the update in module address. The module count will be incremented when the user enters into the specific module. The page rank algorithm is used to deploy the count of all the webpages. It will calculate the uppermost number hits in a module to the lowermost hits in a module. With this algorithm it is easy for us to get the highest and lowest views for the webpage. With this the advertisement coordinator will display the advertisement in highest hit module. This will improve the efficiency in the field of advertisement.

HYPERLINKS

The Hyperlinks allows the user to flow easily between the pages. Intra-document hyperlink will connect the different parts related to the same web page and Inter-document hyperlinks used connect the two different webpages. The hyper link analysis [9] has the major role in ranking the web pages.

HITS

The files requested by the user will be equal to the hit numbers in a webpage. The hits in the webpage determines the request to the webserver for getting a file. The webpages can be downloaded from the webserver when the server gets the hit to that particular file. But one webpage load is not equal to the single hit, it will lead to false measure of a website popularity. The correct measure of particular web traffic will be measured by the number of views for that website.

DOCUMENT STRUCTURE

Based on the XML and HTML tags present in the webpage, the webpage content can be organized as a tree structured format. From the documents the document object model (DOM) structures can be extracted with the help of mining algorithms [10].

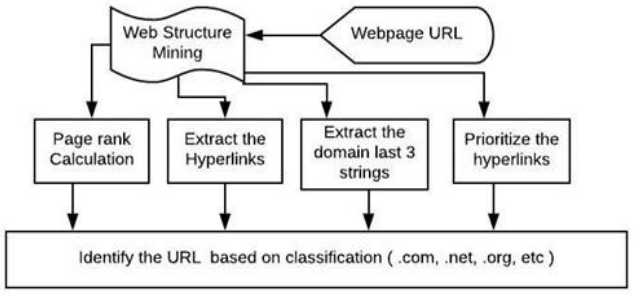

***FIGURE 1**: TREE STRUCTURE OF WEBPAGE CONTENT*

2.1 IMPLEMENTATION

2.1.1 GOOGLE PAGE RANK

The "Vote" for a website relates to the interesting factor of that website and the more interesting website reserves the highest votes which are recommended the most. All the website has its own starting score, which are calculated incremental [11].

- Few links in a webpage → high probability for choosing specific link.
- Many links in a webpage → low probability for choosing specific link.

The Page Rank can be calculated as follows,

$$PR(A) = (1-d) + d\,(PR(T1)/C(T1) + ... + PR(Tn)/C(Tn))$$

The page rank always forms a probability distribution on the webpage. So it will yield the page rank as One by summing all the webpages.

The implementation of web content extraction is been done by the following java code.

```
import java.io.*;
import java.net.URL;
import java.util.*;
import javax.swing.JOptionPane;
import org.apache.tika.exception.TikaException;
import org.apache.tika.io.TikaInputStream;
import org.apache.tika.metadata.Metadata;
import org.apache.tika.parser.AutoDetectParser;
import org.apache.tika.sax.Link;
import org.apache.tika.sax.LinkContentHandler;
import org.xml.sax.SAXException;
public class linkExtract{
 public static void main(String[] args){
  try {
   String addr=JOptionPane.showInputDialog("Web page
address:");
   if(addr!=null)

extractLinks(addr,System.getProperty("user.dir")+"/links.txt");
  }
 catch (Exception e)
{
e.printStackTrace();
  }  }
  public static void linkExtract (String source, String desti)
throws IOException, SAXException, TikaException{
BufferedWriter b=null;
   try{
InputStream isp=TikaInputStream.get(new
URL(src).openStream());
b=new BufferedWriter(new FileWriter(desc));
     Metadata m=new Metadata();
LinkContentHandlerlinkhandler=new LinkContentHandler();
AutoDetectParser pr=new AutoDetectParser();
pr.parse(isp, linkhandler, m);
   List<Link> links=linkhandler.getLinks();
   Iterator<Link> i=links.iterator();
   while(i.hasNext()){
b.write(i.next().toString());
b.newLine();
   }  }
   finally {
b.flush();   b.close();
  } } }
```

TABLE 1

EXAMPLES OF INTERNET DOMAINS CLASSIFICATION

Domain Name	Context
.com	commercial business.
.int	International sites.
.edu	educational institutions.
.gov	Government sites.
.mil	Military sites.
.biz	business sites.
.mobi	for sites like delivering services to mobile devices.
.org	organisational domain.
.net	network organizations.

2.1.2 IMPLEMENTATION OUTCOMES

The analysis of web structure content is successful for a data warehouse from the trustworthy web resource http://www.bannari.com/. The Page rank is calculated and the same is obtained from the Google ranking structure. The page rank is 0.3 for the web resource.

FIGURE 2: *PAGE RANK*

If we execute the above given Java code, it will extract the links connected to specified URL of Bannari Amman Group. It results in,

http://www.bannari.com/exports.html
http://www.bannari.com/granite.html
http://www.bannari.com/power.html
http://www.bannari.com/sugar.html
http://www.bannari.com/Education.html
http://www.bannari.com/seithiithazh.html
http://www.bannari.com/transports.html
http://www.bannari.com/alcohol.html
http://www.bannari.com/whatsnew.html
http://www.bannari.com/InvestorInformation.html
http://www.bannari.com/feedback.html
http://www.bannari.com/contactus.html
http://www.bannari.com/jobs.html

3 CONCLUSION

Web structure mining is used to abstract the information from the ancient behavior of the customers. The page ranking algorithm is used to ranks the appropriate pages by treating the available links equally while distributing the score of that webpage. In this paper, the business domain is identified to find the sub URL related to that website and to find the page rank. The java coding will extract the Sub URL related to the business domain. This page rank will decide the popularity of the webpage and it will help to improve the business accordingly. The Hit count for a particular web page is used to find the popularity of a webpage and it will decide the number of advertisement popups in that webpage.

REFERENCES

[1] Dr. S. P. Victor, Mr. M. Xavier Rex , " Analytical Implementation of Web Structure Mining Using Data Analysis in Educational Domain", International Journal of Applied Engineering Research ISSN 0973-4562 Volume 11, Number 4, pp. 2552-2556, 2016.

[2] Baraglia, R. Silvestri, F. "Dynamic personalization of web sites without user intervention", In Communication of the ACM 50(2): 63-67,2007.

[3] Cooley, R. Mobasher, B. and Srivastave, J. "Web Mining: Information and Pattern Discovery on the World Wide Web" In Proceedings of the 9th IEEE International Conference on Tool with Artificial Intelligence,1997.

[4] Cooley, R., Mobasher, B. and Srivastava, J. "Data Preparation for Mining World Wide Web Browsing Patterns", Journal of Knowledge and Information System, Vol. 1, Issue. 1, pp. 5–32, 1999.

[5] Costa, RP and Seco, N. "Hyponymy Extraction and Web Search Behavior Analysis Based On Query Reformulation", 11th Ibero-American Conference on Artificial Intelligence, 2008.

[6] Kohavi, R., Mason, L. and Zheng, Z. "Lessons and Challenges from Mining Retail E- commerce Data" Machine Learning, Vol 57, pp. 83– 113, 2004.

[7] Lillian Clark, I-Hsien Ting, Chris Kimble, Peter Wright, Daniel Kudenko"Combining ethnographic and clickstream data to identify user Web browsing strategies" Journal of Information Research, Vol. 11 No. 2, 2006.

[8] Eirinaki, M., Vazirgiannis, M. "Web Mining for Web Personalization", ACM Transactions on Internet Technology, Vol. 3, No. 1, 2003.

[9] Mobasher, B., Cooley, R. and Srivastava, J. "Automatic Personalization based on web usage Mining" Communications of the ACM, Vol. 43, No. 8, pp. 142–151, 2000.

[10] Mobasher, B., Dai, H., Luo, T. and Nakagawa, M. "Effective Personalization Based on Association Rule Discover from Web Usage Data" In Proceedings of WIDM 2001, Atlanta, GA, USA, pp. 9–15, 2001.

[11] Nasraoui O., PetenesC.,"Combining Web Usage Mining and Fuzzy Inference for Website Personalization", in Proc. of WebKDD 2003 – KDD Workshop on Web mining as a Premise to Effective and Intelligent Web Applications, Washington DC, p. 37,2003.