International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 6 Issue 06, June - 2017

A Study on Improve Quality of Data for Web Mining Using Data Cleaning Tools

Chirag R. Prajapati
Assistant Professor,
Department of computer science and application,
UCCC & SPBCBA & SDHG COLLEGE OF BCA AND IT, Udhna, Surat, India

Pinal P. Solanki
Assistant Professor,
Department of computer science and application,
V. T. Poddar BCA College , Surat, India

***Abstract*— Web mining is the one important application area of data mining which deals with the extraction of the useful data or facts from the World Wide Web. For web mining, it is required to fetch the quality data from the World Wide Web which helpful from collecting log information from the different locations like web server, client browser, etc. In this paper we have represented how to improve quality of data while retrieving information from the log files using different data cleaning tools, its phases and process.**

Keywords*— Web mining, Web usage mining, Data Pre-processing, Data Cleaning

I. INTRODUCTION

Web mining is mainly used to extract the data or knowledge from web data, web documents available on the web, and links between document usages of different web sites [3]. The web is the collection of so many documents and it is very dynamic, flexible and diverse. In today's world the World Wide Web continuously growing up in many areas which may difficult for identifying relevant data or information present in the web. The web sites are mostly handled three kind of information i.e. Content, Structure and Log Data, Based on these information web mining divided into three parts. First is Web Content Mining, second is Web Structure Mining and third is Web Usage Mining [2]. Web structure mining tries to discover the link structure of the hyperlinks at the inter document level based on the topology of the hyperlinks and generate the information, such as the similarity and relationship between different web sites [4]. Here we are mainly focus on the third one that is log data means Web Usage Mining.

In this paper we have included different sections. In section II we have described about web usage mining. In section III we have described about Data cleaning as a part of Data pre-processing and its different tools which helps to improve quality of data. In section IV we have described Literature Review. In section V we have described Conclusion.

II. WEB USAGE MINING

Web Usage Mining mines the secondary data which is present in log files and derived from the interactions of the users with the web. Web usage Mining techniques are applied on the data present in web server logs, browser logs, cookies, user profiles, bookmarks, mouse clicks etc. [5]. In web area World Wide Web has two sides one is a user side and another one is an information provider. Both sides are face problems while dealing with the web data. So Web Usage mining is useful to retrieves useful data from these [3]. Web Usage Mining includes the following three stages for retrieving data:

1. Data Pre-processing
2. Pattern Discovery
3. Pattern Analysis

Data Pre-processing also described the process of converting the usage, content and structure information into data abstraction. Then after the next phase that is Pattern discovery which focuses on to uncover patterns from the abstractions produced as a result of pre-processing phase. At the last Pattern analysis which separates the interesting and uninteresting patterns from the overall patterns discovered during pattern discovery phase [5]. As per the [2] while mining the data, pre-processing takes the 70% of time.

Data Pre-processing:

The quality data can only be produced by cleaning the data and pre-processing is prior to loading it in the data warehouse. The problems related to the data quality are presents in the single data collections. The single data collection includes the data collections sources like files and databases.

It includes the quality problems like misspellings of data during data entry, missing data or other invalid data. It is also known as data scrubbing, which deals with detecting and removing all the errors and inconsistencies from data in order to improve the quality of data [1].

Data Pre-processing is performing various pre-processing techniques on the log files to improve efficiency and quality of patterns mined and to avoid noisy and dirty data like:

1. Data Cleaning
2. User Identification
3. Session Identification
4. Path completion

Here the data cleaning is the major part of the data pre-processing.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 6 Issue 06, June - 2017

III. DATA CLEANING

Data cleaning is the one important process of data pre-processing for maintaining quality data by identifying incorrect or invalid or may be duplicate entries in the information system management [1]. It is also known as Data cleansing and Data scrubbing. Data cleaning is one of the major techniques used in the Data Pre-processing and Web Usage Mining [6]. We can improve the data quality for the mining by applying data cleaning software and tools.

There are different tools and software available for performing data cleaning which helps to improve web data quality. One of the popular data cleaning software is the R platform [1] and another one important thing is a EPLogCleaner [6].

The R platform is the one Data cleaning software which facilities for data manipulation, mathematical calculation and graphical representation. The R statistical environment provides a good platform for re-correctness data cleaning since all cleaning actions can be scripted on the data and therefore reproduced [1].

EPLog Cleaner is the one which is used to improve the data quality and efficiency of Web Log by filtering some inconsistent, irrelevant data based on the common URLs [6].

When we are working with heterogeneous data sources at that time Data Cleaning is required and it should be addressed together with schema-related data transformations [7]. The data cleaning algorithms can increase the quality and also it reduces the computational cost after finding and removing the outliers. If we talk about the several phases of data cleaning then it includes Data Analysis, Definition of transformation workflow and mapping rules, verification, Transformation, Backflow of cleaned Data [8].

Data Cleaning Tools:

The World Wide Web contains an enormous amount of data. The users of WWW need to use intelligent tools to find, sort, and filter the available data. The Web mining goals is to find and retrieve the information on the Web. By using algorithms we can find the work on web data. The user needs some Data Cleaning tools to find the necessary data [4].

The Following table shows some data cleaning tools available in the market which are mainly used for mining purpose:

Tools Name	Description
Drake	Drake is the one popular data cleaning tool which is very simple to use. It is the tool that organizes the execution of command around data and its dependencies. All the steps of Data pre-processing are defined with their input and output, Drake tool automatically resolves their dependencies and calculates that which command is best for execute, in that order to execute the commands.
	It is somewhere similar to the GNU Make but it designed especially for data workflow management. Drake has HDFS support which allows working on multiple input and outputs.
OpenRefine	OpenRefine is the one which formerly known as Google Refine is a tool for working with the data to clean it, transform it from one format to another format. After cleaning and transforming in to another format, it is extending it with web services and external data.
DataWrangler	DataWrangler is an interactive tool which focuses on the data cleaning and transformation. DataWrangler tool spends less time on the formatting of data and more time on the analysing of the data means more focusing on improving data quality. This tool also allows transformation of big data or messy data into the data tables analysis tools.
DataCleaner	It is a data profiling engine which finds the missing values, patterns, character sets of data and other important characteristics of data. DataCleaner is consider as an engine for analysing and discovering the quality of the data. if we talk about the data quality then profiling is one essential activity which can achieve by DataCleaner.
Winpure Data Cleaning Tool	Winpure Data Cleaning Tool is a good data cleaning tool which focuses on the problems like duplicate data, bad or wrong entries in the database and incorrect information available in the database.

IV. LITERATURE REVIEW

The following section discusses the various works of several authors.

L.Ramesh et al. [1] in this research paper they focuses on the how to perform data cleaning using the IMDA (Identify the Missing Data Algorithm). It also gives the information about data cleaning software.

D. Jayalatchumy et al. [2] in this paper they focuses on the problems and issues related to the data pre-processing, pattern discovery and pattern analysis.

S.Vidya et al. [3] authors focus on the application of the Web mining. In this paper they are provided usage of different applications of web mining.

D. Uma Maheswari et al. [4] in this paper authors focus on the future trends of the web usage mining and also tells to use the different tools for performing data cleaning.

Kamika Chaudhary et al. [5] in this paper they are providing the detail description about the web mining tools and its different techniques.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 6 Issue 06, June - 2017

Ms Shashi Sahu et al. [6] in this paper they are providing information related to how to improve of data quality of log file. In this paper they gives the details about the EPLogCleaner which helpful for performing data cleaning.

Erhard Rahm et al. [7] in this research paper they are giving information about problems and current approaches about the data cleaning. In this paper they are also described about the data transformation.

R. Deepa et al. [8] the research paper focus on the several algorithms used for performing data cleaning on the larger system. In this paper they are also described the several phases of Data cleaning

CONCLUSION

Web Usage Mining is mainly focuses on fetching useful or quality data from the Log File of web. To performing web mining on the data we required to use quality data and data cleaning is the first stage to provide quality data which is included in the data-preprocessing. So there are several data cleaning tools and software available which improve the things helpful for the web mining.

REFERENCES

[1] L.Ramesh, N.Marudachalam," Data Cleaning Using Identify the Missing Data Algorithm (IMDA)", International Journal of Innovative Research in Science, Engineering and Technology, Vol. 4, Issue 7, July 2015

[2] D. Jayalatchumy, Dr. P.Thambidurai,"Web Mining Research Issues and Future Directions – A Survey", IOSR Journal of Computer Engineering (IOSR-JCE) Volume 14, Issue 3 (Sep. - Oct. 2013).

[3] S.Vidya, K.Banumathy, "Web Mining- Concepts and Application", (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 6 (4), 2015.

[4] D. Uma Maheswari, Dr. A. Marimuthu, "A Study of Web Usage Mining Applications and its Future Trends", International Journal of Engineering Research & Technology (IJERT) Vol. 2 Issue 9, September – 2013.

[5] Kamika Chaudhary, Santosh Kumar Gupta," Web Usage Mining Tools & Techniques: A Survey", International Journal of Scientific & Engineering Research, Volume 4, Issue 6, June-2013.

[6] Ms Shashi Sahu,Ass. Prof. Leena Sahu,"A Survey on Frequent Web Page Mining with Improving Data Quality of Log Cleaner", International Journal of Advanced Research in Computer Engineering & Technology (IJARCET) Volume 4 Issue 3, March 2015

[7] Erhard Rahm, Hong Hai Do, "Data cleaning: Problems and Current Approaches", University of Leipzig, Germany

[8] R. Deepa,Dr. R Manicka Chezian, "A Study on Data Cleansing and Classification Algorithms for Large Dataset Systems", International Journal of Research in Advent Technology, Vol.2, No.9, September 2014.