A New Approach for Handling Null Values in Web Server Log

Pradeep Ahirwar [1]
Department of CSE
MANIT, BHOPAL, INDIA
[1]pradeepgec@yahoo.com

Deepak Singh Tomar[2]
Department of CSE
MANIT, BHOPAL, INDIA
[2]deepaktomar@manit.ac.in

Rajesh Wadhvani[3]
Department of IT
MANIT, BHOPAL, INDIA
[3]wadhvani_rajesh@rediffmail.com

***Abstract*— The web log data embed much of the user's browsing behavior and the operational data generated through Internet end user interaction may contain noise. Which affect the knowledge based decision. Handling these noisy data is a major challenge. Null value handling is an important noise handling technique in relational data base system. In this work the issues related to null value are discussed and null value handling concept based on train data set is applied to real MANIT web server log. A prototype system based on Fuzzy C-means clustering techniques with trained data set is also proposed in this work. The proposed method integrates advantages of fuzzy system and also introduces a new criterion, which enhances the estimated accuracy of the approximation. The comparisons between different methods for handling null values are depicted. The result shows the effectiveness of the methods empirically on realistic web logs and explores the accuracy, coverage and performance of the proposed Models.**

Keywords- Null value, web mining, k-means clustering, fuzzy C-means clustering, log records, log parser.

I. INTRODUCTION

Web mining involves a wide range of applications that aims at discovering and extracting hidden information in data stored on the Web. It is also provide a mechanism to make the data access more efficiently and adequately and discover the information which can be derived from the activities of users, which are stored in log files, by the log records, investigator can vigil on the client's activities but most of the time connection fails due to which it is not possible store entire log record entry in the database, this creates many discrepancies for investigation of the client activities, to enhance the investigation one of the significant process is estimate the null values.

The mining process will be ineffective if the input data represent the raw content. Database systems have much more operational inefficiencies, therefore, for many applications in the area of data analysis, data pre-processing is an essential step. In addition, the handling of null values is the major task of the data quality. The poor data quality can cause negative results, including lost information. Inaccurate, inconsistent or the null data in the database can hamper research's ability to ascertain useful knowledge. An effective data quality strategy can help researchers to find knowledge in database, allowing them to make right decision and reduce costly operational inefficiencies. This paper included a trained data set method, which operated on web mining for processing relational database. Instead of directly taking the database we take live log record and ascertain the relational database we parse these log record by the assistance of Web log Analyser [7]. The structure of the proposed method can be composed of four phases, comprising log records and parsing these log records, applying clustering algorithm with trained data set method to form clusters, estimating null values through these clusters and comparing our possible techniques with other existing methods.

II. PRELIMINARIES

In present, one of the leading research areas in web mining is estimating null values using different techniques [1,2,3]. In this section, we describe briefly about the null value, web mining, k-means clustering, fuzzy c-means clustering, log records and log parser.

A. Null Values

A null value shows the dearth of information or the value is unknown. A value of null is apart from zero or empty. Operation among the null values or any other values, return null results and the value of each null is unknown or unavailable, so it is one of the tedious task to get the knowledge from such type of data record also the mining process will be ineffective if the samples are not a good representation of the data.

Database systems have much more operational discrepancies, therefore, for many applications in the area of data analysis, data pre-processing is an essential step. The availability of null value may be due to the data was not captured, due to faulty equipments, inconsistencies with other recorded data, the application program might have deleted the data, the data were not entered due to confusion, certain data may not have been considered important enough at the time of entry and the data was not registered in the database.

Sometimes null values represent significant and crucial data, and may need to be inferred and approximated. In recent years many researchers pay heed to the estimating null values in relational database systems. There are a number of approaches to deal with this problem; overlook objects containing null values, fill the gaps physically, substitute the missing values

with a continuous, use the mean of the objects and use the most probable value to fill in the missing values, but these methods are not completely sufficient to handle Null value problems [2]. Therefore it is one of the primary research areas in the analysis of the data.

B. Web Mining

Web mining is the branch of the data mining used for discovering the patterns from the web, due to the heterogeneous and devoid of structure of the web; data mining is a challenging task. Through the web server many kinds of data are generated, one of them is log records. Due to many security reasons it is necessary to analyze web log record but many times by the failure in the networks some field of these log records are unknown but this unknown information is very important in the field of investigation, so we propose some possible methods to estimate this unknown value in log records with high accuracy. With the huge amount of information available online the World Wide Web is the fertile area for data mining research. The web mining research is at the cross road of research from several research communities, such as database, information retrieval, and with in AI, specially the sub-areas of machine learning and natural language processing.

C. Log Records and Log Analyzer

Log files formed by web or proxy servers are text files with a row for each http transaction on the network. Web log files contain a large amount of incorrect, ambiguous, and deficient information. The web log gives information about the site's visitors, activity statistics, accessed files, paths through the site, information about referring pages, search engines, browsers, operating systems, and more. Here we describe a log entry.

```
61.2.55.81 - - [04/May/2009:22:53:32 -
0700] "GET
/templates/ja_pollux/ja_transmenu/ja-
transmenuh.css HTTP/1.1" 200 5413
"http://manit.ac.in/content/view/297/123/"
"Mozilla/5.0 (Windows; U; Windows NT 5.1;
en-US; rv: 1.9) Gecko/2008052906
Firefox/3.0"
```

Here “61.2.55.81” is the IP address of the client (the system which is initializing the request). “04/May/2009:22:53:32” is the date time of the transaction, GET is the method of the transaction (GET and POST are the methods which are used to interact with the server), “/templates/ja_pollux/ja_transmenu/ja-transmenuh.css” is the URL requested by the client, “HTTP/1.1” is the http protocol, “200” is the http return code it means ok, “5423” is the size in bytes of the response sent to the client, “http://manit.ac.in/content/view/297/123/” is the cookie at the client browser, “Mozilla/5.0 (Windows; U; Windows NT 5.1; en-US; rv:1.9) Gecko/2008052906 Firefox/3.0” is the client environment specification on provided by the client browser.

A colossal amount of transactions occur day by day on every server. This generates log records endlessly on the server, it is not a simple task to examine these log records by these simply reading text files because these log files are not in well formed so we need a log analyser or a log parser to parse these log records, this makes simple task to analyse these tabular or well structured form of the logs which is generated by the log parsers. The log analyser also supports creating click overlay reports. These interactive reports allow you to navigate your site and view links marked according to the frequency they were clicked, as well as get additional statistics for site pages.

D. Clustering Techniques

Clustering is the process of organizing data items into groups where members are possess posses some similarity among them. A cluster is therefore a collection of similar data items or a collection of dissimilar data items belonging to different clusters.

1) K-means Clustering: The K-means clustering is one of the initials clustering algorithms proposed, it is one of the easiest type of unsupervised learning techniques that solve the clustering problem. It allows the partition of the given data into the k-clusters, the number of clusters is previously decided, after that each cluster randomly guesses centre locations and each data item finds out which centre it is closest to, thus each centre owns a set of data items, now each centre finds its centroid and jumps there, this process is repeated until terminates.

Although it has the advantage that it is easy to implement, it has two drawbacks. First, it is really slow as in each step the distance between each point to each cluster has to be calculated due to this phenomenon it is expensive in a large dataset. Secondly, this method is very susceptible because we provide the initial value of the clusters.

2) Fuzzy C-Means Clustering: In Fuzzy C-Means a data is formed into c-clusters with every data value in the dataset belongs to all clusters with certain degree. It lies under the unsupervised method and is inherited from fuzzy logic; it is capable of solving the multiclass and ambiguous clustering problems. Fuzziness measures the degree to which an event occurs due to this we are able to increase the probability as respective to the normal probability calculation. In the traditional clustering we assign the each data item to only one cluster. In this clustering we assign different degrees of membership to each point. The membership of a particular data item is shared between various clusters. This creates the concept of fuzzy boundaries, which differs from the traditional concept of well-defined boundaries. The well-defined boundary model does not reflect the description of real datasets. This contention led a new field of clustering algorithms based on a fuzzy extension of the least-square error criterion.

III. Problem Description

Including unknown values within your database can have an undesirable effect when using this data within any calculative operations. Any operation that includes a unknown value will result is a null; this being logical as if a value is unknown then the result of the operation will also be unknown [6]. The result below shows how using a null in a Boolean will alter the outcome, this is also known as three-value logic.
The data below shows how using a null in a calculation will alter the outcome:

(40 x 3) + 10 = 130 (Null x 4) + 8 = Null
(7 x Null) + 4 = Null (8 x 9) + Null = Null

The problem of providing a formal treatment for incomplete database information and null values represents one of the thorniest research issues in database theory. A particular source of difficulties has been the generalization of the relational model to include null values. At present some methods exist to estimate null values [4] from relational database and data mining systems.

IV. Proposed Approach

As described in the introduction, the method is structured on a trained data set methodology, including fuzzy c-means clustering and relational database estimation. Assume that there is a null value in the log record; due to the presence of these null values it is not possible to investigate the client activities on the servers or the web, so we estimate these null values. The purpose of the possible proposed algorithm is to process the relational database estimation with a highly estimated accuracy rate by integrating the advantages of the fuzzy c-means clustering algorithm with the trained data set and relational database estimation simultaneously. In the phase of fuzzy c-means with trained data set we already trained some data point which gives surety that, they will always produce some optimal output than the core fuzzy c-means clustering algorithm. When we implement our proposed method and compare it with the previously implemented K-means and C-means algorithm, [3,5] we find that this method gives better results than others.

To show this procedure we take live web log records of Maulana Azad National Institute of Technology, Bhopal, India, for processing our possible approach, first we parse these log records by the tool WebLogExpert [7], in Fig-2(A), Fig-2(B) we show daily search phrases for the Maulana Azad National Institute of Technology, Bhopal, server, in Fig-3(A), Fig-3(B) we show some possible errors on the web log, these possible errors may create a null entry in the web log. For measuring performance, accuracy of the proposed method, we perform our operation on the 500 records of the MANIT, Bhopal, India log records. Finally Fig-1 shows that how we resolve this null value by our proposed possible approach.

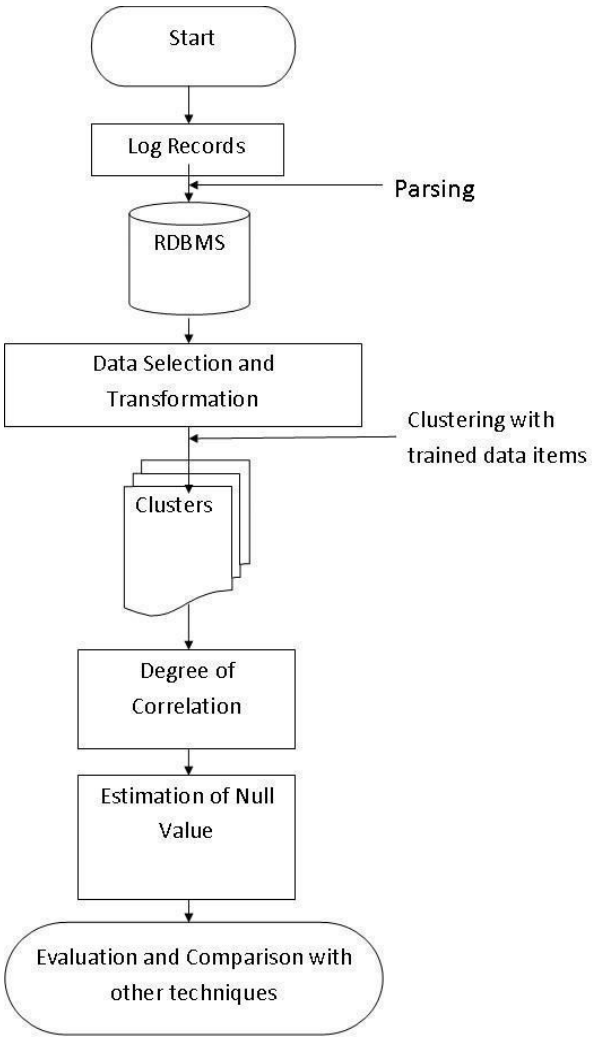

Fig. 1. Proposed possible approach for null value estimation.

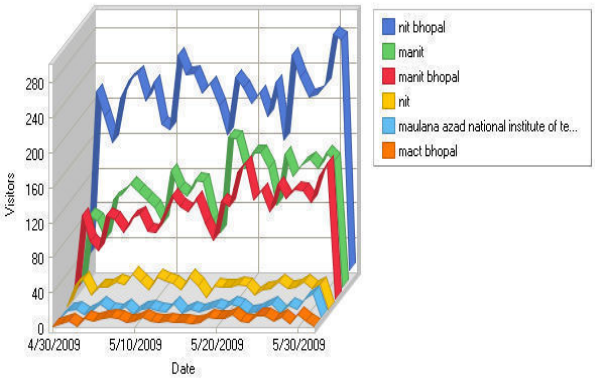

Fig. 2(A). Daily Search Phrases in MANIT Log Records.

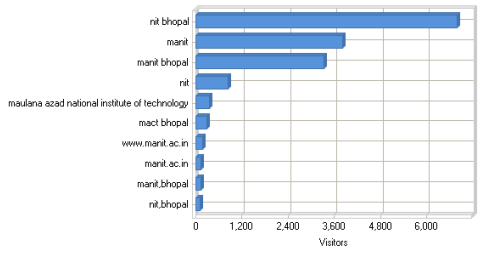

Fig. 2(B). Top Search Phrases in MANIT Log Records.

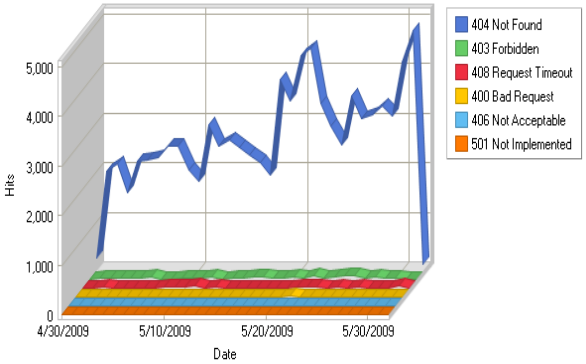

Fig. 3(A). Daily Error Types in MANIT Log Records.

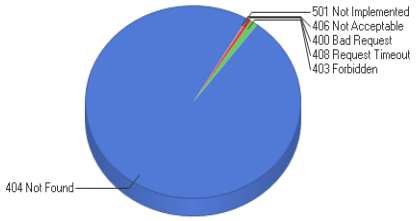

Fig. 3(B). Error Types in MANIT Log Records.

V. Results And Comparision

The results of proposed method compared with the other Implemented methods, the K-means Clustering (where the user need to enter the Cluster value and Seed value) another hands our proposed trained Data Clustering methodology where no need of giving the Seed and Cluster value instead of only generating point needed. In Table 5.1 clearly shown that our proposed method has better results than the implemented methods.

The resultant Iteration and Error rate are generated by our proposed architecture, which identifies most accurate results than the respective and previous implemented method. For measuring performance, accuracy of our proposed method, the Operation performed on 500 records of the MANIT, Bhopal, India log records.

Methods	No. of Iteration	Error Rate
Implemented Method (K-means Clustering)	10	4.93
	09	5.70
Proposed Method (Trained Data based Method)	Ini. Val .45 7	2.73
	.24 5	2.68
	.53 6	2.93

Table 5.1 Performance Comparisons for Implemented and Proposed Method.

VI. Conclusion

In this paper, the proposed method is based on trained data set to estimate null value in relational database systems by constructing fuzzy c-means clustering algorithm, and which integrates advantages of fuzzy system. Due to this trained data set, the proposed method can effectively achieve better performance on relational database estimation. Now we are able to get the unknown values present in the log records, which were the biggest hurdle in the field of analyzing the log record.

Acknowledgment

The work presented in this paper would not have been possible without our college, at MANIT, Bhopal. We wish to express our thankfulness to all the people who helped turn the World-Wide Web into the useful and popular distributed hypertext it is. We also wish to thank the anonymous reviewers for their valuable suggestions.

References

[1] Pradeep Ahirwar, Deepak Singh Tomar, Rajesh Wadhvani, "*Handling Null Values in Web Server Log*", National Conference on Recent Trends and Challenges in Internet Technology, RTCIT, 19-20 march 2010.

[2] Ching-Hsue Cheng, Liang-Ying Wei, Tzu-Cheng Lin, "*Improving Relational Database Quality Based on Adaptive Learning Method for Estimating Null Value*", IEEE Innovative Computing, Information and Control, 2007. ICICIC apos; 07. Second International Conference on Volume, Issue, 5-7 Sept. 2007.

[3] S.M. Chen, and H.R. Hsiao, "A new method to estimate null values in relational database systems based on automatic clustering techniques", IEEE Information Sciences: an International Journal, 169, 2005, pp. 47-60.

[4] Shin-Jye Lee, Xiaojun Zeng, *"A Modular Method for Estimating Null Values in Relational Database Systems"* Eighth International Conference on Intelligent Systems Design and Applications, Nov. 2008.

[5] Muhammad Nazrul Islam, Pintu Chandra Shill, Muhammad Firoz Mridha, Dewan Muhammad, Sariful Islam and M.M.A Hashem, *"Generating Weighted Fuzzy Rules for Estimating Null Values Using an Evolutionary Algorithm"* in 4th International Conference on Electrical and Computer Engineering, 19-21 Dec 2006.

[6] Claude Rubinson, "Nulls, three-valued logic, and ambiguity in SQL: critiquing date's critique," ACM SIGMOD Record, v.36 n.4, p.13-17, December 2007.

[7] WebLogExpert, http://www.weblogexpert.com.

AUTHORS PROFILE

[1]**Mr. Pradeep Ahirwar** is a student of MTech Computer Science, MANIT, Bhopal, MP, INDIA. His Research activities are based on web log mining, digital forensics.

[2]**Mr. Deepak Singh Tomar**, working as Assistant Professor in Department of Computer Science at MANIT Bhopal, MP, INDIA. His Research activities are based on digital forensics, data mining and network security.

[3]**Mr. Rajesh Wadhvani, working as** Assistant Professor in Department of Information Technology at MANIT Bhopal, MP, INDIA. His Research activities are based on database, data mining, and network security.