International Journal of Pure and Applied Mathematics
Volume 117 No. 20 2017, 785-794
ISSN: 1311-8080 (printed version); ISSN: 1314-3395 (on-line version)
url: http://www.ijpam.eu
Special Issue

ijpam.eu

A Survey on Data Preprocessing Techniques for Bioinformatics and Web Usage Mining

[1]A. Sivakumar and [2]R.Gunasundari

[1]Department of Computer Science,

Karpagam University,

Coimbatore.

sivamgp@gmail.com

[2]Department of Information Technology,

Karpagam University,

Coimbatore.

Abstract

It is well known that over 80% of time is required to carry out any real world data mining project is usually spent on data preprocessing. Data Preprocessing lays the groundwork for data mining. Before the discovery of useful information/knowledge, the target dataset must be properly prepared. But it is unfortunately ignored by the most researchers on data mining due to its perceived difficulty. This paper describes an efficient approach for data preprocessing for mining based bioinformatics and web usage mining data in order to speed up the data preparation process. This paper surveys the data preprocessing activities like data cleaning, data reduction and related algorithms. It is not only providing flexibility for data preprocessing, but also reduces complexity and difficulty in preparing mining data.

Key Words: Data mining, data preprocessing, data cleaning, bioinformatics, web usage mining.

1. Introduction

Data Preprocessing is required and it is an important phase in Bioinformatics and Web Usage Mining. Data Cleaning and User Identification are the methods in Data Preprocessing. The purpose of data cleaning is to eliminate the irrelevant items. This current research is persisting with data preprocessing methods which include data cleaning, data integration, data transformation and data reduction. Different techniques are provided for data cleaning but there are some problems in data collection and accurate metric of user identification. This paper provides the review on algorithm and different techniques are used in Data Preprocessing which is in turn used for Bioinformatics and Web Usage Mining.

Data Preprocessing [1] is used to clean the data, when it provides the pattern to discover, it identifies the technique which will further be used to discover the user's navigational pattern. After the processing, it passes to pattern analysis which takes only relevant pattern and removes irrelevant pattern. Data Mining is the data-driven technique [1] to discover patterns in large volumes of raw data. Bioinformatics Mining is performed in three steps – Data Preprocessing, Pattern Discovery and Pattern Analysis. The results of the pattern discovery directly influence the quality of the data processing. Good data sources discover the quality patterns and also improve the bioinformatics algorithm.

Hence, Data Preprocessing is a main activity to complete Bioinformatics Mining processes and plays vital role in determining the quality of patterns. In Data Preprocessing, the collection of data differs not only in the type of data available but also the data source site, the data source size and the way it is being implemented. The Data Preprocessing of Bioinformatics Mining is usually complex.

The purpose of Data Preprocessing is to offer reliable, structural and integrated data source to pattern discovery. Pattern discovery is the key process of the Bioinformatics Mining, which covers the algorithms and techniques from several research areas, such as data mining, machine learning, statistics and pattern recognition. The techniques such as statistical analysis, association rules, clustering, classification, sequential pattern and dependency modelling are used to discover rules and patterns. The knowledge is discovered that can be represented in the form of rules, tables, charts, graphs, and other visual presentation forms for characterizing, comparing, predicting, or classifying data from the bioinformatics 1og.

The final stage of this Bioinformatics Mining is pattern analysis. The aim of this process is to extract the rules or patterns from the output of pattern discovery by eliminating the irrelative rules or patterns. Here, the focus is on data preprocessing method of Bioinformatics.

2. Necessity for Bioinformatics Data Set Preprocessing

The Bioinformatics datasets collected from the field, are very raw and have the tendency of following characteristics [2]. This data has to be processed before analyzing through Data Mining Techniques.

Incomplete

When collecting the data of any domain or bioinformatics data from the field, there is a possibility of lacking in attribute values or certain attributes of interest, or containing only aggregate data. Missing data, particularly for tuples with missing values for some attributes, may need to be inferred.

Noisy

Noisy data means, the data in the tuples containing errors, or outlier values that deviate from the expected. Incorrect data may also produce the result from inconsistency in naming conventions or data codes used, or inconsistent formats for input fields, such as date. Hence it is necessary to use some techniques to replace the noisy data.

Inconsistent

Inconsistent means, the data source containing discrepancy between different data items. Some attributes representing the given concept may have different names in different databases, causing inconsistency and redundancy. Naming inconsistency may also occur in attribute values. Therefore the inconsistency in data needs to be removed.

Aggregate Information

It would be useful to obtain aggregate information such as the bioinformatics data sets something that is not a part of any pre-computed data cube in the data warehouse.

Enhancing Mining Process

Large number of data sets may make the data mining process slow. Hence, reducing the number of data sets to enhance the performance of the mining process is important.

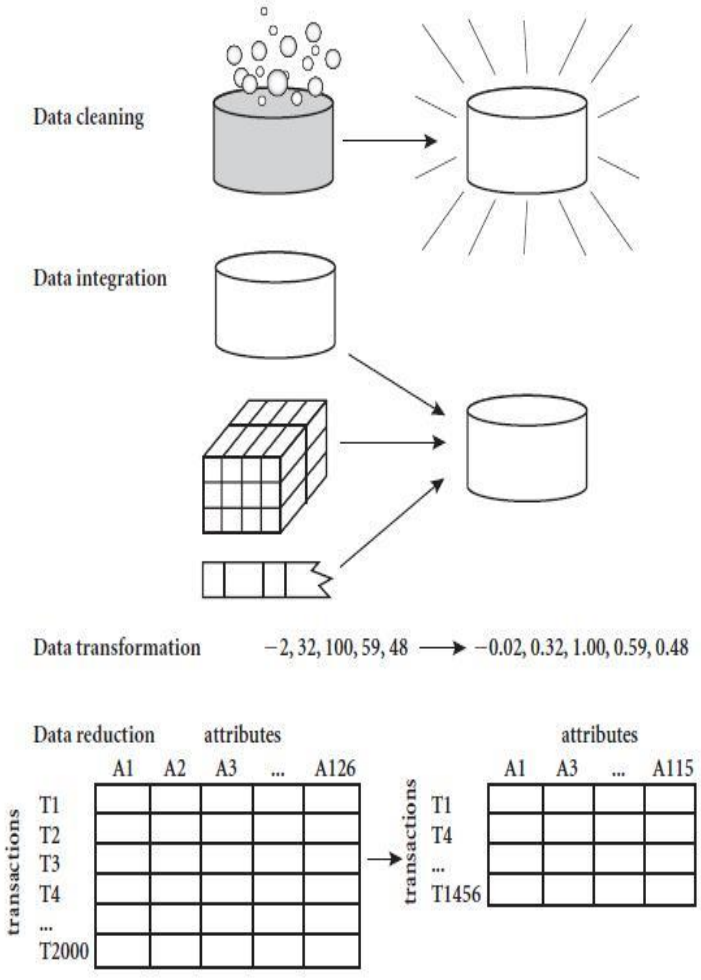
Improve Data Quality

Data Preprocessing techniques can improve the quality of the data, thereby help to improve the accuracy and efficiency of the subsequent mining process. Data Pre-processing is an important step in the knowledge discovery process, because quality decisions is based on the quality data. The detecting data become anomalies and rectifying them can lead to improve the accuracy and efficiency of the data analysis.

3. Data Preprocessing Methods

Raw data is highly susceptible to noise, missing values, and inconsistency. The quality of data affects the data mining results. In order to improve the quality of the data and the mining results, raw data is pre-processed so as to improve the efficiency and ease the mining process. Data Preprocessing is one of the most critical steps in data mining process which deals with the preparation and transformation of the initial dataset. Data preprocessing methods are divided into the following categories. They are:

1. Data Cleaning.
2. Data Integration.
3. Data Transformation.
4. Data Reduction.



Data Cleaning

Data is analyzed by data mining techniques which may incomplete, noisy, and inconsistent. Real-world data tend to be incomplete, noisy, and inconsistent.

Data cleaning (or data cleansing) routines attempt to fill in missing values, smooth out noise while identifying outliers, and correct inconsistencies in the data. This section, gives basic methods for data cleaning. Incomplete, noisy, and inconsistent data are the common properties of real-world databases and data warehouses.

Data can be noisy, having incorrect attribute values. Owing to the following, the data collection instruments used may be fault. There may be human or computer errors occurred at data entry. Errors in data transmission can also occur. There may be technological limitations, such as limited buffer size for coordinating synchronized data transfer and consumption. Incorrect data may also produce the result from inconsistency in naming conventions or data codes.

Data cleaning routines work to "clean" the data by filling in missing values, smoothing noisy data, identifying or removing outliers, and resolving inconsistency. Dirty data can cause confusion for the mining procedure. Although most mining routines have some procedures, they deal incomplete or noisy data, which are not always robust. Therefore, a useful pre-processing step is to run the data through some data cleaning routines.

Data Integration

It is likely that the data analysis task involves in data integration, which combines the data from multiple sources into a coherent data store, as in data warehousing. These sources may include multiple databases, data cubes, or flat files. There are number of issues to consider during data integration. Schema integration is tricky. How can real world entities from multiple data sources be 'matched up'? This is referred to as the entity identification problem. For example, how can the data analyst or the computer be sure that customer id in one database, and cust_number in another refer to the same entity? databases and data warehouses typically have metadata - that is, data about the data. Such metadata is used to help avoid errors in schema integration. Redundancy is an another important issue. An attribute may be redundant, if it is derived from another table, such as annual revenue. Inconsistencies in attribute or dimension naming can also cause redundancies in the resulting data set.

Data Transformation

In data transformation, the data are transformed or consolidated into appropriate forms for mining. Data transformation involves the following:

1. In Normalization, where the attribute data are scaled so as to fall within a small specified range, such as -1.0 to 1.0, or 0 to 1.0.
2. Smoothing works remove the noise from the data. Such techniques include binning, clustering, and regression.
3. In Aggregation, summary or aggregation operations are applied to the data. For example, the daily sales data may be aggregated so as to compute monthly and annual total amounts. This step is typically used

in constructing a data cube for analysis of the data at multiple granularities.

4. In Generalization of the data, low level or 'primitive' (raw) data are replaced by higher level concepts through the use of concept hierarchies. For example, categorical attributes are generalized to higher level concepts street into city or county. Similarly, the values for numeric attributes may be mapped to higher level concepts like, age into young, middle-aged, or senior.

Data Reduction

Complex data analysis and mining on huge amounts of data may take a very long time, making such analysis impractical or infeasible. Data reduction techniques is helpful in analyzing the reduced representation of the dataset without compromising the integrity of the original data and yet producing the qualitative knowledge. The concept of data reduction is commonly understood as either reducing the volume or reducing the dimensions (number of attributes). There are number of methods that facilitate in analyzing the reduced volume or dimension of data and yield useful knowledge. Certain partition based methods work on partition of data tuples. That is, mining on the reduced data set should be more efficient yet produce the same (or almost the same) analytical results. Strategies for data reduction include the following,

1. In Data cube aggregation, aggregation operations are applied to the data in the construction of a data cube.
2. In Dimension reduction, irrelevant, weakly relevant or redundant attributes or dimensions may be detected and removed.
3. Data compression - encoding mechanisms are used to reduce the data set size. The methods are used for data compression are wavelet transform and principle component analysis.
4. Numerosity reduction - the data is replaced or estimated by alternative, smaller data representations such as parametric models (which store only the model parameters instead of the actual data e.g. regression and log-linear models), or nonparametric methods such as clustering, sampling, and the use of histograms.
5. Discretization and concept hierarchy generation - raw data values for attributes are replaced by ranges or higher conceptual levels. Concept hierarchies allow the mining of data at multiple levels of abstraction, and are powerful tools for data mining.

4. Related Work

Navin Kumar Tyagi [4], some data preprocessing activities like data cleaning and data reduction is surveyed and presented the algorithms for data cleaning and data reduction. It is important to note that before applying data mining techniques to discover user access patterns from web log, the data must be processed because quality of results is based on the data is mined. This paper

presents two algorithms for preprocessing, ie, algorithm for data cleaning and data reduction are proposed in data cleaning algorithm, the records with extension .jpg, gif, .css are removed but records with irrelevant status code are not removed in this algorithm so, the status code can be removed in improved algorithm in data reduction algorithm identifies the sessions and removes the incomplete session entries are removed.

According to Ke Yiping [5], surveyed a typical sequence of tasks of web usage mining preprocessing. The techniques and methods are used in each individual task which are also presented in great details. One important thing to point out is that each task relies heavily on each other. In practical application, some of the tasks are carried out together and do not distinguish with each other, clearly. Moreover, for specific mining applications, the procedures of preprocessing may be in little variation. As for the application of web personalization, the preprocessing steps include data selection, cleaning, transformation and the identification of users and user sessions.

Huaqiang Zhou, Hongxia Gao and Han Xiao [6] focussed on data preprocessing methods such as Data Cleaning, User Identification, Path Completion, Session Identification and Transaction Identification. In transaction identification, data in user session may be too large and it is needed to convert into smaller transaction by using segmentation algorithm to identify. In analysis of preprocessing method, the Frame Page Filter and Time Out Threshold Value Setting are used.

Theint Aye [7], proposed a technique for preprocessing ie., Field Extraction and Data Cleaning. Main task is cleaning the web log file and inserting the processed data into a relational database so that the data mining technique can be applied on it. Field Extraction is the process of separating field from the single line of log file. Field Extraction algorithm is used to open database connection and create a table to store log data file.

Data Cleaning algorithm is used to eliminate irrelevant or unnecessary items in the analyzed data. Web log file also records the failed HTTP status codes and suffix. So data cleaning inconsistency is detected and removes to improve the quality of data. Here in algorithm, for data cleaning log table as an input which is used to generate after the field extraction.

In the opinion of Harmit kaur and Hardeep singh [3], data preprocessing in log files is explained. The preprocessing from data fusion and cleaning is started. The data are combined from different sources and then irrelevant entries are also removed. In Session identification, time oriented and structure oriented are suggested for using graphs.

According to Dafa-Alla, Mirghani. A. Eltahir and Anour [8], analysed that techniques of preprocessing are used in data cleaning, data filtering, path

completion, user identification, session identification and web session clustering. The different sources of log files, log file formats, preprocessing techniques, algorithms applied and data support to data preprocessing phase are described. A survey is done by the authors on preprocessing techniques used in preprocessing phase.

In the opinion of Tasawar hussain, Dr.asghar and Dr.masood [9], Web log data preprocessing is divided into steps such as log consolidation, data cleaning, user identification and transaction identification. Log consolidation is the first step in preprocessing in which the logs from different servers are combined into one place for data cleaning. Next step is data cleaning which is divided into two parts ie., page element cleaning in which files with extension.gif, jpeg, .jpg are removed and cleaning other information such as files with extension .css, xsl, .xsd, .dll.

Data preprocessing is used to clean the data so that when it provides to the independent component analysis, it identifies the technique which is used to discover the pattern after the processing, it passes that to pattern analysis so that it takes relevant pattern, removing irrelevant pattern.

5. Conclusion

Data preparation is an important issue for both Bioinformatics and Web Usage Mining, as real-world data tend to be incomplete, noisy, and inconsistent. Data preparation includes data cleaning, data integration, data transformation, and data reduction. Preprocessing improves the performance of bioinformatics and web mining data. Data cleaning routines can be used to fill in missing values, smooth noisy data, identifying outliers, and the correct data inconsistencies. Data integration combines data from multiple sources to form a coherent data store. Metadata, correlation analysis, data conflict detection, and the resolution of semantic heterogeneity contribute towards smooth data integration. Data transformation routines confirm the data into appropriate forms for mining. Data reduction techniques such as data cube aggregation, dimension reduction, data compression, numerous reduction, and discretization can be used to obtain a reduced representation of the data, while minimizing the loss of informative content. Although several methods of data preparation are developed, data preparation remains an active and important area of research. To sum up shortly, the concepts of data mining and improving the performance of bioinformatics and web mining data by using preprocessing techniques is analysed and presented.

References

[1] Chandrama W., Devale P.R., Ravindra M., Survey on Data Preprocessing Method of Web Usage Mining, International

Journal of Computer Science and Information Technologies 5 (3) (2014), 3521-3524.

[2] Baskar S.S., Arockiam L., Charles S., A Systematic Approach on Data Pre-processing in Data Mining, An International Journal of Advanced Computer Technology 2 (11) (2013).

[3] Harmit K., Hardeep S., A Survey of Preprocessing Method for Web Usage Mining Process, International Journal of Computer Trends and Technology (IJCTT) 9 (2) (2014).

[4] Navin Kumar T., Solanki A.K., Sanjay T., An Algorithmic Approach to Data Preprocessing in Web Usage Mining, International Journal of Information Technology and Knowledge Management 2 (2010).

[5] Ke Y., A Survey on Preprocessing Techniques in Web Usage Mining, Computer Science Department, The Hong Kong University of Science and Technology, (2003).

[6] Zhou H., Gao H., Xiao H., Research on improving methods of preprocessing in web log mining, IEEE 2nd International Conference on Information Science and Engineering (ICISE) (2010), 1472-1474.

[7] Aye, T.T., Web log cleaning for mining of web usage patterns, IEEE 3rd International Conference on Computer Research and Development (ICCRD) 2 (2011), 490-494.

[8] Eltahir M.A., Dafa-Alla A.F., Extracting knowledge from web server logs using web usage mining, IEEE International Conference on Computing, Electrical and Electronics Engineering (ICCEEE) (2013), 413-417.

[9] Sheetal A.R., Shailendra J., Efficient Preprocessing technique using Web log mining, International Journal of Advancements in Research & Technology 1 (6) (2012), 1-5.

[10] Chitraa V., Antony S.D., A Survey on Preprocessing Methods for Web Usage Data, International Journal of Computer Science and Information Security 7 (3) (2010).

[11] Renso C., Introduction to Data mining: Data Preprocessing, UFPE, (2012).

[12] Jun D., Data Preprocessing, The University of Western Ontario (2013), 1-48.

[13] Vu A.T., Osamu H., A Data Preprocessing Algorithm to Improve the Performance of Clustering, Journal of Software Engineering and Applications 7 (2014), 639-654.