IJARSE
ISSN (O) 2319 - 8354
ISSN (P) 2319 - 8346

Web Mining Techniques and Preprocessing

[1]Dharmendra Kaushik,[2]Gargishankar Verma,[3]K.Ravikant

Department of CSE ,CIET Raipur ,CSVTU Bhilai ,(India)

ABSTRACT

The spread and involvement of Internet has grown amazingly in all walks of life in last 10 to 20 years. The web has become a prime source of information in every field of work. Influencing and interacting with all kinds of surfers, from children to old men, from housewives to industrialists. This pervasiveness of Internet in such rapid pace has been possible due to availability of information in any subject at any expert level to any person of any expertise level. Its usefulness is obvious and un-doubtful.To retrieve all type of information various techniques are available as 'web pages' for that preprocessing is required.

Keywords: Web mining,preprocessing,merging,formatting,pattern,classification,hyperlink.

I. INTRODUCTION

According to William J. Frawley, Gregory Piatetsky-Shapiro and Christopher J. Matheus 'Data Mining, or Knowledge Discovery in Databases (KDD) as it is also known, is the nontrivial extraction of implicit, previously unknown, and potentially useful information from data. This encompasses a number of different technical approaches, such as clustering, data summarization, learning classification rules, finding dependency networks, analyzing changes, and detecting anomalies'.According to Marcel Holshemier and Arno Siebes "Data mining is the search for relationships and global patterns that exist in large databases but are 'hidden' among the vast amount of data, such as a relationship between patient data and their medical diagnosis. These relationships represent valuable knowledge about the database and the objects in the database and, if the database is a faithful mirror, of the real world registered by the database".

Data mining refers to "using a variety of techniques to identify nuggets of information or decision-making knowledge in bodies of data, and extracting these in such a way that they can be put to use in the areas such as decision support, prediction, forecasting and estimation. The data is often voluminous, but as it stands of low value as no direct use can be made of it; it is the hidden information in the data that is useful".Data mining is also called knowledge discovery in databases (KDD). It is commonly defined as the process of discovering useful patternsor knowledge from data sources, e.g., databases, texts, images, the Web, etc. The patterns must be valid, potentially useful, and understandable. Data mining is a multi-disciplinary field involving machine learning, statistics, databases, artificial intelligence, information retrieval, and visualization. There are many data mining tasks. Some of the common ones are supervised learning (or classification), unsupervised learning (or clustering), association rule mining, and sequential pattern mining. We will study all of them in this book. A data mining application usually starts with an understanding of the application domain by data analysts (data miners), who then identify suitable data sources and the target data. With the data, data mining can be performed, which is usually carried out in three main steps:

The whole process (also called the data mining process) is almost always iterative. It usually takes many rounds to achieve final satisfactory results, which are then incorporated into real world operational tasks.Traditional data mining uses structured data stored in relational tables, spread sheets, or flat files in the tabular form. With the growth of the Web and text documents, Web mining and text miningare becoming increasingly important and popular.

1.1 Traditional Web Mining Techniques

Traditional information retrieval methods represent plain-text documents using a series of numeric values for each document. Each value is associated with a specific term (word) that may appear on a document, and the set of possible terms is shared across all documents. The values may be binary, indicating the presence or absence of the corresponding term. The values may also be a non-negative integers, which represents the number of times a term appears on a document (i e. term frequency). Non-negative real numbers can also be used, in this case indicating the importance or weight of each term. These values are derived through a method such as the popular inverse document frequency model [Sa189], which reduces the importance of terms that appear on many documents. Regardless of the method used, each series of values represents a document and corresponds to a point (i e. vector) in a Euclidean feature space; this is called the vector-space model of information retrieval. This model is often used when applying machine learning techniques to documents, as there is a strong mathematical foundation for performing distance measure and centroid calculations using vectors.

II. LITURATURE STUDY

A lot of contents are available on the web for this perspective and various work has been done in this direction, for web personalization. Personalization was tried to be provided on available data about user. But the ability to track user behavior in detailed level has raised a great possibility for a very efficient and user friendly environment to user.

In year 2011 RamyaC,Kavitha G [1] In this paper, a complete preprocessing methodology for discovering patterns in web usage mining process to improve the quality of data by reducing the quantity of data has been proposed. A dynamic ART1 neural network clustering algorithm to group users according to their Web access patter ns with its neat architecture is also proposed. Several experiments are conducted and the results show the proposed methodology reduces the size of Web log files down to 73-82% of the initial size and the proposed ART1 algorithm is dynamic and learns relatively stable quality clusters. Web log data is usually diverse and voluminous. This data must be assembled into a consistent, integrated and comprehensive view, in order to be used for pattern discovery. Withou t properly cleaning, transforming and structurin g the data prior to the analysis one cannot expect to find the meaningful patterns. Rushing to analyze usage data without a proper p reprocessing method will lead to poor results or even to failure. So we go for preprocessing meth odology. The results show that the proposed methodology reduces thesize of Web access log files down to 73-82% of the initial size and offers richer logs th at are structured for further s tages of Web Usage Mining .

In year 2011 Mrs. G. Sudhamathy and Dr. C. JothiVenkateswaran [2]As more organization rely on the Internet and the World Wide Web to conduct business, the proposed strategies and techniques for market analysis need to be revisited in this context. We therefore present a survey of the most recent work in the field of Web usage mining, focusing on three different approaches towards web logs clustering. Clustering analysis is a widely used data mining algorithm which is a process of partitioning a set of data objects into a number of object clusters, where each data object shares the high similarity with the other objects within the same cluster but is quite dissimilar to objects in other clusters. In this work we discussthree different approaches on web logs clustering, analyze their benefits and drawbacks. We finally conclude on the most efficient algorithm based on the results of experiments conducted with various web log files.

In year 2010 A. Anitha Member IEEE [3] Web usage mining is applied as data mining techniques[] to discover usage patterns from web data, in order to understand and better serve the needs of web based application. The aim of this paper is to discuss about a system proposed which would perform clustering of user sessions extracted from these web logs for each user which constitutes the dataset. Clustering is then performed on these datasets based on the key attributes to partition the users into several homogenous groups such that similar user access patterns belong to the same cluster Implementation and the results are also discussed.This paper discusses the design steps and result applied to the dataset and how it was clustered in one of the homogenous groups with similar activities. Segmentation or clustering is based on IP Address and time in the system. Several other attributes can also be chosen for clustering. After segmentation has been done on the web logs based on a defining attributes this can be utilized in variety of ways. If clustering is done on the basis of time, it can be used for the purpose of website modification and if it is done on the basis of IP addresses, it can be used to identify users with similar access patterns and can be used for the purpose of marketing.

In year 2009 K.R.Suneetha,Dr. R. Krishnamoorthi[4] Major part of the filtered entries containedextensions gif, jpeg, jpg, swf, cgi, asp, css, mov, ico, dll, exe, etc. The number of entries was reduced to43368 entries, which is approximately 82%reduction rate (see Table).At this stage the information available from the field cs-method was used to filter out all entries containingmethod of access either HEAD or OPTIONS (see Table).The next preprocessing stage, user identification, checked the entries for their client IP address (field c-ip) andclient agent type (field cs (User-Agent)) in order to identify users and to group together all entries of each ofthe users. If either changes, a new user was registered.

In year 2006 M Kellar, C Watters, M Shepherd [5] aimed at understanding of the high level tasks in which users engage on the Web. They conducted a field study in which participants were asked to annotate all web usage with a task description and categorization. Based on our analysis of participants' recorded tasks during the field study, as well as previous research.They developed a classification of web information tasks. The classification consists of three information goals: information seeking, information exchange, and information maintenance. Web information tasks consist of the set of tasks in which users engage on the Web that deal with some aspect of information, from acquisition, consumption, and distribution of information.

In year 2005 Jeffrey Heer, E d H. Chi [6] Recent research has explored web user session clustering as a means of understanding user activity and interests on the World Wide Web. Though the proposed techniques have proven to be useful and effective they require that one either specify the number of clusters in advance or browse a large hierarchy of clusters to find the optimal depth at which to describe user activity. In this paper, we examine the utility of a stability-based technique for automatically determining th e optimal number of clusters in the context of web user session clustering. We present two case studies evaluating the technique's effectiveness.A web user session clustering is gain acceptance as a means for understanding user activity and goals, it is important that these techniques are made easy and effective for web masters and web analysis to use Accordingly, automated tools for determining the structure of the clustered data would be of great benefit to those trying to understand user's interactions with the Web, directing the analyst immediately to the d esired level of presentation and allowing web designers to un derstand the high-level structure of user activity in site.

III. WEB MINING TECHNIQUES

Web mining aims to discover useful information or knowledge from the **Web hyperlink structure**, **page content**, and **usage data**. Based on the primary kinds of data used in the mining process, Web mining tasks can be categorized into three types: Web structure mining, Web content mining and Web usage mining.

3.1 Web Structure Mining (WSM):

Web Structure Mining is concerned with discovering the model underlying the link structure of web. It is used to study the topology of the hyperlinks with or without the description of the links. This model is used to categorize web pages. It is useful to generate information such as the similarity and relationship between different web sites. Web mining is also used to discover authority sites for the subjects and overview (or hub) sites for the subjects that point to many authorities. It is seen that Web content mining attempts to explore the structure within a document (intra-document structure). Web structure mining studies the structure of documents within the web itself (inter-document structure).Web structure mining discovers useful knowledge from hyperlinks (or links for short), which represent the structure of the Web. For example, from the links, we can discover important Web pages, which, incidentally, is a key technology used in search engines. We can also discover communities of users who share common interests. Traditional data mining does not perform such tasks because there is usually no link structure in a relational table.

The challenge for Web structure mining is to deal with the structure of the hyperlinks within the Web itself. Link analysis is an old area of research. However, with the growing interest in Web mining, the research of structure analysis had increased and these efforts had resulted in a newly emerging research area called Link Mining, which is located at the intersection of the work in link analysis, hypertext and web mining, relational learning and inductive logic programming, and graph mining. There is a potentially wide range of application areas for this new area of research, including Internet. The Web contains a variety of objects with almost no unifying structure, with differences in the authoring style and content much greater than in traditional collections of text documents. The objects in the WWW are web pages, and links are in-, out- and co-citation (two pages that are both linked to by the same page). Attributes include HTML tags, word appearances and anchor texts. This diversity of objects creates new problems and challenges, since is not possible to directly made use of existing techniques such as from database management or information retrieval. Link mining had produced some agitation on some of the traditional data mining tasks. As follows, we summarize some of these possible tasks of link mining which are applicable in Web structure mining.

1. Link-based Classification: Link-based classification is the most recent upgrade of a classic data mining task to linked domains. The task is to focus on the prediction of the category of a web page, based on words that occur on the page, links between pages, anchor text, html tags and other possible at- tributes found on the web page.

2. Link-based Cluster Analysis: The goal in cluster analysis is to find naturally occurring sub-classes. The data is segmented into groups, where similar objects are grouped together, and dissimilar objects are grouped into different groups. Different than the previous task, link-based cluster analysis is unsupervised and can be used to discover hidden patterns from data.

3. Link Type: There are wide ranges of tasks concerning the prediction of the existence of links, such as predicting the type of link between two entities, or predicting the purpose of a link.

4. Link Strength: Links could be associated with weights.

5. Link Cardinality: The main task here is to predict the number of links between objects.

There are many ways to use the link structure of the Web to create notions of authority. The main goal in developing applications for link mining is to made good use of the under- standing of these intrinsic social organization of the Web.

3.2 Web Content Mining (WCM):

Web content mining extracts or mines useful information or knowledge from Web page contents. For example, we can automatically classify and cluster Web pages according to their topics. These tasks are similar to those in traditional data mining. However, we can also discover patterns in Web pages to extract useful data such as descriptions of products, postings of forums, etc, for many purposes. Furthermore, we can mine customer reviews and forum postings to discover consumer sentiments. These are not traditional data mining tasks.Web content mining targets the knowledge discovery, in which the main objects are the traditional collections of text documents and, more recently, also the collections of multi- media documents such as images, videos, audios, which are embedded in or linked to the Web pages. Web content mining could be differentiated from two points of view: the agent-based approach or the database approach. The first approach aims on improving the information finding and filtering and could be placed into the following three categories

1. Intelligent Search Agents: These agents search for relevant information using domain characteristics and user profiles to organize and interpret the discovered information.

2. Information Filtering/ Categorization: These agents use information retrieval techniques and characteristics of open hypertext Web documents to automatically retrieve, filter, and categorize them.

3. Personalized Web Agents: These agents learn user preferences and discover Web information based on these preferences, and preferences of other users with similar interest.

The second approach aims on modeling the data on the Web into more structured form in order to apply standard database querying mechanism and data mining applications to analyze it. The two main categories are Multi-level databases and Web query systems.Its goal is gathering data and identifying patterns related to the contents of the web and the searches performed on them. There are two main strategies: Web page mining, extracting patterns directly from the contents existing in web pages. In this case the data in use can be

- Free text
- HTML pages
- XML pages
- Multimedia elements
- Any other type of contents existing in the web site.

Search results mining, intending to identify patterns in the results of the search engines. Mining, extraction and integration of useful data, information and knowledge from Web page contents.

3.3Web Usage Mining (WUM):

Web usage mining focuses on techniques that could predict the behavior of users while they are interacting with the WWW. Web usage mining collects the data from Web log records to discover user access patterns of Web pages. There are several available research projects and commercial products that analyze those patterns for different purposes. The applications generated from this analysis can be classified as personalization, system improvement, site modification, business intelligence and usage characterization

The challenges involved in web usage mining could be divided in three phases:

Pre-processing: The data available tend to be noisy, incomplete and inconsistent. In this phase, the data available should be treated according to the requirements of the next phase. It includes data cleaning, data integration, data transformation and data reduction.

Pattern discovery: Several different methods and algorithms such as statistics, data mining, machine learning and pattern recognition could be applied to identify user patterns.

Pattern Analysis: This process targets to understand, visualize and give interpretation to these patterns.

Web usage mining depends on the collaboration of the user to allow the access of the Web log records. Due to this dependence, privacy is becoming a new issue to Web usage mining, since users should be made aware about privacy policies before they make the decision to reveal their personal data.Every user leaves a path through the pages they have accessed when they visit a Web site. Web usage mining tries to discover the useful information such as navigation patterns from the secondary data derived from the interactions of the users while surfing on the Web. It focuses on the techniques that could predict user behavior while the user interacts with the Web. E-commerce specialists have recently focused on this behavioral data to understand how users decide to buy something.

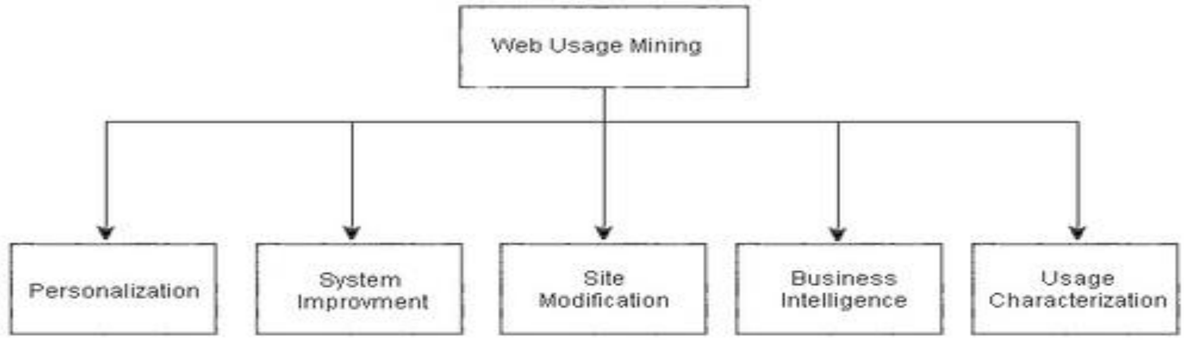

Fig 3.1.1: Web Usage Mining

Data warehouse reporting systems aggregate and report facts over different dimensions. These systems are commonly called online analytic processing (OLAP) systems. OLAP systems can report only on directly observed and easily correlated information. They provide information like most requested pages, top entry and top exit pages, most downloaded files, top visitors, geographic regions of visitors, client and server errors, top search keywords, most used browsers, most used platforms, etc. Data mining techniques are performed to this data to find new information. For example by performing association technique Amazon.com identifies items that are likely to be purchased or viewed in the same session and says when comparing Web content mining and Web structure mining, Web usage mining is a considerably new research area and it has gained more attention in recent years. Here the goal is to dive into the records of the servers (log-files) that store the information transactions that are performed in the web in order to find patterns revealing the usage the customers make of it. For example the most visited pages, usual visiting paths, etc. We can also distinguish here:

- **General access pattern** tracking: Here the interest doesn't rely on the access patterns of a particular visitor but on the integration of them into trends allowing us to re-structure the web in order to facilitate our customer's access and utilization of our web site.
- **Customized** access pattern tracking: Here what we look for is gathering data about the individual visitor's behavior and their interaction with the website. This way we can establish access/purchase profiles so that we can offer a customized experience to every customer. An archetypical case of this is amazon.com and its purchase advice and suggestions.

Web usage mining is the process of extracting useful information from server logs i.e. user's history. Web usage mining is the process of finding out what users are looking for on the Internet. Some users might be looking at only textual data, whereas some others might be interested in multimedia data.

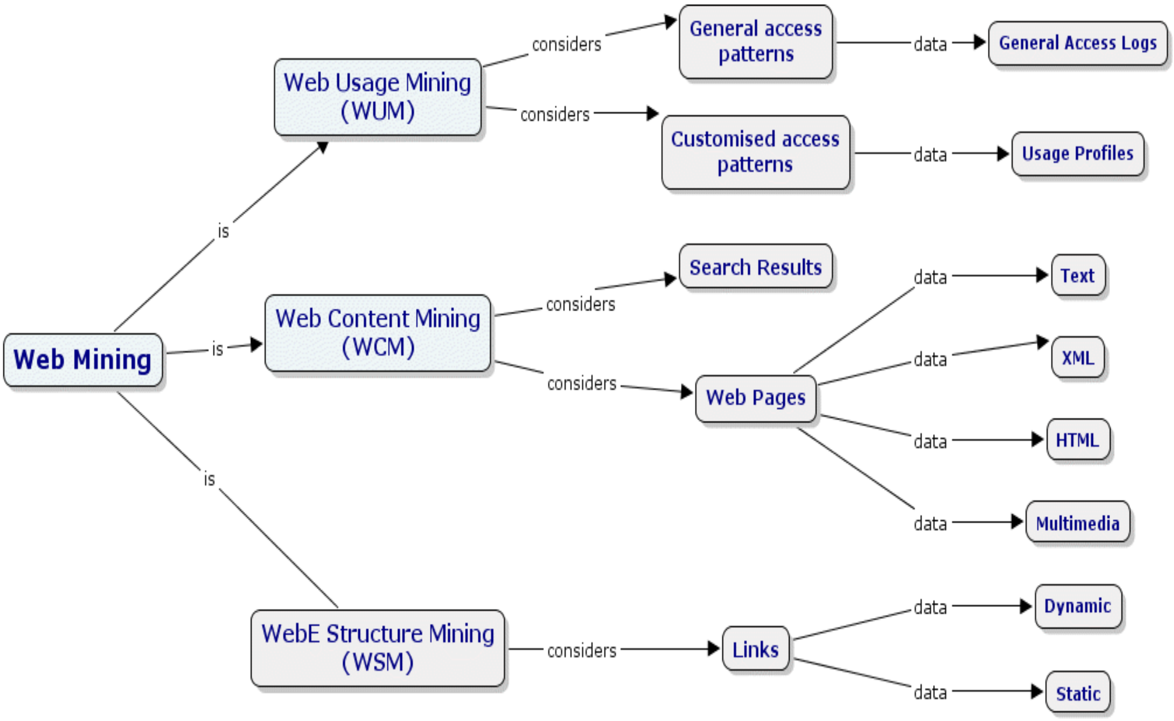

Figure 2: Web Mining

3.4 Challenges of Web Mining:

The web is a huge, distributed information center for different kind of information services such as government, electronic commerce, news, etc. The web is also a rich resource for linkage and usage information, providing data for data mining. However there are some difficulties to categorize and collect data for data mining and knowledge discovery. The Web is too big and it grows very fast for effective data warehousing and data mining. According to Jacob Nielsen from Sun Microsystems the number of servers on the Web is doubling every 53 days.

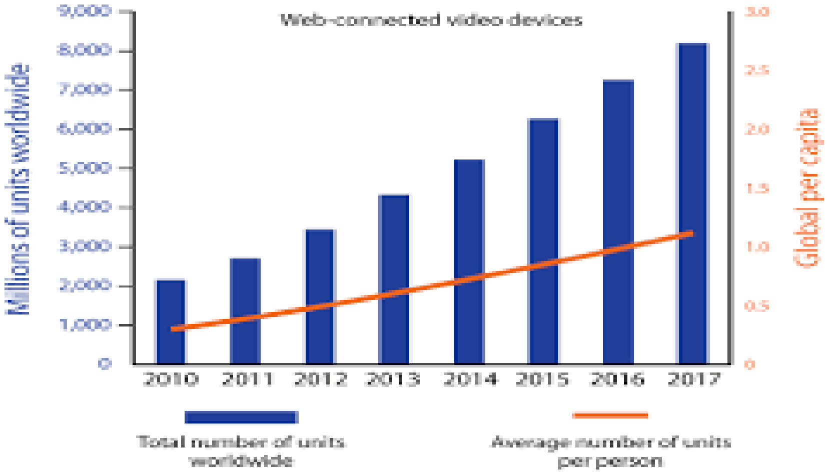

Fig 3: Growth of the Web between year 2000 and 2017

According to a study released by Bright Planet, an Internet content company, search engines like Google and Yahoo index only a little part of a huge reservoir of information. It is estimated that the Web is 500 times larger than what popular search engines have measured. The Web contains19 terabytes, compared to 7,500 terabytes hidden in the "deep" Web, according to the study. Because of the growth rate and mass of the Web it is almost impossible to set a data warehouse to store all of the data on the Web.The complexity of Web pages is another challenge. While some Web pages are just text-based some of them contain multimedia data such as images and videos. There is no index by category, author, and title and so on. Extensible Markup Language (XML) is an attempt to reduce the complexity of web pages. For example Financial Services have an XML standard. So the structure of the information exchanged is in a format that is agreed upon by the communicating parties and that is easily manipulated programmatically. It is obvious that the complexity of Web pages in general cannot be solved with XML.

The Web is a highly dynamic information source. It receives constant updates on the information it holds. Both Web page content and linkage information are updated frequently.The Web is a place where a broad diversity of user groups is involved. Users may have different backgrounds, usage purposes, languages and cultures. This variety, especially language issue, is a great challenge for data mining.It is said that 99% of the information on the Web is useless to 99% of the Web users. A particular person is only interested in a tiny part of the Web. Is it possible to determine the most relevant portion of the Web content for each individual’s interest? Can keyword-based search engines do that? Today’s search engines index Web pages depending on certain keywords. However, there are several deficiencies for keyword-based search engines. When the phase data mining is searched it may give some results related to other mining industries but fail to turn up results related to knowledge discovery. This is referred to as the polysemy problem. Homonym is a word having the same sound as another, but differing from it in meaning. Search results for the word Turkey contains data for both the country Turkey and the animal turkey. Last but not least synonymy is another problem for keyword-based search engines. Two synonym words are equivalent of each other in meaning but different in writing. Search results for a word do not contain its synonyms.

IV. DATA PREPROCESSING

The raw data is usually not suitable for mining due to various reasons. It may need to be cleaned in order to remove noises or abnormalities. The data may also be too large and/or involve many irrelevant attributes, which call for data reduction through sampling and attribute selection. Details about data pre-processing can be found in any standard data mining textbook.

The data in the log files of the server that stores details about the actions of the users can not be used for mining purposes in the form as it is stored. For this reason a preprocessing step must be performed before the pattern discovering phase.

It comprises following steps:

1. Merging of log files from different web servers
2. Data cleaning
3. Session identification of users.
4. Data formatting and summarization.

Web mining: The processed data is then fed to a data mining algorithm which will produce patterns or knowledge.Web mining involves six common classes of tasks:

- **Anomaly detection** (Outlier/change/deviation detection) – The identification of unusual data records, that might be interesting or data errors and require further investigation.
- **Association rule learning** (Dependency modeling) – Searches for relationships between variables. For example a supermarket might gather data on customer purchasing habits. Using association rule learning, the supermarket can determine which products are frequently bought together and use this information for marketing purposes. This is sometimes referred to as market basket analysis.
- **Clustering** – is the task of discovering groups and structures in the data that are in some way or another "similar", without using known structures in the data.
- **Classification** – is the task of generalizing known structure to apply to new data. For example, an e-mail program might attempt to classify an e-mail as "legitimate" or as "spam".
- **Regression** – Attempts to find a function which models the data with the least error.
- **Summarization** – providing a more compact representation of the data set, including visualization and report generation

4.2 Post Processing

In many applications, not all discovered patterns are useful. This step identifies those useful ones for applications. Various evaluation and visualization techniques are used to make the decision.KDD post-processing components can be categorized into the following groups, Buha and Famili [2000]: knowledge filtering; interpretation and explanation; evaluation; and knowledge integration. In the case of machine learning algorithms such as trees or decision rules trained with noisy data, the results are generated covering few training data. This is because the induction algorithms try to subdivide the training data set. To overcome this problem the decision trees or rules should be shrunk, by either post pruning (decision trees) or truncation (decision rules). After obtaining new knowledge, this can be either implemented in an expert system or used by an end user. In this last case, the knowledge results should be documented for the end user interpretation. Another possibility is to display the knowledge and transform it into a form understandable to the end user. We can also check the new knowledge for potential conflicts with previously induced knowledge. In this step, we can also summarize the rules and combine them with a domain-specific knowledge provided for the given task. Once a learning system has induced concept hypotheses (models) from the training set, their evaluation (or testing) should take place. There are several criteria used for this purpose: classification accuracy, understanding, computational complexity, and so on.

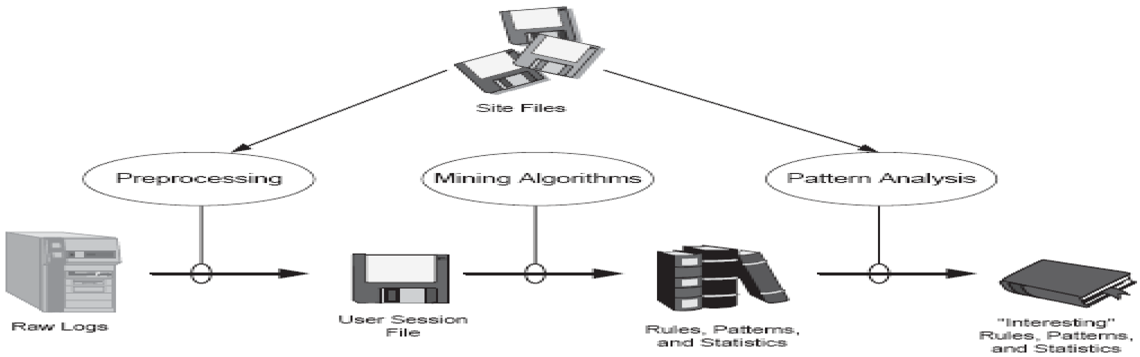

Fig 4: Data Mining Processes

Step 1:Merging

In this step we merge all log files in one log, after merging all log we sort them by time and then we got our main file in which we perform rest of the processes. Algorithm for merging log is as follows

```
Step 1:   create a new file named merge_log
Step 2:    initialize its curser
Step 3: initialize i=1
Step4:    read the log entries from log file Li
Step 5: append to merge_log
Step 6: repeat 3 and 4 until i <=N
Step 7:   Sort the merge_log entries in ascending order based on access time
Step 8: return merge_log
```

Step 2:Data Cleaning

Second step consist of cleaning useless entries from the log file. Since all the log entries are not valid we need to eliminate the irrelevant entries. Usually, this process removes requests concerning non-analyzed resources such as images, multimedia files, and page style files. But in our case we also have to remove all those entries which are generated during navigating from one page to another such as mouse move events occurring several times repeatedly.

Step 3:Session Identification

The different sessions belonging to different users should be identified. A session is understood as a group of activities performed by a user when he is navigating through a given site. To identify the sessions from the raw data is a complex step, because the server logs do not always contain all the information needed. If there is no other user identification available we can use IP address of System although we know that one IP address is used by several users through multiuser system such as a system in cyber café, or in a library.

Step 4: Data Formatting

This step is required only when we need to identify and group events by their category. We can also give them a unique identification code to summarize if required. Here we provide the table for identification of event according to over log file:

This thesis has presented the details of preprocessing tasks that are necessary for performing Click stream Data Mining in the application of data mining and knowledge discovery techniques for web log data collected by web servers. The contribution of the paper is to introduce the process of preparing web log data for different processes of click stream data mining, and then use web log for mining. The experimental results presented the importance of the data preprocessing step and the effectiveness of our methodology, by reducing not only the size of the log file but also increasing the quality of the data available through the new data structures that we obtained. The process itself does not fully guarantee that we identify correctly all the transactions (i.e. user sessions & visits). This can be due to the poor quality of the initial log file Therefore, we need a solid procedure that guarantees the quality and the accuracy of the data obtained at the end of data preprocessing. In conclusion, our methodology is more complete because:

- It offers the possibility to analyze several click stream data.
- It employs the effective cleaning of unnecessary data.
- It gives the formatted log file in a text file format.

Huge amounts of data are stored in different forms such as structured or unstructured, hypertext, and multimedia. Thus, mining complex types of data has become an increasingly important task in data mining. This paper summarizes the challenge, data sources and taxonomy of Web mining. Web mining is a wide research area in the cross-section of artificial intelligence, machine learning, neural networks, database, information retrieval, and semantic web search. Different categories of Web mining are related to different kind of these research topics

V. CONCLUSION

This paper has presented the details of preprocessing tasks that are necessaryforperforming Click stream WebMininginthe application of data mining and knowledge discovery techniques forweb log data collected by web servers. Thecontribution of the paper is to introduce the process of preparing web log data on the web for different processesof click streamdata mining, and then use web log for mining.

VI. FUTURE SCOPE

Here we perform a simple laboratory based experiment which may be defer for real log, it may be modified some morerealistic data. We can experiment for more click stream log and realistic and sophisticated data. There may be somechanges occurs on web log data. Some new phases may be introduced for time series data.The process itselfdoes not fully guarantee that we identify correctly all the transactions. This can be due tothe poorquality of the initial log file Therefore, we need a solid procedure that guarantees the quality and the accuracyof the data obtained at the end of data preprocessing.

VII. REFERENCES

[1] Fifth International Conference on Information Processing, Augus-2011,Bangalore, INDIA An Efficient Preprocessing Methodology for Discovering Patterns and Clustering of Web Users using a Dynamic ART1 Neural Network 2011

[2] Web Log Clustering Approaches – A SurveyG. Sudhamathy et al. / International Journal on Computer Science and Engineering (IJCSE) 2011.

[3] Mining Access Patterns using clustering(Mrs. Kiruthika M and Mrs. Dipa Dixit) International Journal of Computer Applications (0975 – 8887) Volume 4– No.11, August 2010.

[4] Identifying User Behavior by Analyzing Web ServerAccess Log File(K. R. Suneetha and Dr. R. Krishnamoorthi)IJCSNS International Journal of Co mputer Science and Network Security, VOL.9 No.4, April 2009.

[5] M Kellar, C Watters, M Shepherd aimed at understanding of the high level tasks in which users engage on the Web.2006.

[6] Mining the Structure of User Activity using Cluster Stability Jeffrey Heer, E d H. Chi 2005.

[7] A Web Page Prediction Model Based on Click-Stream Tree Representation of User Behavior MMM.TamerOzsu S. Gundus, 2003.

[8] Data Mining For Browsing Patterns In Weblog Data By Art2 Neural Networks (A.Nachev , NI.Ganchev) International Journal "Information Theories & Applications" Vol.10 2001.