Extraction and Analysis of User Profile from Event Logs

Anjali Jachak
Department of Information Technology
University of Pune

Anuj Sharma
Department of Information Technology
University of Pune

ABSTRACT
This paper discusses the analysis of data from log files and presents novel methods and ideas of analyzing these log files. Huge amount of data is generated on daily basis in every IT or non-IT organization. This data is stored in logs. These logs contain data which can prove valuable. These log files can be analyzed for different reasons. This data is initially collected and then this raw data is organized in such a way that frequent patterns can be recognized from this data set. Various techniques and algorithms are used for finding such patterns , but many of them don't prove useful on huge data sets. A number of products are available in the market .A number of algorithms have been proposed so far for mining frequent patterns. The data in these logs if used properly can prove useful in improving system performance and generating various reports on the usage of data. This information provides insights into user behaviors as well. To make this analysis easier, we need a tool which will extract the data, organize it, analyze it and generate suitable reports. Historical reports having older data also can prove vital if analyzed for drawing certain conclusions and predicting future use. Thus , our aim should be to provide with a platform-independent, low cost tool which can be affordable by smaller organizations as well. It should also be able to analyze the huge data sets generated by large organizations efficiently.

1. INTRODUCTION
A number of products are available in the market which contribute in analyzing patterns in huge datasets. The main advantage of this tool is that it is platform independent and easily affordable for smaller IT systems like schools, colleges and smaller institutes. This reduces the overhead of installing extra material and also is affordable. This tool is also user friendly and it not only analyzes the data in the data sets but also generates reports depicting the common and uncommon patterns which helps us in identifying certain parameters for better utilization of the resources. Most of the products available in the market today are platform dependent and costly. This makes them in-efficient for institutions which cant afford them. Finding frequent patterns is important but mining in-frequent patterns is also important. In this clustering plays an important role. Various clustering algorithms have been used so far but all of them have gained little success in finding less common patterns. Spatial data mining in particular is the discovery of interesting relationships that may exist in huge databases. Spatial data mining aims in a) extracting patterns b) Finding relationship between spatial or frequent and non-spatial data c) helps to present data regularity concisely at higher levels d) helps to recognize spatial databases to achieve better performance. Most of the algorithms suffer from problems like the user must provide the algorithms with spatial concept hierarchies, which are not available in most of the applications. Discovering the hierarchy is the most difficult problem in many cases. To deal with such problems we need some good cluster analysis techniques which have the ability to find structures without relying on the hierarchies. For cluster analysis to work, there should be natural similarity between the objects to be clustered. Whereas the traditional clustering techniques do not work for large data sets. Here we will be discussing one of the clustering algorithms. PAM is one such algorithm used for data mining[2]. It is used for searching hidden patterns that may be present in huge databases. Spatial data mining in particular is the discovery of interesting relationships that may exist in huge databases. PAM is a clustering algorithm based on partitioning.

2. ALGORITHM PAM
1. Select k representative objects arbitrarily.
2. Compute TC_{ih} for all pairs of objects O_i, O_h where O_i is currently selected and O_h is not.
3. Select the pair O_i ,O_h which corresponds to $\min_{Oi,Oh} TC_{ih}$. If the minimum TC_{ih} is negative, replace O_i with O_h and go back to step 2.
4. Otherwise for each non-selected object find, find the most similar representative object and halt.

O_i is the number of selected objects whereas O_h is number of unselected objects.

$TC_{ih}= j\ C_{jih}$

C_{jih} is the cost for all non-selected objects O_j.

3. CLUSTERING
There are many ways for defining frequent patterns from event logs. One of the ways is to find overlapping slices of data and other way is to find relation between overlapping and non-overlapping slices of data[4]. As already discussed , non-co-relating data also proves to be useful in finding the health of the system. The main highlight of our paper is the use of clustering which uses the PAM algorithm which we have discussed earlier[2]. Clustering is putting all the relevant objects in one cluster, putting all the irrelevant objects in another cluster and then finding the relation between them. Over here we will be categorizing the data into different categories and then we will be finding relations between the data. We will be profiling the data. Data profiling is the process of examining the available data and collecting statistics and information about that data. The main use of data profiling may be to:

- Find the whether the data can be used for developing some new applications.
- Find whether the available data consists of any patterns.
- For data management where data plays an important role in finding network faults or any other faults.
- For finding the usage patterns.

- For finding the amount of data being generated or used.
- For monitoring the quality of data.
- For generating periodic reports on the amount of data being used and the usage patterns which might help in various ways like restricting the usage and time of use in various institutions like academic institutions or corporate institutes.

Data profiling makes use of many statistics such as:

- Frequency
- Variation
- Minimum
- Maximum
- Length
- Type

The main purpose of data profiling is improving and understanding of data to the users.

Data governance ensures control over the data. It is the execution of authority over the data. It ensures that the data assets are formally managed by authorized people. It is about giving specific rights to specific people for handling the data and making it available to the entire organization without any misuse of the data. It includes the consistent and proper management of data . The master data management includes :

4. ANALYTICAL DATA: USED FOR ANALYZING THE DATA

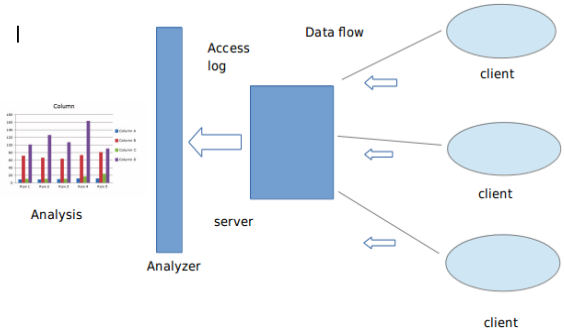
This analytical data is used for decision making. This decision making involves deriving conclusions from the analysis and finding out the relevant objects and irrelevant objects and generating reports based on the conclusions which might help for further maintenance or usage of data. Master data is the data generated or collected by the access log server[1]. This server receives the data from different client hosts of a organization. The server maintains this data. The data from different clients hosts is collected and maintained in one single file called the master file and then the analysis on this master data is done. The analysis includes inspecting ,cleaning, modeling and transforming the data with the aim to extract useful information. In the first step, the data is extracted from raw data then this data is analyzed for decision making. After drawing certain conclusions this data is processed and transformed into meaningful information[5]. The data is collected from varied sources. For example ,the data may be collected from access logs ,satellites, cameras, event logs etc.

After collecting the data , the data is organized and processed. This might involve placing the data into different fields etc. After the data is processed and organized, the data is checked for any errors, incompleteness or redundancy. The over here is checked for all these points and then further processed. In the next step, various algorithms may be applied to the data to identify relationships. The various data processing may involve:

- Sorting
- Categorizing
- profiling
- clustering
- co-relating

Concluding by finding facts and relating them.

We will be using data from the access logs. We will be processing this data, co-relating it and then generate reports after drawing certain conclusions. To manage a web server effectively and to make the use of the statistics, it is necessary to get the feedback about the activities and performance of the server. Storing the information in the access log is very useful in log management. The server access log records all the requests processed by the clients and the server. Statistical analysis plays an important role to study the interpretation, collection, analysis of the raw data available to us[1]. Statistical reports help us improvise our system in terms of quantitative analysis. Be it statistical analysis or any other analysis, the data that is to be analyzed should be divided in parts of relevance and then examined or analyzed. Here clustering plays a role for grouping the data. Clustering divides the data and forms groups of relevant and non-relevant data sets. This results in different types of clusters, each cluster containing same type of objects. Similarly ,a cluster of non-relevant objects can be formed. These clusters of non-co-relating and co-relating objects also can be analyzed and related



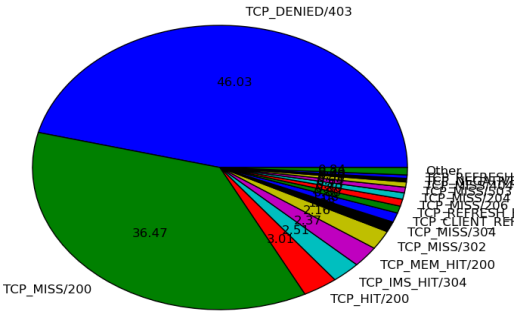
4.1 Finding Frequent Patterns

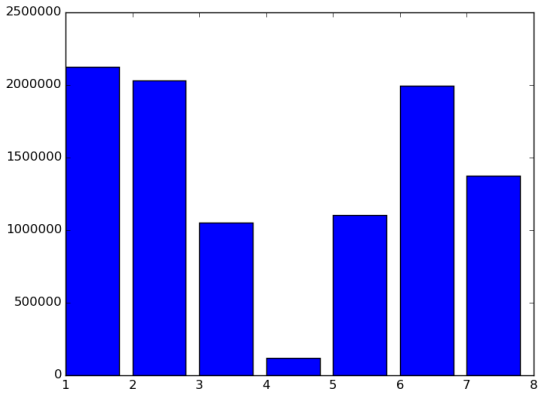
A huge data set comprises of various objects and variants of these objects. It becomes difficult to examine such huge amounts of data. Then data mining comes into picture where we find frequent patterns from such data sets and draw meaningful results out of these frequent patterns. This initially involves data extraction and then this raw data is converted into understandable structure so as to study it further. This helps in managing the data in a efficient manner. Processing data involves many steps as extracting, warehousing, analyzing, modeling, statistics and reporting. Various algorithms can be used for finding peculiar structures in the data sets. Automatic analysis of large quantities of data is very essential today taking in account the huge amount of data generated on daily basis. Analyzing this huge amount results in finding some interesting patterns which were undiscovered earlier or which could have never been found and proved useful otherwise. Clustering as we know performs group by mechanism but at the same time anomaly detection of unusual patterns or records is equally vital[2]. This may discover certain system flaws as well which would help us implementing steps to improve our system performance and make the most out of the data generated.

A compact representation of data is a must for making it easier every time we access it. Such as tables can be created and summarized data can be stored in it. When query is fired for a purpose the processing time can be reduced down and

the results can be displayed quickly, consuming less time. But this in turn may generate an overhead of extra memory usage.

For finding frequent patterns, we have to apply algorithms on these data sets. Before the algorithms are applied, the data must be arranged so as to make it easier for processing and less time consumption. Data mining plays its role in many fields be it medicine, law, arts, business, engineering, geography, weather and many more. In terms of business, periodic customer reports may contain certain patterns which may describe customer behavior, customer needs; in brief it results in better customer relationship management.





Graphs 2. Bytes vs Hours

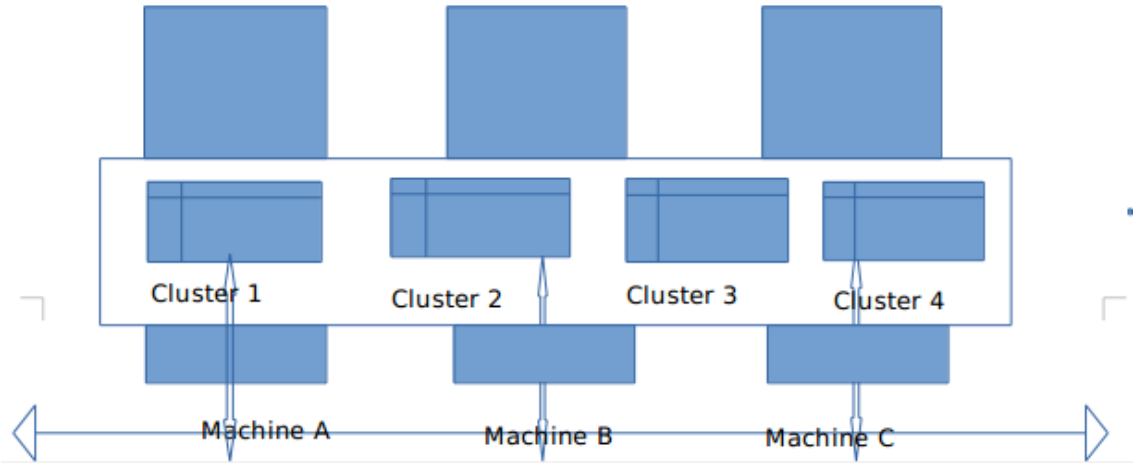


Fig 2. Clustering

5. CONCLUSION

In this paper we have developed a automated tool for analyzing data from access logs which is available on github.com(www.github.com/inheritedburrp/squidanalysispyt hon) Extracting this data, analyzing it and generating useful reports for further better usage of data is the basic task of this tool. For this purpose we have considered various parameters like time vs bandwidth, duration vs bandwidth etc. Previous work does not support analyzing huge amount of data, this paper discusses various ways and algorithms for analyzing huge amount of data and is platform independent and cost effective.

6. ACKNOWLEDGMENTS

The efforts taken in this research would not have been possible without the kind support and help of many individuals. We would like to extend our sincere thanks to all of them. We are highly indebted to Mr. Pawan Wawage for his invaluable guidance and constant supervision as well as for providing necessary support in completion of this research. We would also like to thank Department Of Information Technology of Vishwakarma Institute of Information Technology, Pune for providing the required facilities and information.

7. REFERENCES

[1] Risto Vaarandi. “A Data Clustering Algorithm for Mining Patterns From Event Logs". Proceedings of the 2013 IEEE Workshop on IP Operations and Management, pp. 119-126

[2] Raymond T.Ng, Efficient and effective clustering methods for data mining

[3] G.Jackobson, M.Weissman, L.Brenner, C.Lafond, C.Matheus.2000.GRACE: building next generation event correlation services.

[4] Xindong Wu,Xingquan Zhu ; Gong-Qing Wu ; Wei Ding “Knowledge and Data Engineering", Sch. of Comput. Sci. & Inf. Eng., Hefei

Univ. of Technol., Hefei, China, Jan 2014;

[5] Qingguo Zheng, Ke Xu, Weifeng Lv, and Shilong Ma. ``Intelligent Search of Correlated Alarms from Database Containing Noise Data". Proceedings of the 8th IEEE/IFIP Network Operations and Management Symposium,2012, pp 405-419.

[6] Dan Gorton.`` Extending Intrusion Detection with Alert Correlation and Intrusion Tolerance." Licentiate project, Chalmers University of Technology,2003.

[7] Risto Vaarandi. ``Platform Independent Event Correlation Tool for Network Management" Proceedings of the 8th IEEE/IFIP Network Operations and Management Symposium,2002 pp. 907-910.

[8] Sheng Ma and Joseph L. Hellerstein: ``Mining Partially Periodic Event Patterns with Unknown Periods." Proceedings of the 16th International Conference on Data Engineering (2000) 205-214.

[9] H.Mannila, H.Toivonen, and A.I.Verkamo, ``Discovery of frequent episodes in event sequences." Data Mining and Knowledge Discovery" 1(3) (1997) 259-289.

[10] Rakesh Agrawal and Ramakrishnan Srikant. ``Fast Algorithms for Mining Association Rules." Proceedings of the 20th International Conference on Very Large Data Bases, 1994 pp. 478-499.

[11] S. A. Yemini, S. Kliger, E. Mozes, Y. Yemini, and D. Ohsie. ``High speed and robust event correlation.