Identifying the Visitors with Data Mining Methods from Web Log Files

Uğur Gürtürk
Department of Software Engineering, Fırat University, Elazig, Turkey
23ugurgurturk23@gmail.com

Muhammet Baykara
Department of Software Engineering, Fırat University, Elazig, Turkey
mbaykara@firat.edu.tr

Murat Karabatak
Department of Software Engineering, Fırat University, Elazig, Turkey
mkarabatak@firat.edu.tr

Abstract – The usage of data stored in web search engines and on transaction logs of websites can provide valuable information for researchers related to the searched information and user behavior analysis. Within this context, some information, which is important for a network structure, can be obtained such as access time and access type. It can be especially beneficial in designing the information system, developing the interface, and improving the information architecture for content collections. In this paper, a set of samples of one month access log records of Fırat University website is collected and used. The set of samples is cleaned up with the log parser application developed in the data cleansing phase, which is the core of data mining. The cleaned data were converted to CSV format and analyzed using the BayesNet classifier method, which provides the best performance in the WEKA Software. As a result of the analysis it is seen that the future behavior of website users can be correctly estimated based on RemoteHostname.

Index Terms – Behavior Analysis; Data Mining; Information Extraction; Internet Access Log.

1. INTRODUCTION

The use of computers and the Internet in today's information age has started to rapidly spread in parallel to the development of technology. This widespread use has led many processes to be electronically implemented. This, allows the data in digital environments to rapidly grow on a daily basis and in a timely manner bringing about the complexity of data. The complex digital data generally need to be analyzed. For this reason, the necessity of applying data mining techniques to web technologies has increased. Accordingly, an increasing number of researchers has focused on this issue.

Data mining can be expressed as a process used to transform raw data into useful information [1]. Data mining is a powerful technology that helps businesses, organizations or individuals focus on the most important information in data warehouses, future vision, trends and behaviors, and making useful and right business decisions [2]. Data mining basically uses sophisticated mathematical algorithms to divide data into parts, evaluate the likelihood of future events, and provide efficient decisions and steps [3]. Data mining techniques can be implemented quickly on existing software and hardware platforms to enhance the value of existing information resources and can they be integrated with new products and systems. The basic features of data mining are:

- Authors automatic detection of patterns,
- Estimation of potential results,
- Creation of usable information,
- Focus on large datasets and databases,
- Answering unresolved questions with simple queries and reporting techniques [4].

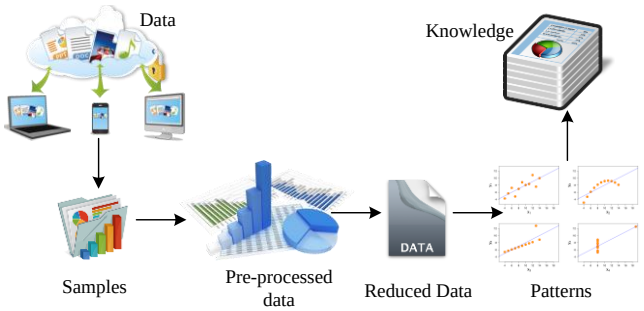

Figure 1 Data Mining Process Steps

Figure 1 shows data mining process steps. Among the given data mining steps, the first step and most important part of data mining is data selection process in order to be analyzed for making a decision. The data selection step is one of the most time-consuming steps among data mining and processing steps. In this process step, data generated in the system should be

well-chosen, and a good analysis should be applied to suit the correctness of the decision to be taken [5].

Another important step in implementing a successful data mining method is the pre-processing. In pre-processing, data should be presented in a convenient form for the future use [6]. The success achieved at this phase affects the success of the result at a high rate. As a result of the pre-processing, that is performed correctly and efficiently, a clear and precise result can be obtained.

After the pre-processing phase, the phase of obtaining useful and realistic information is the step of data reduction. This step ensures the reduction of data that is not suitable to use in next processing steps. This is applicable even for data that passed through certain pre-processing steps and transformed into required format [7].

In order to properly implement a data mining method, it is necessary to apply to use certain approaches to obtain the reduced data [8]. In this step, one or more known data mining techniques can be applied to the reduced data. More accurate and clear information can be obtained by combining different data mining methods.

Once data mining techniques have been applied to the obtained data, the produced results can be interpreted. Whether the comments to be made are correct can also be determined by the results of other data mining techniques applied to the same data [8]. In other words, it should be determined which of the methods applied to the database obtains a more accurate result. Comparing the success rate achieved by the applied methods with the success rate of other methods found in the literature ensures that the best results are obtained and these results can be verified.

2. RELATED WORK

Network devices used in information systems and computer networks keep a track of events developed on them. The records stored on these devices, called the log, examine and record the behavior of the users coming to that system. By means of these records, it is possible to determine the security incidents in the network and to take various precautions. The log, which is called the transaction log, is a file that contains communications between a system and its users. It can also be described as a data collection method that automatically captures type, content and time of a process done between a terminal with a system or a person. More explicitly, a transaction log is an electronic record of interactions between a web search engine and users searching for information on that web search engine [9].

Logging is the process of saving activity information to a web log file called a log when the web user sends a request to the web server. The main source of raw data is in the web access log, which will be referred to as the log file. These log files are initially kept for errors analysis, yet they spread across a wide range of applications with the increased size and complexity of electronic processes.

Network and security components running in information systems also generate many logs every day. However, when the logs of servers and clients are added, it is almost impossible to report this amount of information that may be valuable about visitors' movements and traffic. These actions can be associated with other data movements and traffic information, to be analyzed, in order to produce meaningful results. When the information cannot be assessed as effectively as it is, controlling process won't be achieved in a real and complete sense, and investments in uncontrollable information systems are not satisfactory.

The format of the log files, which are sensitive to the management and compulsory to be kept restricted by the law, may vary according to the type of daily resources. The text-based log files, which are independent of the server platform, can be found in three different places. These are [10]:

1) Web Servers,
2) Web Proxy Servers,
3) Client Browsers.

Server log files: It usually provides the most complete and accurate data usage. However, these log files have two important deficiencies. These deficiencies are:

- These logs contain special individual information. For this reason server owners usually keep them restricted.
- These log files do not record visited cached pages.

Cached pages are called from the local store of browsers or proxy servers, not web servers.

Proxy log files: They retrieve HTTP requests from users, forward them to a web server, and send the results to the user who is communicated by the web server. It has three major disadvantages:

- The proxy server is hard to deal with. Dealing with it needs advanced network knowledge and programming skills such as the know-how needed for TCP/IP.
- Request blocking is largely limited to most requests.
- The proxy log file on the web is used when a weblogging system performance is degraded because each page's request must be processed by a proxy simulator.

Client log files: Participants remotely test a website by downloading a special software that stores the web usage, or by changing the source code of an existing browser. HTTP cookies can also be used for this purpose. This type of log is a part of the information that is generated by a web server and stored for future access on computers. However, the drawbacks of this approach are:

- The design team needs to deploy a custom software and install end users.
- The technique here makes it difficult to acquire compatibility with a range of operating systems and web browsers.

A lot of valuable information can be gathered from the log records that are generated by the server and the client, which are held by many information that may be valuable about visitor's movements and traffic. Here are some statistics that can be obtained [10]:

- Who logs in at a certain time interval,
- Whether or not there is a change in the hostname, IP address, MAC address, etc. of the connecting computer,
- Information about who gets which IP address,
- Which pages do these IP addresses provide access to,
- Whether remote connection to the system occurs or not,
- Who established the VPN connection at which time in the system,
- •Information on who's accessing which file,
- Whether any of the accessed files have been deleted,
- Password change information successfully,
- Information about whether a computer account or user account has been created.
- The characteristics of the weblog access file that can provide such information are given in Table 1.

LOG FIELD NAMES	MEANING
date:	Date of activity
time:	Hour of activity
c-ip:	The ip address of the requesting user
s-ip:	The IP address of the server where the web site resides
cs-uri-stem:	Requested web address
cs-status:	The code of the given case
cs(user-agent):	Knowledge of the browser that the client used
cs-referer:	The knowledge of which source is the active adrese
cs-uri-query:	The query the client attempted to perform
cs-username:	The name of the authenticated user who accesses the server
cs(cookie):	The content of the cookie sent or received
sc-bytes:	The number of bytes sent by the server
cs-bytes:	The number of bytes received and processed by the server
time-taken:	Time taken by process (milliseconds)

Table 1 Log File Fields

3. PROPOSED METHODS

The classification can be defined as a prediction of a specific result according to some qualities, starting from educational data. To estimate the result, a particular classification algorithm operates on a set of qualifications and a set of training that includes the relevant result, often called the target or predictive quality. The algorithm tries to explore relationships between attributes that are likely to predict the results. Then, the algorithm is given a previously unseen dataset, called the set of estimates, containing the same set of qualities, except for an unknown prediction set qualifier. The algorithm analyzes the input and generates an estimate. The accuracy of the estimate indicates that the used algorithm is "good" [11].

After preliminary steps of data mining phases, parameter selection and the dataset selection to be tested will affect the performance of the model that will appear in the applications. Therefore, the result of the comparison will be dependent on the chosen classification algorithm. The used dataset in this study consists of a log file. The used log file is resulted from accessing Fırat University website. In our work, the proxy server handles log data that is in a text format with the .log file extension, which consists of Internet user access logs. This file, in the NCSA log format, is more than 1 GB in size. However, data mining methods have been applied select a specific set of samples. As a result of the operations performed on this dataset, analyzed by WEKA program, the highest performance is obtained from the BayesNet classification method. In the following sections, these operations are discussed in details.

3.1 Bayesian Network Classifiers

Bayesian network classifiers are a special type of Bayesian networks designed for classification problems. The randomly obtained set of variables and conditional dependencies are defined as a probabilistic graph model directed by Directed Acyclic Graph (DAG) [12]. Characterized as a graphical model that effectively encodes the common probability distribution for a large set of variables, the Bayesian classifier provides an efficient representation of the multivariate probability distribution of a set of random variables and makes various calculations based on this representation [13].

BayesNet is a model class that is used extensively to represent probabilistic information. In BayesNet, edges represent conditional dependencies, non-connected nodes represent conditionally independent variables.

$$p(X) = \prod_{i=0}^{n-1} p(X_i \mid \prod\nolimits_{X_I}) \quad (1)$$

In formula 1, X = (X0… Xn-1) is evaluated as the vector of the variables. ∏xi is representing the clusters of the parents of XI in the network. p(Xi | ∏xi) is the conditional probability of Xi. The distribution here can be used to create new examples from conditional and marginal probabilities. The benefits of Bayesian Networks are given as follows:

- They do not need to have a prior information about the problem.
- They have the ability to protect a high level of interaction.
- They enable the architectural blocks to be efficiently combined and fused together in accordance with a specified layout.
- They use data modeling to estimate the common distribution of the solution that seems positive in terms of the outcome [14].

4. DEVELOPED APPLICATION AND ANALYSIS

All operations made by the visitors on the site are recorded by the server where the user access information called log is stored. These recorded data are made into useful information by carrying out specific analyses. This phase is included in the literature as web mining. These activity logs are used by system administrators to record interesting events that occur on the system. The type of information stored in any activity log is often a function of the purpose of the monitoring application/tool used to create and update the log. In other words, creating activity logs for different types of system activity is achieved using different types of monitoring tools [15].

These daily events' data, which are held by different monitoring tools, are kept on different file types. Significant data can be obtained from the held access records. The acquisition of these data has gained importance both in terms of statutory requirements and standards, as well as in terms of system performance, and has caused it to be regarded as a responsibility. In this regard, in accordance with the Turkish law no 5651, daily log analysis should be performed professionally in order to keep the log files together with the time stamps of the obtained files. The first step of these analyses, the Log Parser process, is important in extracting meaningful data from log records. Because these records are confusing and have some parts that are irrelevant [16].

In order to obtain meaningful data from the existing log records, the first step is to separate the log file. For this purpose, a Log Parser application was implemented with C# programming language on Visual Studio 2015. Using this application, the log file is initially loaded into the system and then parsed with the Parse Library and transferred to the appropriate form for analysis. The produced log data, which is suitable for analysis, can be saved with the extension .XLS for the file to be sent by the users. Figure 2 shows a general diagram of the developed system in this study.

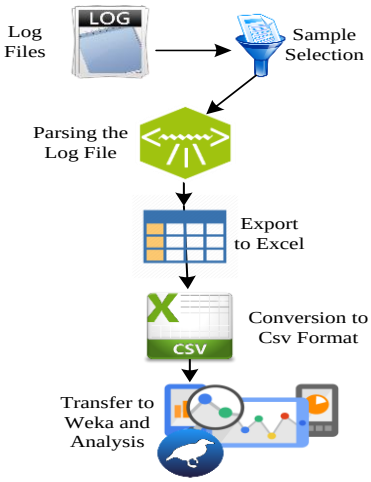

Figure 2 Developed Application Steps

A screenshot of this software is shown in Figure 3.

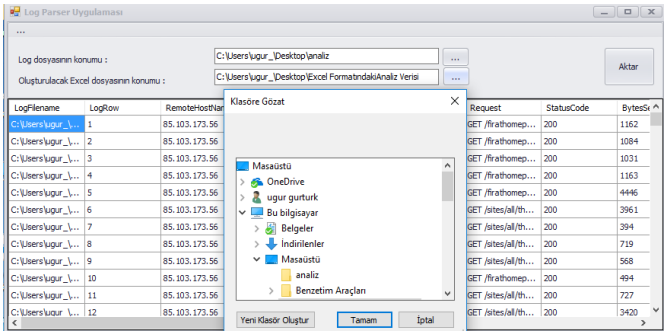

Figure 3 Screenshot of the Developed Log Parser Application

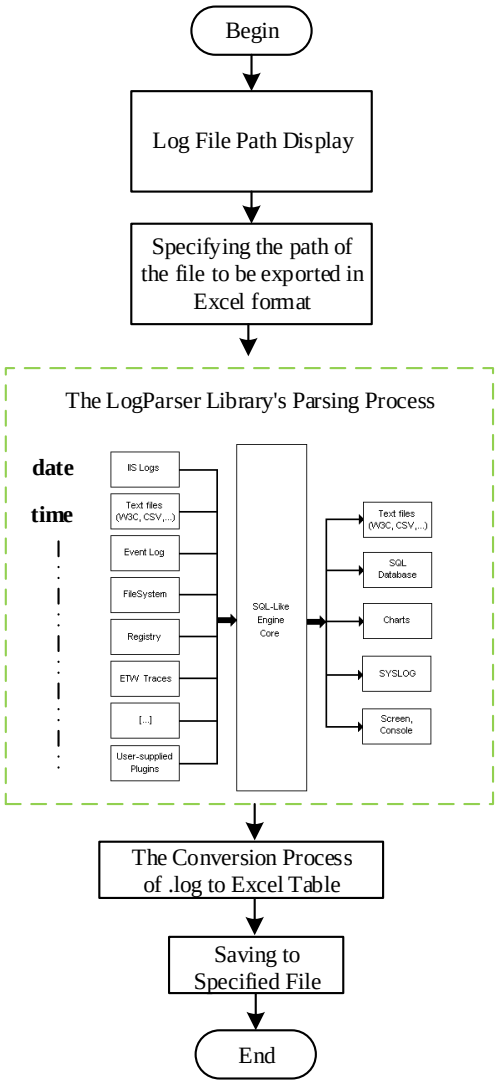

Figure 4 Flowchart of the Developed Log Parser Application

The existing dataset is cleaned with the developed log parser software. Since the required data for the implemented data

mining techniques is represented in some parts of the access records, consistent and necessary data should be taken from the log records and remnants should be cleaned up. Figure 4 shows the flow chart of the developed log parser software.

With this application and similar studies, as initial steps of data mining, data reduction and processing have been realized in order to provide important information on website administrators to system administrators. [17].

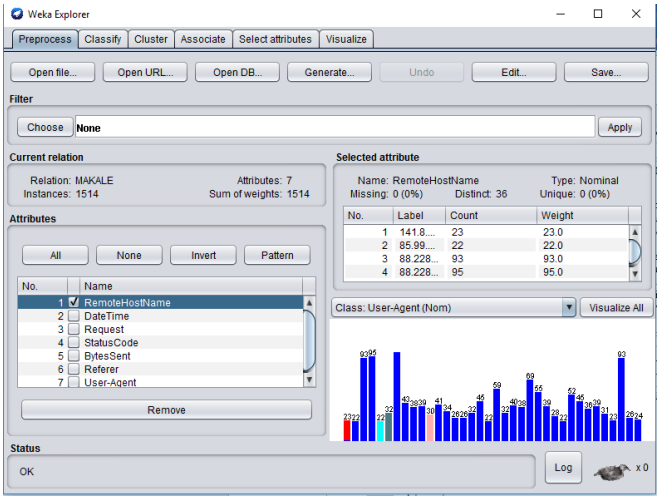

Figure 5 Opening of the Log Dataset in Weka And Distributions of Remotehostname Class

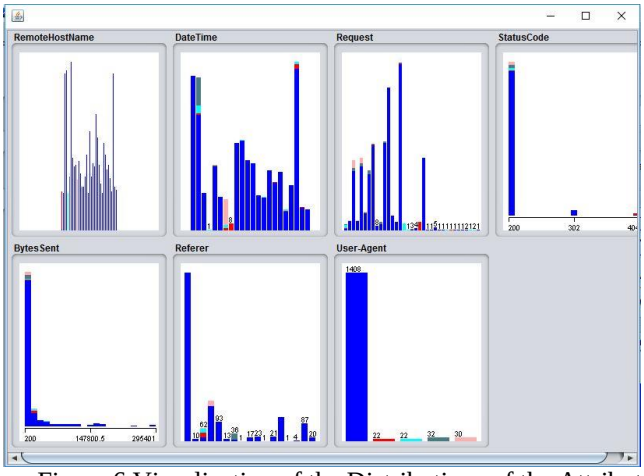

Figure 6 Visualization of the Distributions of the Attributes of the Dataset

The log data converted into the appropriate file format is converted to .CSV format online. Then, the .CSV format data file is opened with WEKA Software. All available log data instances are classified with all algorithms in WEKA and the result of this classification are compared according to the performance values. Table 2 shows the performance results. The most successful result of this comparison is provided by the BayesNet classifier presented under section 3. By BayesNet classifier, the predictions, which will be mentioned in the conclusion section, are realized in the direction of some information in the accesses to the Fırat University website on certain days and dates. In WEKA, the RemoteHostName property is handled during the classification phase with BayesNet and the estimates are made accordingly. Table 2 gives the results of classification with other classifiers.

Classification Method	Accurate Classified Samples	Performance Rate
BayesNet	1197	79%
NaiveBayes	1136	75%
RandomForest	1078	71%
RandomTree	1023	68%
Ibk (Nearest K-Neighbor)	1004	66%

Table 2 Performance of Different Classification Algorithms

As shown in Figure 5, the date-time information in the log file is DataTime, the Request is the page the user has requested, the requested response status is StatusCode, BytesSent is the number of bytes sent by the server, and the page that the user has visited before arriving at the page Referer a classification was made by referring to the information of the browser that the user was using as well as the User-Agent properties. 1514 rows of data are used in this analysis process; these data instances are subject to a cleaning process to discard the information that can be extracted during the analysis phase. The analysis is performed on 36 different IP addresses provided access more than 20 of the IP addresses on the transmitted dataset. The information of users provided access 20 and below has been deleted. According to the information contained in the log file, the results of this analysis achieved 79.0621% performance value in 36 different user classifications. The results of the performance rate of the analysis are shown in Table 3.

=== Stratified cross-validation ===		
=== Summary ===		
Correctly Classified Instances	1197	79.0621 %
Kappa statistic	0.7826	
Mean absolute error	0.0163	
Root mean squared error	0.084	
Relative absolute error	30.4717 %	
Root relative squared error	51.2984 %	
Total Number of Instances	1514	

Table 3 The Performance Rate of the Analysis

In Table 4, a confusion matrix, which is also known as the error matrix in machine learning and statistical classification problem, is obtained according to the classification result. This matrix, called the confusion matrix, is a table structure that allows you to visualize the performance of the algorithm. Each column of the matrix represents instances of a class that performs an estimate, and each row represents instances of a real class [18] or this operation is taken as the opposite. Since this matrix contains 36 different IP addresses and it is not possible to retrieve the entire matrix, a specific part is shown in Table 4.

Table 5 shows results of the Model and Evaluation on Training Set Clustered Instances obtained from WEKA software. As it seen in Table 5, 1197 rows were correctly classified, and 317 rows were incorrectly classified of total 1514 rows. Figure 7 shows general results of the developed system. In this study, there are 36 different classes of RemoteHostName that we classify in used log file. These classes were classified with using the classifiers given in Table 2. Finally, after this classification it was seen that the best results were obtained from BayesNet algorithm. Among the five algorithms that provide the best results, the algorithm with the least performance is K Nearest Neighbors algorithm.

=== Confusion Matrix ===

a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	q	r	s	t	u	v	w	x	y	z	aa	ab	ac	ad	ae	af	ag	ah	ai	aj	<-- classified as
22	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	a = 141.8.143.47
0	0	0	0	0	22	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	b = 85.99.220.171
0	0	13	77	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	c = 88.228.204.13
0	0	52	40	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	d = 88.228.91.150
0	0	0	0	22	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	e = 10.1.1.28
.....	.....
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	22	0	0	0	0	0	0	0	0	0	aa = 37.201.169.10
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	52	0	0	0	0	0	0	0	0	ab = 88.175.93.137
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	6	0	0	0	0	0	0	0	38	0	0	0	0	0	0	0	0	0	ac = 85.106.201.187
0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	34	0	0	0	0	0	0	0	0	0	ad = 176.219.167.101
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	39	0	0	ae = 88.228.102.152
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	23	3	5	0	0	af = 95.13.110.219
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	3	15	3	0	0	ag = 194.27.118.38
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	91	0	0	ah = 81.213.47.152
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	26	0	ai = 85.98.79.75
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	24	aj = 85.104.147.182

Table 4 Confusion Matrix

Model and Evaluation on Training Set		
Clustered Instances		
0	1197	79%
1	317	21%

Table 5 Training Set Model and Evaluation

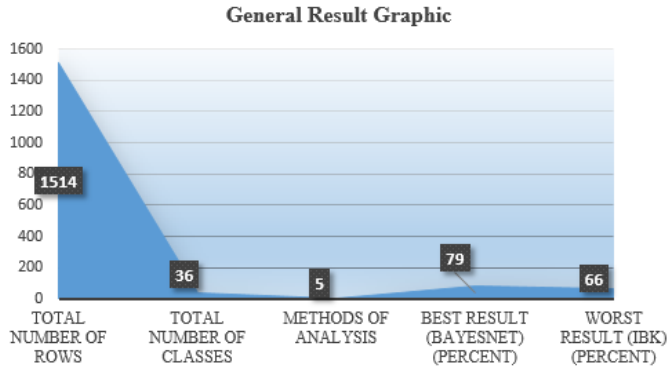

Figure 7 General Results

5. CONCLUSION AND EVALUATION

As a result of the widespread use of the Internet, the analysis of the log files stored on web servers has become important. The increasing use of the Internet makes it necessary to analyze log files in order to understand and resolve many events, especially security issues, performance, judicial processes and so on. We are informed about many issues such as analyzing log files, detecting security violations and collecting evidence, monitoring performance, detecting successful and unsuccessful accesses. This information shows how important log files are in terms of system administrators. Data mining techniques such as Merging, Clustering, and Classification can be applied only to interested regular user groups to find frequently accessed patterns, resulting in less time consuming and memory usage, which in turn produces higher accuracy and performance.

In this study, an analysis of collected log files was carried out. A sample set was selected from the access log data, with the developed parser application, it was segmented according to the columns and transferred into the appropriate format. This data is converted to CSV format over the Internet. It was then transferred to WEKA software. Less than 20 of these access information that received at different times were removed from the system, and more than 20 requests were included in the study. According to the information contained in the log file, the results of this analysis achieved 79.0621% performance value in 36 different user classifications. Accordingly, it is predicted that in the case of any malicious activity to be carried out in future, the user who performs this activity can be detected using the stored log records. For example, it may be possible to determine which malicious user or activity is taking place rather than from which computer in a laboratory environment. It is suggested to enlarge the size of the log file uploaded to the system when the Log Parser application is parsed and transferred to the appropriate form in progress with the implemented application. The size of the basic log file used in this study was determined to be more than 1 GB. However, the developed software for the parser does not support the decomposition of data with such a large size. Therefore, further work will be needed to eliminate this deficiency and achieve higher performance with more data.

REFERENCES

[1] B. J. Jansen, "Search log analysis: What it is, what's been done, how to do it." Library & information science research 28.3, pp. 407-432, 2006.

[2] M. Spiliopoulou, "The laborious way from data mining to web log mining." Computer Systems Science and Engineering 14.2, pp. 113-126, 1999.

[3] I. Çınar, M. S. Çınar, H. Ş. Bilge, "Web Sunucu Loglarının Web Madenciliği Yöntemleri ile Analizi", Academic Informatics 14 - XIV. Academic Informatics Conference Reports, 2014.

[4] G. G. Emel, Ç. Taşkın, "Veri madenciliğinde karar ağaçları ve bir satış analizi uygulaması." Eskişehir Osmangazi University Social Science Journal 6 (2), 2005.

[5] M. Çağlayan, B. Çekirge, D. Birant, P. Yıldırım, "Mobil Uygulama ile Görüntü İşleme ve Veri Madenciliği Tekniklerine Dayalı Melanom

Tahmin Desteği Sağlanması.", Innovations and Applications in Intelligent Systems Symposium (ASYU), 2014.

[6] U. Ekim, "Veri madenciliği algoritmalarını kullanarak öğrenci verilerinden birliktelik kurallarının çıkarılması", Ph.D. Thesis Selçuk University Graduate School of Natural Sciences, 2011.

[7] M. S. Aktas, O. Kalıpsız, "Veri Madenciliğinde Özellik Seçim Tekniklerinin Bankacılık Verisine Uygulanması Üzerine Araştırma ve Karşılaştırmalı Uygulama", Proceedings of the 9th Turkish National Software Engineering Symposium (UYMS 2015), Yasar University, Izmir, Turkey, September, 9-11, 2015.

[8] Akçapınar, Gökhan. "Çevrimiçi Öğrenme Ortamındaki Etkileşim Verilerine Göre Öğrencilerin Akademik Performanslarının Veri Madenciliği Yaklaşımı İle Modellenmesi." (2014).

[9] B. Bart, G. Verstraeten, D. V. Poel, M. E. Petersen, P. V. Kenhove, J. Vanthienen, "Bayesian network classifiers for identifying the slope of the customer lifecycle of long-life customers." European Journal of Operational Research 156.2, pp. 508-523, 2004.

[10] S. Çalışkan Kırmızıgül, İ. Soğukpınar, "K×Knn: K-Means Ve K En Yakın Komşu Yöntemleri İle Ağlarda Nüfuz Tespiti." EMO Proceeding, pp. 120-124, 2008.

[11] Friedman, Nir, Dan Geiger, and Moises Goldszmidt. "Bayesian network classifiers." Machine learning 29.2-3, pp. 131-163, 1997.

[12] Muralidharan, V., and V. Sugumaran. "A comparative study of Naïve Bayes classifier and Bayes net classifier for fault diagnosis of monoblock centrifugal pump using wavelet analysis." Applied Soft Computing 12.8, pp. 2023-2029, 2012.

[13] Bouckaert, Remco R. "Bayesian network classifiers in weka." Hamilton: Department of Computer Science, University of Waikato, 2007.

[14] E. Ardıl, "Esnek hesaplama yaklaşımı ile yazılım hata kestrimi." (2009).

[15] Giuseppini, Gabriele. "Log parser." U.S. Patent Application No. 10/461,672, 2003.

[16] M. Baykara, R. Daş, "Web Sunucu Erişim Kütüklerinden Web Ataklarının Tespitine Yönelik Web Tabanlı Log Analiz Platformu", International Journal of Science and Technology, Vol. 28 No. 2, pp.291-302, 2016.

[17] T. Özseven, M. Düğenci, "Log Analiz: Erişim Kayıt Dosyaları Analiz Yazılımı ve GOP Üniversitesi Uygulaması." International Journal Of Informatıcs Technologıes 4.2, 2011.

[18] S. V. Stehman, "Selecting and interpreting measures of thematic classification accuracy." Remote sensing of Environment 62.1, pp. 77-89, 1997

Authors

Uğur Gürtürk was born in Elazig, Turkey. He received his BS in Software Engineering from Fırat University in 2016. Currently, he is a MSc student in the Department of Software Engineering at Fırat University and works as a software engineer in Netix Information Technologies. His research interests are Data Mining, Big Data, Information Security, Intrusion Detection and Prevention Systems.

Muhammet Baykara was born in Elazig, Turkey. He received his BS and MSc. in Computer Engineering from Fırat University in 2006, 2009 respectively. He received his PhD. in Software Engineering from Fırat University in 2016. Currently, he is a research assistant in the Department of Software Engineering at Fırat University. His research interests are Information Security, Honeypots, Intrusion Detection and Prevention Systems.

Murat Karabatak received BS degree in 1999, Master degree in 2002 and Ph.D. title in Electric and Electronics Engineering Department at Fırat University in 2008. He worked for University of Sam Houston State (Texas-USA) as visitor researcher in 2014. Assoc. Prof. Dr. Karabatak is vice president of Software Engineering Department now. His interest areas are Data Mining, Big Data, intelligent systems, image processing and e-learning.