Sudan University Of Science And Technology
College Of Graduate Studies

College of Computer Science and Information Technology

Proposing Enhancement Of Websites According To User Behavior Prediction basing Web Usage Mining

A dissertation Submitted in Partial Fulfillment of the Requirements for MSc Degree in Information Technology

مقترح تحسين مواقع الويب وفقًا لـلتنبؤ بسلوك المستخدمين باستخدام تقنية تعدين الويب

BY :

Yassir Yousif Saeed Al-Haj

Supervised by

Dr. Wafaa Faisal Mukhtar

August 2021

Introductive

بسم الله الرحمن الرحيم

قال تعالى: **{ اقْرَأْ بِاسْمِ رَبِّكَ الَّذِي خَلَقَ ﴿١﴾**
خَلَقَ الْإِنسَانَ مِنْ عَلَقٍ ﴿٢﴾
اقْرَأْ وَرَبُّكَ الْأَكْرَمُ ﴿٣﴾
الَّذِي عَلَّمَ بِالْقَلَمِ ﴿٤﴾
عَلَّمَ الْإِنسَانَ مَا لَمْ يَعْلَمْ ﴿٥﴾}

سورة العلق - آية (1 – 5)

Dedication

I dedicate this work to my beloved family, my wife , friends, college,

classmates and everyone who support me …

Thanks to all of you “

To mister. Hisham Abdullah

Acknowledgements

I would like to express my special thanks of gratitude to my

Teacher's as well as my supervisor (**Dr. wafaa fisal , Hisham Abdullah**) whose gave me the golden opportunity to work with them on this thesis

which also helped me in doing a lot of research and i came

to know about so many …

Abstract

Web usage mining is the application of data mining techniques to discover usage patterns from Web data, in order to understand and better serve the needs of Web-based applications. Web usage mining consists of three phases, namely preprocessing, pattern discovery, and pattern analysis. Web Usage Mining Web servers, proxies, and client applications can quite easily capture data about Web usage. Web server logs contain information about every visit to the pages hosted on a server. Some of the useful information includes what files have been requested from the server, when they were requested, the Internet Protocol (IP) address of the request, the error code, the number of bytes sent to the user, and the type of browser used. The problem of this research is need to analyse the web site of (sudan university of science and technology) in order to enhance the site. In this research firstly collect data then prepare (preprocess) data to analysis (applying clustering) and evaluating the result (up to down methodology) . According to result give recommendation. By using flexable design and Offer more information in each page this will make the site more effective.

المستخلص

التنقيب عن استخدام الويب هو تطبيق لتقنيات استخراج البيانات لاكتشاف أنماط الاستخدام من بيانات الويب ، من أجل فهم احتياجات التطبيقات المستندة إلى الويب وخدمتها بشكل أفضل. يتكون تعدين استخدام الويب من ثلاث مراحل ، وهي المعالجة المسبقة واكتشاف الأنماط وتحليل الأنماط. تعدين استخدام الويب يمكن لخوادم الويب والوكلاء وتطبيقات العميل التقاط البيانات حول استخدام الويب بسهولة تامة. تحتوي سجلات خادم الويب على معلومات حول كل زيارة للصفحات المستضافة على الخادم. تتضمن بعض المعلومات المفيدة: ما هي الملفات التي تم طلبها من الخادم ، ومتى تم طلبها ، وعنوان بروتوكول الإنترنت (IP) الخاص بالطلب ، ورمز الخطأ ، وعدد وحدات البايت المرسلة إلى المستخدم ، ونوع المتصفح المستخدم . مشكلة هذا البحث ضرورة تحليل موقع (جامعة السودان للعلوم والتكنولوجيا) من أجل تحسين الموقع. في هذا البحث ، تم أولاً جمع البيانات ثم تحضير البيانات (ما قبل المعالجة) للتحليل (تطبيق التجميع) وتقييم النتيجة (منهجية من أعلى إلى أسفل). وفقا للنتيجة تعطي التوصية. باستخدام التصميم المرن وتقديم مزيد من المعلومات في كل صفحة ، سيجعل هذا الموقع أكثر فعالية.

List of Tables

List of Figures

Contents

CHAPTER ONE

INTRODUCTION

CHAPTER ONE

INTRODUCTION

1.1 Research background

Internet has become increasingly important as a medium for life, work and study as well as for dissemination of information. Web mining is the mining of data related to the World Wide Web. It is categorized into three active research areas according to what part of web data is mined, of which Usage mining, also known as web-log mining, which studies user access information from logged server data in order to extract interesting usage patterns. Web mining is the intelligent analysis of Web data.. By the use of Web browsing patterns, business organizations can perform mass customization and personalization, adapt their Web sites, and further improve their marketing strategies, product offerings, and promotional campaigns. Therefore, Web browsing pattern mining has special meaning for business organizations. Thus, it has attracted much attention from data mining, machine learning, and other research communities for many years. Of the existed methods, some arenon-sequential, such as association rule mining and clustering; and some are sequential,such as sequential or navigational pattern mining. Web mining is the application of data mining techniques to automatically discover and to extract knowledge from web data.

1.2 Problem statement

The majority of institutions in sudan have official websites , but they are ineffective (meaning that they are not used by the beneficiaries of the service) and this may be due to several reasons.

Sudan university of science and technology website data (the logfile) was taken as a dataset.

The research problem can be viewed from three aspects:

1. There is a need to analyze the web site of sudan university of science and technologyin order to recommend guide lines to enhance the web site .
2. Not resorting to the official websites of institutions in general .
3. difficulty of obtaining a questionnaire from the site to find out the problem.

1.3 Research objectives

analize sust website log file ,in order to decover patterns to help developers making it easy to use and manage. To achieve that can be viewed from three aspects:

1. Know the activity of user on web site.

2. Site analysis to help find the required modifications on the web site.

3. Help in decision making.

1.4 Research Methodology

In this research firstly collect data then prepare (preprocess) data to analysis (applying clustering) and evaluating the result (up to down methodology) . According to result give recommendation.

1.5 Research scope

This research focus on applying data mining techniques (web usage mining) for supporting official sust-website by extracting knowledge from website logfile which obtained from Sudan university for science & technology (sust)(access log file 14-2-2017 to 1-5-2017).

1.6 Thesis Organization

In chapter one, A comprehensive way about mining in the Web.

In chapter two, more detail about mining in web usage its process,techniques, tools and some research issues.

In chapter three, clustering and classification technique

In data mining it’s defined, types, explain how the algorithm work (only one algorithm for each type) and the comparison between them.

In Chapter four, implementation of research tool (as detail). talked first about the ‘ORANGE, WEEKA and WEBLOG 'as a tool, then for how to use them inthe clustering process and analysis, and we discussed the results and Synopsis ofthe application.

In Chapter five, Conclusions and recommendation for future work.

CHAPTER TWO

LITERATURE REVIEW AND RELATED WORK

CHAPTER TWO

LITERATURE REVIEW AND RELATED WORK

2.1 Introduction

This chapter discusses the literature concerned with Data Mining and Knowledge Discovery techniques used in user behavior analysis (usage mining). The chapter firstly show importance of data Secondly, it discusses using Data Mining techniques (usage mining) in log file Analysis, and some useful tools.

2.2 Important of Data

With the continued growth and proliferation of e-commerce, Web services, andWeb-based information systems, the volumes of clickstream and user datacollected by Web-based organizations in their daily operations has reachedastronomical proportions. Analyzing such data can help these or-ganizations determine the life-time value of clients, design cross-marketing strategies acrossproducts and services, evaluate the effectiveness of pro-motional campaigns,optimize the functionality of Web-based applications, provide more personalized content to visitors, and find the most effective logical structure for their Web space. This type of analysis involves the automatic discovery of meaningful patterns and relationships from a large collection of primarily semi-structured data, often stored in Web and applications server access logs, as well as in relatedoperational data sources.

2.3 Data mining

Discovering new information or knowledge from data. For example, data retrieval techniques are mainly concerned with improving thespeed of retrieving data from a database, whereas data mining techniques analyze the data and try to identify interesting patterns.

It should be noted, however, that the distinction between information retrieval andtext mining is not clear. Many applications, such as text classification and text clustering, are often considered both information retrieval and text mining Similarly, Web retrieval and Web mining share many similarities. Web document clustering has been studied both in the context of Web retrieval and Web mining. On the other hand, Web mining is not simply the application of information retrieval and text mining techniques to Web pages; it also involves non-textual data such as Web server logs and other transaction-based data. Although Web mining relies heavily on

data mining and text mining techniques, not all techniques applied to Web mining are based on data mining or text mining. Some techniques, such as Web link structure analysis, are unique to Web mining. In general, it is reasonable to consider Web mining as a subfield of data mining, but not a subfield of text mining, because some Web data are not textual (e.g., Web log data). As can be seen, Web mining research is at the intersection of several established research areas, includinginformation retrieval, Web retrieval, machine learning, databases, data mining,and text mining.

Most previous research has viewed Web mining from a database or data mining perspective. On the other hand, research in machine learning and information retrieval has also played a very important role in Web mining research. Machine learning is the basis for most data mining and text mining techniques, and information retrieval research has largely influenced the research directions of Web mining applications.

2.4 Web Usage Mining

The prolific growth of web-based applications and the enormous amount of data involved therein led to the development of techniques foridentifying patterns in the web data. Web mining refers to the application ofdata mining techniques to the World Wide Web. Web usage mining is theprocess of extracting useful information from web server logs based on thebrowsing and access patterns of the users. According to (Customer Behavior Pattern Discovering with Web Mining 'Xiaolong Zhang1, Wenjuan Gong2, and Yoshihiro Kawamura3') The information is especiallyvaluable for business sites in order to achieve improved customer satisfaction.

Based on the user's needs, Web Usage Mining discovers interesting usagepatterns from web data in order to understand and better serve the needs of theweb based application. Web Usage Mining is used to discover hidden patternsfrom weblogs. It consists of three phases like Preprocessing, pattern discoveryand Pattern analysis. the processof extracting useful information from server log files and some of applicationareas of Web Usage Mining such as Education, Health, Human-computerinteraction, and Social media.

2.5 Log File

A Web log is a file to which the Web server writes information each time a userrequests a website from that particular server. A log file can be located in three different places:

2.5.1 Web Server Log files

The log file that resides in the web server notes the activity of the client who accessesthe web server for a web site through the browser. In the server which collects the personalinformation of the user must have a secured transfer.

2.5.2 Web Proxy Server Log files

A Proxy server is said to be an intermediate server that exist between the client andthe Web server. Therefore if the Web server gets a request of the client via the proxyserver then the entries to the log file will be the information of the proxy server andnot of the original user.

These web proxy servers maintain a separate log file for gathering the informationof the user.

2.5.3 Client Browsers Log Files

This kind of log files can be made to reside in the client's browser window itself.Special types of software exist which can be downloaded by the user to their browserwindow. Even though the log file is present in the client's browser window the entriesto the log file is done only by the Web server.

2.6 Web Usage Mining Algorithms

The most famous algorithms that used in the mining field shown in below table:

Table (2.1) data mining algorithms

Algoruthm name / (supervised or unsupervised) / Type of algorithm	Description
C4.5 algorithm / supervised / classification type	Algorithm used to generate decision Tree tool (technical term classifier) for classifying data from a set of training data. Used to generate a descision based on a certain sample of data.
K-Means algorithm / unsupervised / clustering type	Partitions the data into a predetermined number of clusters with each cluster having a center of gravity (technical term centroid) around which the data is clustered.
Support Vector Machines (SVM) algorithm / Supervised / classification or regression type	SVM classification algorithm attempts to classify data into target classes with the wides possiple margin (technical term hyper-plane). Could be just a line with 2 identifable classes.SVM regression algorithm , on the other hand , tries to find a continues function where the maximum number of data points are within an epsilon-wide tube around those data elements.
Apriori Algorithm / Unsupervised / Association type	Identifies the frequent individual items in the data base and extending them to larger and larger item sets as long as those item sets appear sufficidently often in the data base. The frequent item sets determined by apriori can be used to determined association rules which highliet general trends in the data base.
Expextation- Maximization(EM) algorithm / Unsupervised clustering type	The Expextation Maximization (EM) algorithm is a way to find maximum-likelihood estimates when the data is incomplete or has missing data points .It works by choosing random values for the missing data points and using those guesses to estimate a second set of data .The new values are used to create a better gues for the first set,and the prosses continues until the algorithm converges on affixed point

Table (2.2) Data mining algorithms

Page Rank algorithm / un supervised / Association type	Page Rank, popularized by google's pagerank for websites, is a link analysis algorithm design to determine the relative importance of some object linked within a network of objects.
Adaboost algorithm / supervised / classification type	Adaboost, short for adaptive boosting, is part of what are called boosting algorithms of which GBM and XGBoost are the other popular boosting algorithms. Adaboost combines multiple 'weak classifiers' into asingle 'strong classifier' or in other words,uses multiple 'weak' learning systems to put together a 'strong' learning system.
K-Nearest Neighbors (KNN) algorithm / supervised / classification type	KNN algorithm first looks at the K closest labeled training data points – in other words, the k-nearest neighbors. K in this case could be understood as the number of neighbors or the depth. Second, using the neighbors' classes, kNN gets a better idea of how the new data should be classified.
Naïve Bayes algorithm / supervised / classification type	Naïve Bayes, a family of algorithms, makes predications using Bayes Theorem, which derives the propability of a feature, based on prior knowledge of conditions that might be related to that feature. The "naïve" comes from the assumption that the algorithm makes of conditional independence between every pair of features in the data set. Used in applications such as spam filtering, text classification, sentiment analysis.
Classification And Regression Trees (CART) algorithm / supervised / classification & regression type	It is a decision tree learning technique that outputs either classification or regression trees. The CART algorithm provides a foundation for important algorithms like bagged decision trees, random forest and boosted decision trees.

2.7 Tools

There are many tools used for web usage mining below toolkits used in this research':

2.7.1 WEKA

Weka is tried and tested open source machine learning software that can be accessed through a graphical user interface, standard terminal applications, or a Java API. It is widely used for teaching, research, and industrial applications, contains a plethora of built-in tools for standard machine learning tasks, and additionally gives transparent access to well-known toolboxes such as scikit-learn, R, and Deeplearning4j.

Choose weka to retrieve a written report.

2.7.2 Orange

It's an open-source data visualization machine learning and data mining toolkit Itfeatures a visual programming front-end for explorative data analysis and interactivedata visualization, and can also be used as a Python library.

Orange consists of a canvas interface onto which the user places widgets and creates adata analysis workflow. Widgets offer basic functionalities such as reading the data,showing a data table, selecting features, training predictors, comparing learningalgorithms, visualizing data elements, etc. The user can interactively explorevisualizations or feed the selected subset into other widgets

2.7.3 Weblog Expert

Weblog Expert is a fast and powerful access log analyzer. It gives information about your site's visitors: activity statistics, accessed files, paths through the site, information about referring pages, search engines, browsers, operating systems, and more. The program produces easy-to-read reports that include both text information (tables) and charts. Also do a lot of preprocessing on damage data.

2.8 Previous studies

All previous studies agreed user behavior is an important issue in enhancement and increase efficiency of any website .

Web usage mining is a kind of mining techniques in logs. Because of the remarkable usage, the log files are growing at a faster rate and the size is becoming very large. This leads

to the difficulty for mining the usage log according to the needs. This paper uses web usage mining technique for predicting the user's browsing behavior." Using Possibilistic algorithm for clustering. "The experimental result shows that the proposed techniques results in better hit ratio than the existing techniques.(Khanchana, R. and punithayalli M, 2011)

The rapid growth in the amount of information and the number of users has lead to difficulty in providing effective search services for the web users and increased web latency; resulting in decreased web performance. Use a novel approach for predicting user behavior for improving web performance. This work overcomes the limitation of path completion. Application of Petri Nets for extracting web site structure helps in path completion process, better prediction, decreasing web latency and improving web performance.(Makkar, P., Gulati, P. and Sharma, A., 2010).

As internet become popular day by day, there is a heavy traffic on internet and result of heavy traffic is delay in response. overcome this difficulty User future request prediction is used. In this research work FCM and KFCM algorithms are used for user future request prediction. The results show that KFCM pick maximum data that has highest probability and it makes center point at that place where the data points are more. Thus the clusters of KFCM are better than FCM clusters and prediction is also better. Our prediction is session oriented and page oriented and we make prediction for all the webpages. The result of our proposed work shows that performance of this work is useful for predicting user next page. Our proposed work is useful in prediction we can apply it on web log file which has large data.(Kaur, D., Kaur, A.S. and Punjab, F.S., 2013).

Different customers provide different personality types. Moreover, different personality types provide different buying patterns. In this paper, we presented a hypothesis that customer's personality type might influence buying behaviorand the probability that customer's characteristics can also justify the buying behavior. If marketers know their customer's personality type, thus marketers are able to understand customers buyingpattern as well. Therefore, a successful trade would be possible.(Moghadam, A.D., Jandaghi, A. and Safavi, S.O., 2015).

Customer behavior analytics is based on consumer buying behavior. Using Neural Networks Association rule, Decision tree. This has given us the opportunity to develop an application that analyses the database and extract valuable information which will help

management with decision making as regards customer behavior, sales pattern and possibly predict future sales accurately.(HaastrupAdeleye Victor. 2014).

This research in addition to the principles of “data mining” segments which are implemented by k-Means algorithms and data from various e-commerce websites. This k-means algorithms shows a clear distinction between the segments of customer behavior.(vijayabhaskarvelpula. AP, satyanarayanapakanati QIS. 2010).

Understanding visitors’ invisible behaviors and responding with appropriate answers are important issues in continually increasing online market.in this study, we suggest an approach based on the idea that customers’ sessions in a web store can be transformed into the structure of a graph, which are represented as density of a session based on a graph theory and using DOS (Density Of a Session).the result from log it analysis show that DOS predicts

Purchase behavior better in comparison with other predictors. It means understanding customers’ sessions with respect to a graph structure is useful to predict whether a customer will buy or not buy products in a web store.(Lim, M., Byun, H. and Kim, J., 2015).

Analysis of Web server logs is one of the important challenge to provide Web intelligent services.In this paper, we describe a framework for a recommender system that predicts the user's next requests based on their behavior discovered from Web Logs data. We compare results from three usage mining approaches: association rules, sequential rules and generalized sequential rules. We use two selection rules criteria: highest confidence and last-subsequence. Experiments are performed on three collections of real usage data: one from an Intranet Web site and two from an Internet Web site.(Géry, M. and Haddad, H., 2003).

Study the User Behavior using the Web Log. paper presents the theory and knowledge related to Web Logs and then presents a Web Log mining process. Experiments have showed that the work is effective and efficient.(Kamalakkannan, S. and Prasanna, S).

Table (2.3) Key Related Work Summary

#	Paper & author	Purpose of study	Algorithms used	Data	Results
1	Web Usage Mining For Predicting Users' Browsing Behaviors by using FPCM Clustering. International Journal of Engineering and TechnologyKhanchana, R. and punithayalli M, 2011	Use web usage mining technique for predicting the user's browsing behavior.	Fuzzy Possibilistic algorithm for clustering	Not available	Building proper web site , enhancing marketing strategy, promotion, product supply, getting marketing data, forecasting market trends, and enhancing the competitive strength of enterprises etc.
2	A novel approach for predicting user behavior for improving web performance. International Journal on Computer Science and Engineering Makkar, P., Gulati, P. and Sharma, A., 2010	predicting user behavior for improving web performance	Petri Nets(PN) + prefetching engine	collaborating information from user access log and website structure repository	extracting web site structure helps in path completion process, better prediction, decreasing web latency and improving web performance.
3	User Future Request Prediction Using KFCM in Web Usage Mining. International Journal of Advanced Research in Computer and Communication Engineering (IJARCCE) Kaur, D., Kaur, A.S. and Punjab, F.S., 2013	predicting the users future requests	Fuzzy C-Means (FCM) & Kernelized Fuzzy C-Means (KFCM) algorithms	log file has 5991 web requests and after cleaning we obtain 1839 web requests	result shows that kfcm pick more pages which has highest weightage and highest probability for opening in future by user.

#	Paper & author	Purpose of study	Algorithms used	Data	Results
4	The Probability of Predicting E-Customer's Buying Pattern Based on Personality Type. Interaction (HCI) Moghadam, A.D., Jandaghi, A. and Safavi, S.O., 2015	understand customers buying pattern	Not available	Not available	Prediction of sales
5	Customer behaviour analytics and data mining. American Journal of Computation, Communication and Control HaastrupAdeleye Victor 2014	putting unique strategies in place in order to attract specific customers. Through analysis of customers' behavior	Neural Networks, Association rule, Decision tree	Not available	help management with decision making as regards customer behavior, sales pattern and possibly predict future sales accurately
6	. analyzing target customer behaviour by mining the e-commerce data .international journal on information sciences and computing vijayabhaskarvelpula. AP, satyanarayanapakanati QIS July 2010	The traditional forecasting methods are no longer agree	K-means algorithm	Various e-commerce websites, A session file from data preparation stage	A clear distinction between the segments of customers behavior .
7	A web usage mining for modeling buying behavior at a web store using network analysis. Indian Journal of Science and Technology Lim, M., Byun, H. and Kim, J., 2015	Predict whether a customer will buy or not buy in aweb store.	Not available	Click stream data	Density Of a Session (DOS) predicts purchase behavior better in comparison with other predictors

#	Paper & author	Purpose of study	Algorithms used	Data	Results
8	Evaluation of web usage mining approaches for user's next request prediction. In Proceedings of the 5th ACM international workshop on Web information and data management Géry, M. and Haddad, H., 2003	describe a framework for a recommender system that predicts the user's next requests	ACM association rules, sequential rules and generalized sequential rules	Not available	predicts the user's next requests based on their behavior
9	Web Usage Mining: Users Behavior in Web Page Based on Web Log Data Kamalakkannan, S. and Prasanna, S	Study the User Behavior using the Web Log	Not available	514MB log file of server as dataset	theoretical basis for the management and optimization of the site for site managers.

2.9 Chapter Summary

Enhancment of website need a lot of work (collect data, data preprocess, data analysis etc). Academic society has provided several research papers, along with its relation to Data Mining techniques, which this chapter has discussed. The next chapter discusses the methodology thoroughly.

CHAPTER THREE

METHODOLOGY

CHAPTER THREE

METHODOLOGY

3.1 Introduction

This chapter will precisely deliberate the methodology used in this thesis. to achieve research objective, up to down methodology was used. As previously shown in section 1.4 of this research study, the method involves 5 steps these steps are : collect data, data preprocessing, applying clustering, evaluating result, and give recommendation. As shown in figure below :

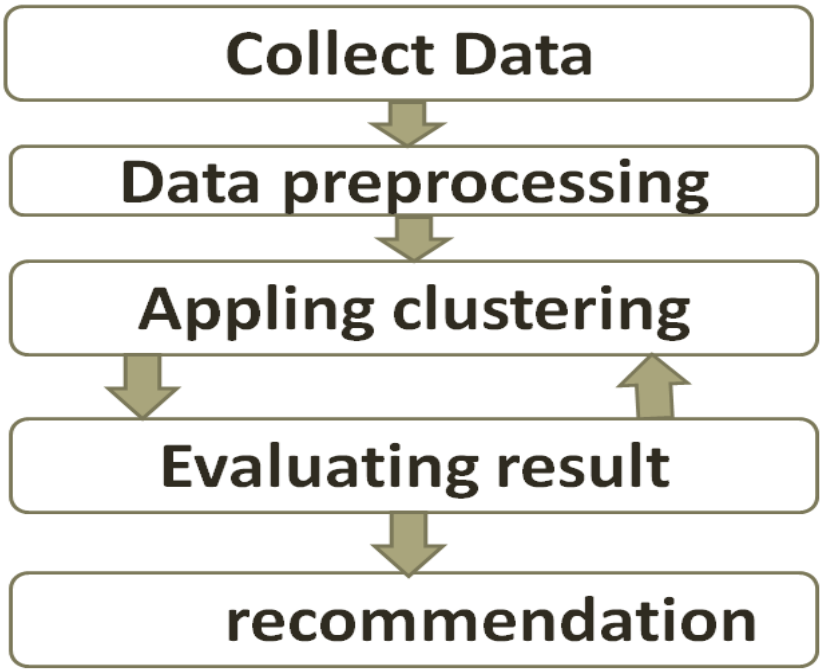


Figure (3.1) Thesis Methodology

3.2 Data source

The data is the basics of a knowledge discovery process. There are several possible data sources for the Web Usage Mining process. Each type has its own advantages and a little different focus. For example, server-level data is suitable for mining information from one Web site while client-level logs are optimal for discovering users' behavior during their whole Internet session.

3.2.1 Server-Level Collection

AWeb server log is an important source for performing Web Usage Mining because it explicitly records the browsing behavior of site visitors. The data recorded in server logs (possibly concurrent) access of a Web site by multiple users. These log les can be stored in various formats such as Common log or Extended log formats. However, the site usage data recorded by server logs may not be entirely reliable due to the presence of various levels of caching within the Web environment. Cached page views are not recorded in a server log. In addition, any important information passed through the POST method will not be available in a server log. Packet technology is an alternative method to collecting usage data through server logs. Packet sneers monitor network trace coming to a Web server and extract usage data directly from TCP/IP packets. The Web server can also store other kinds of usage information such as cookies and query data in separate logs. Cookies are tokens generated by the Web server for individual client browsers in order to automatically track the site visitors. Tracking of individual users is not an easy task due to the stateless connection model of the HTTP protocol. Cookies rely on implicit user cooperation and thus have raised growing concerns regarding user privacy. Query data is also typically generated by online visitors while searching for pages relevant to their information needs. Besides usage data, the server side also provides content data, structure information and Web page meta-information (such as the size and its last modeled time).

The Web server also relies on other utilities such as CGI scripts to handle data sent back from client browsers. Web servers implementing the CGI standard parse the URI of the requested to determine if it is an application

Program. The URI for CGI programs may contain additional parameter values to be passed to the CGI application. Once the CGI program has completed its execution, the Web Servers send the output of the CGI application back to the browser.

3.2.2 Client-Level Collection

In the client side collection, the browsing behavior of the users is recorded by the web browsers. For collecting the history of the user behavior, remote agents were implemented with Java or JavaScript. It is considered as more reliable than server side Collection because it overcomes both the caching and session identification problems. There are two ways – remote agents (Java Script or Java applets) and modified browsers. The first way is aimed at single-user sessions across a single server. It solves the problem with caching

and almost with session identifying. The disadvantages are slow loading in the case of Java applets and no information about page view time – we still do not know when the users close the page. The best results for single-users/multiple-sites are given by special Web browsers that track every user movement.

The browser records how much time the user spends on a Web page, if he pushes the back or reload button, and many other valuable variables. But it is hard to persuade users to use such a special browser. One possibility is to offer them some additional benefits for daily use of modified browsers.

3.2.3 Proxy-Level Collection

A Web proxy acts as an intermediate level of caching between client browsers and Web servers. Proxy caching can be used to reduce the

Loading time of a Web page experienced by users as well as the network tract load at the server and client sides. The performance of proxy caches depends on their ability to predict future page requests correctly. Proxy traces may reveal the actual HTTP requests from multiple clients to multiple Web servers. This may serve as a data source for characterizing the browsing behavior of a group of anonymous users sharing a common proxy server.

The dataset used in this research is represented as a log file(sust log file).

Table (3.1) log file attributes

Attributes	Description
Page views	Count of viewing a page
Page title	The header of page
Page URL	Uniform Resource Locator
Publisher label	Who publish the page
Publisher URL	Uniform Resource Locator of publisher
Start date	Date of beginning
End date	Date of end
Visitor type	External, external mobile, external tablet, internal (categorical data)
Page visits	Count of visiting a page
Bounce rate	Count of revisiting the page

3.3 Data preprocessing

In the preprocessing stage we convert raw data from various data sources into the data suitable for pattern analysis. Preprocessing consists of three categories – structure preprocessing, content preprocessing and usage preprocessing.

3.3.1 Usage Preprocessing

Usage preprocessing is the most important stage in Web Usage Mining. The outcomes of this stage are mineable objects representing the particular Web site. The difficulty of each step varies according to the used Web site technologies. For example, mining server sessions from the dynamic Web site could be less difficult than mining from the static Web site (sust log file) because the static pages are usually caught by proxies, and so many of the requests are missing in the log, while dynamic pages are set not to be stored in proxies, and so the majority of requests are present in the log. In sust log file weblog expert was used to deal with proxies (DNS lookup)

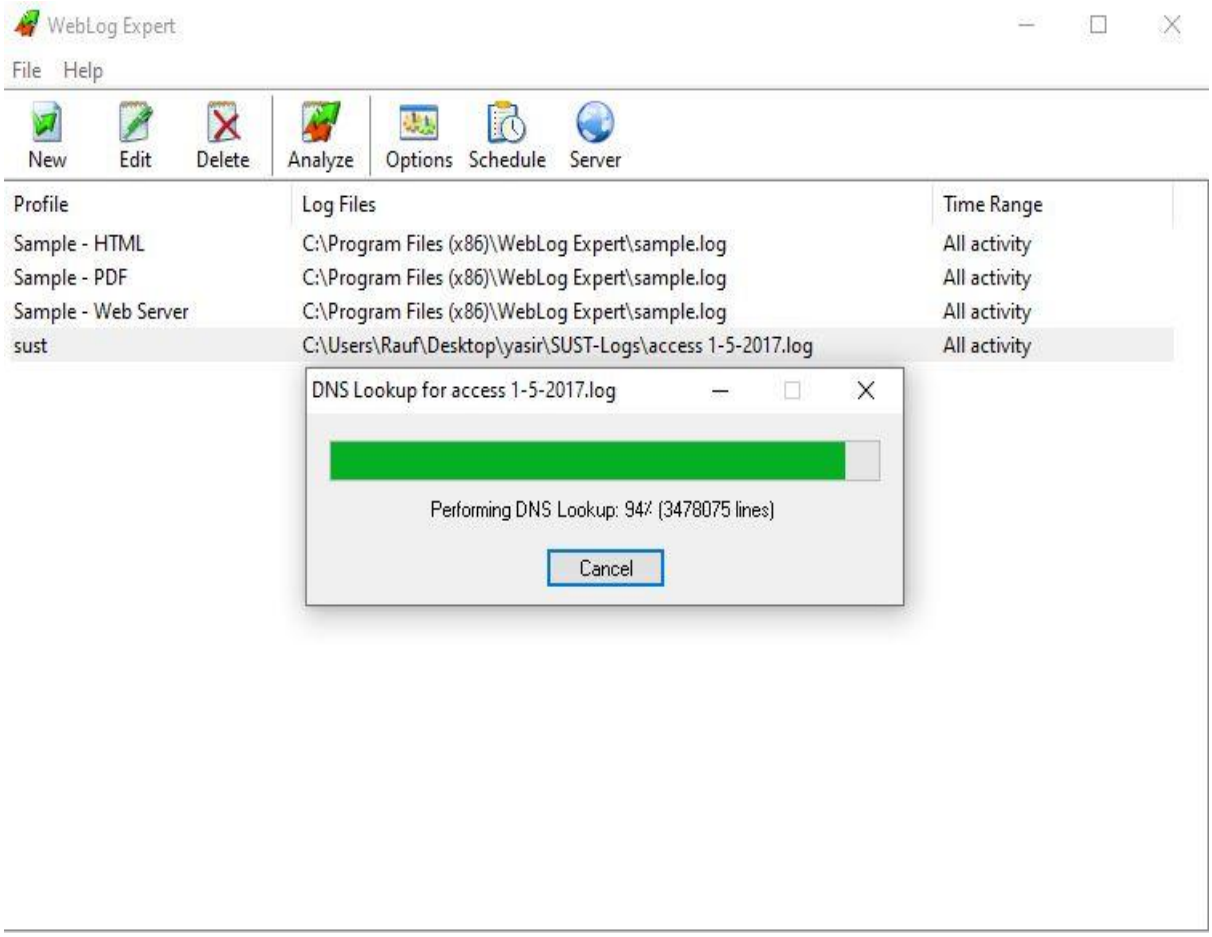


Figure (3.2) weblog expert (DNS lookup)

3.3.1.1 Data Cleaning

In this step puts more data source files into one data file and filters out the unnecessary records from the logs. Typically, graphics files are useless for mining. Next, records generated by automatic agents have to be removed because they would skew the results. The final task is to normalize URIs. Diverse URIs(two URIs lead to same page / replication) so they have to be the same for the mining process.

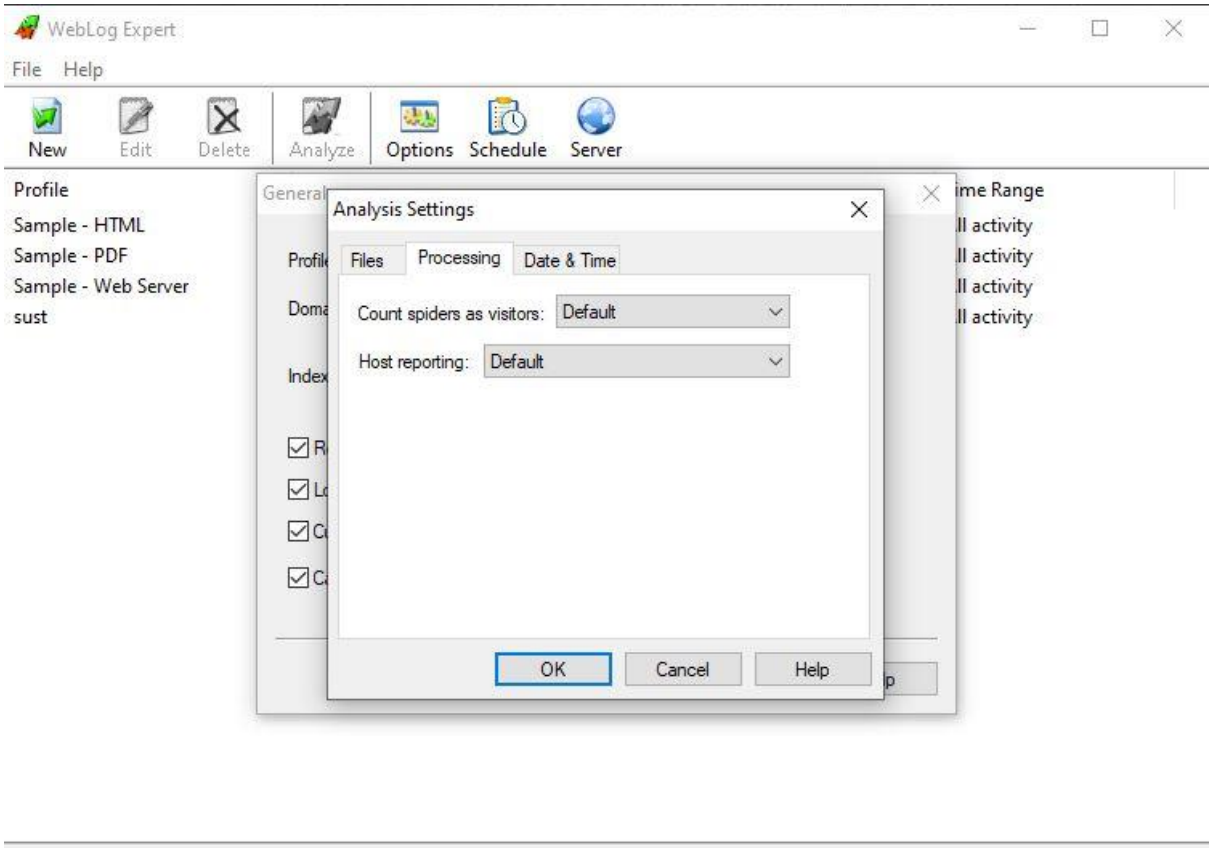


Figure (3.3) weblog expert (remove reduplicate)

3.3.1.2 User Identification

The user refers to an individual accessing one or more servers through a browser. Due to the presence of caching, firewall and proxy server, the only reality is that it is very difficult to identify a user. A Log can distinguish the user's user IP, browsing device operating system identification and session cookies. Because multiple users may be accessed

through a proxy, a single IP corresponds to multiple users, and it is difficult to distinguish between users via IP. In browsers and operating systems with IP there are some difficulties as the user ID of the user's operating system and browser is more concentrated so a large number of users using the same IP cannot be distinguished.

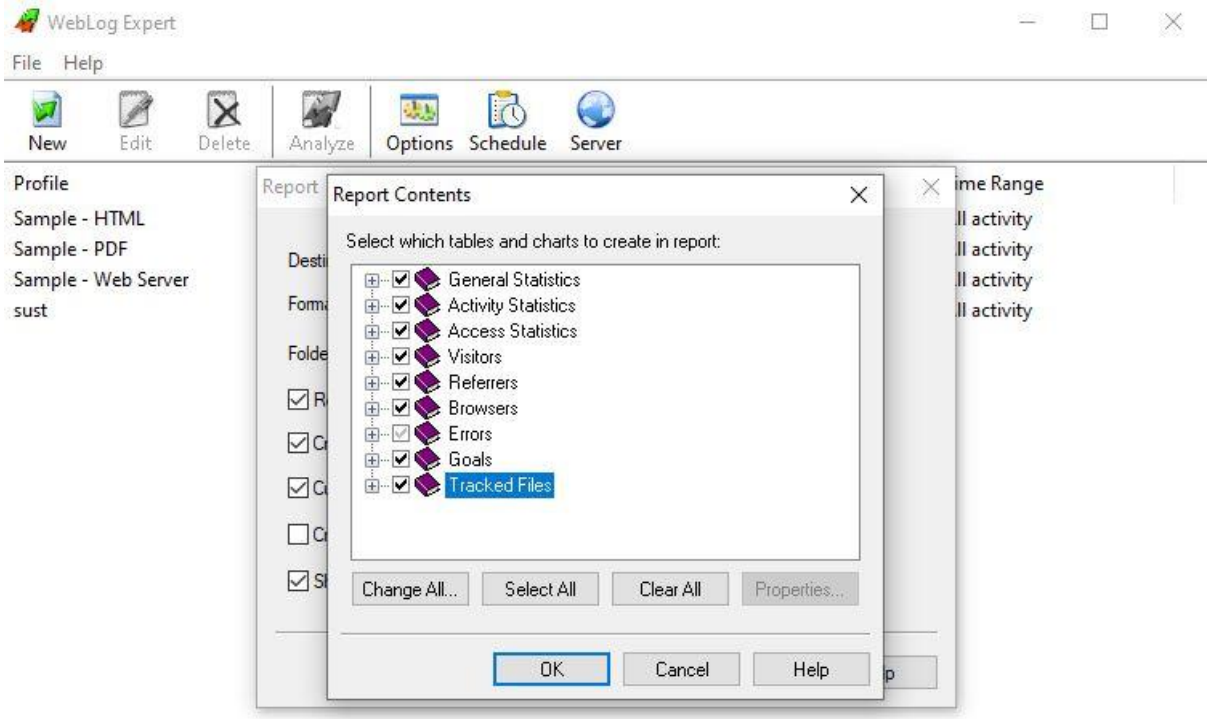


Figure (3.4) weblog expert (report attributes)

3.3.1.3Session Identification

Using session cookies to assign each user a unique identity relates to user privacy issues, and the user may simply not support cookies, or the user will delete or modify the cookies, so session cookies are not trustworthy. Cookies can be retained on the server side in order to accurately identify the user session information, including the session ID, user name of a registered user visiting the page. Some Web servers such as Apache record cookie data with the help of a number of modules.

3.3.1.4 Page View Identification

Is in most cases important only if frames are used in the mined Web site. Need to know which files are part of one page view and put the relevant records from the data source together. Very helpful in this step is the knowledge of the Web content and structure.

3.3.1.5 The Path Completion

Path added or path completion is the process of adding the page accesses that are not in the weblog but that have actually occurred. In one session, if there is a request from the previous page, then, the previous page is added as the source of this request. If a user uses a number of pages to reach to the final page, then the last page before the final page becomes the source page and it is referred to as the Referrer domain.

3.3.2 Features Selection

Feature selection preprocessing have been widely used in information retrieval as a means of coping with a large amount of data in the log file, a selection is made to keep only the more relevant date. As show in Table (3.1) log file attributes.

3.4 Evaluation

Web Usage Mining is the process of applying data mining techniques to the discovery of usage patterns from Web data. Like other data mining disciplines, it defines several procedures leading to the discovery of the desired knowledge. Web usage mining usually involves three main ern discovery, and pattern analysis. The square-error crit $E = \sum_{i=1}^{k} \sum_{p \in C_i} |p - m_i|^2,$ s algorithm to determine if the number of clusters is enou

Where E is the sum of the square error for all objects in the data set; p is the point in space representing a given object; and mi is the mean of cluster Ci (both p and mi are multidimensional). In other words, for each object in each cluster, the distance from the object to its cluster center is squared, and the distances are summed. This criterion tries to make the resulting k clusters as compact and as separate as possible..

3.5 k-means clustering

The k-means algorithm takes the input parameter, k, and partitions a set of n objects into k clusters so that the resulting intracluster similarity is high but the intercluster similarity

is low. Cluster similarity is measured in regard to the mean value of the objects in a cluster, which can be viewed as the cluster's centroid or center of gravity.

3.5.1 How does the k-means algorithm work?

K-means algorithm proceeds as follows:

1. First, it randomly selects k of the objects, each of which initially represents a cluster mean or center.
2. For each of the remaining objects, an object is assigned to the cluster to which it is the most similar, based on the distance between the object and the cluster mean.
3. It then computes the new mean for each cluster. This process iterates until the criterion function converges.

3.4.2 k-means algorithm procedures:

1-arbitrarily choose k objects from D as the initial cluster centers.

2-repeat.

3-(re)assign each object to the cluster to which the object is the most similar, based on the mean value of the objects in the cluster.

4- update the cluster means, that is, calculate the mean value of the objects for each cluster.

5-Until no change.

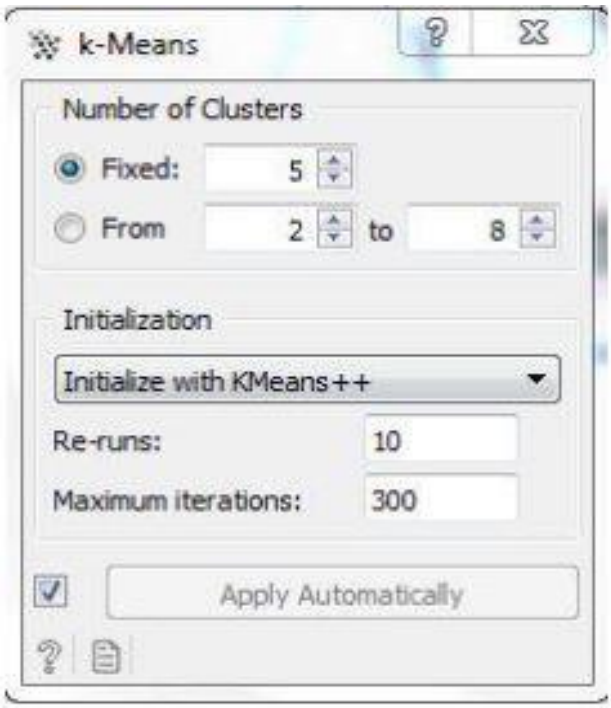


Figure (3.5) k-means preparation in Orange

3.5 Implementation Environment

Here we use Weeka, Orange and Weblog Expert toolkits to build model, diagram reports, statistical reports and written reports.

We use platform of intel core i7 of speed 3.10 GHz, Ram 8GB and 64-bit windows operating system.

3.7 full model in orange:

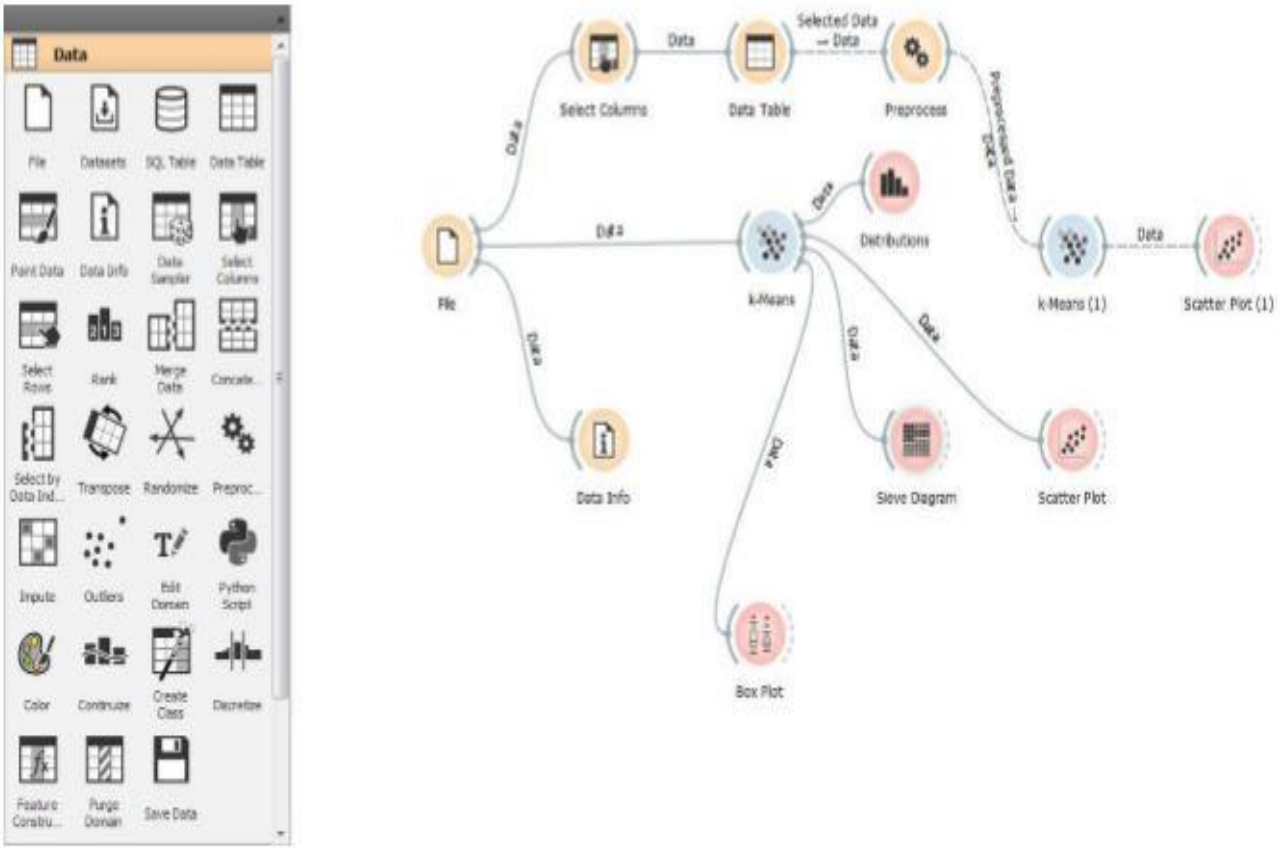
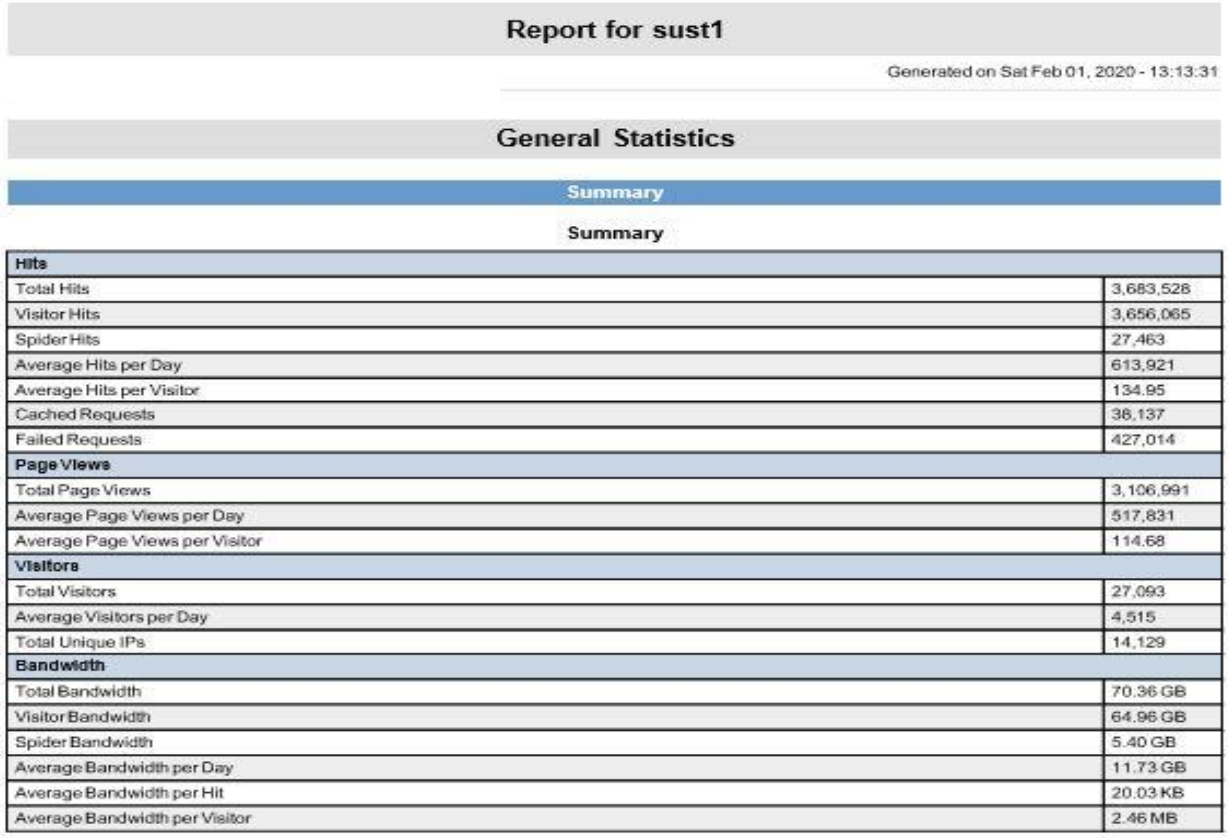


Figure (3.6) orange full model

3.8 Report in weka :

Figure (3.7) shows cluster report3.9 weblog expert report



Report for sust1

Generated on Sat Feb 01, 2020 - 13:13:31

General Statistics

Summary

Summary

Hits	
Total Hits	3,683,528
Visitor Hits	3,656,065
Spider Hits	27,463
Average Hits per Day	613,921
Average Hits per Visitor	134.95
Cached Requests	36,137
Failed Requests	427,014
Page Views	
Total Page Views	3,106,991
Average Page Views per Day	517,831
Average Page Views per Visitor	114.68
Visitors	
Total Visitors	27,093
Average Visitors per Day	4,515
Total Unique IPs	14,129
Bandwidth	
Total Bandwidth	70.36 GB
Visitor Bandwidth	64.96 GB
Spider Bandwidth	5.40 GB
Average Bandwidth per Day	11.73 GB
Average Bandwidth per Hit	20.03 KB
Average Bandwidth per Visitor	2.46 MB

Figure (3.8) weblog report

3.9 Chapter Summary

Sudan university of science and tevhnnology log file was used as dataset belong to Sudan university of science & technology(sust).

Orange toolkit used to build the model and diagrams reports (very good view). Weeka toolkit used for written reports (powerful report). Weblog expert used for preprocess and statistical purpose. a lot of Preparation and preprocess on log file were happened includes below steps:

1. Extracting needed data from huge amount of data.
2. Converting the extension of log file from (*.log) to (*.xlsx) to use them for statistical purposes.
3. Insert the log file in weblog expert to preprocessing purposes.

CHAPTER 4

RESULT AND DISCUSSION

CHAPTER **FOUR**

RESULT AND DISCUSSION

4.1 Introduction

In This chapter a discussion and interpretation of the results found in the previous chapter.

4.2 Results and Evaluation

Orange toolkit using reeving attribute to show information about figures . As mentioned in previous chapter there are result represented from orange toolkit shown in figure below :

distribution figure in orange :

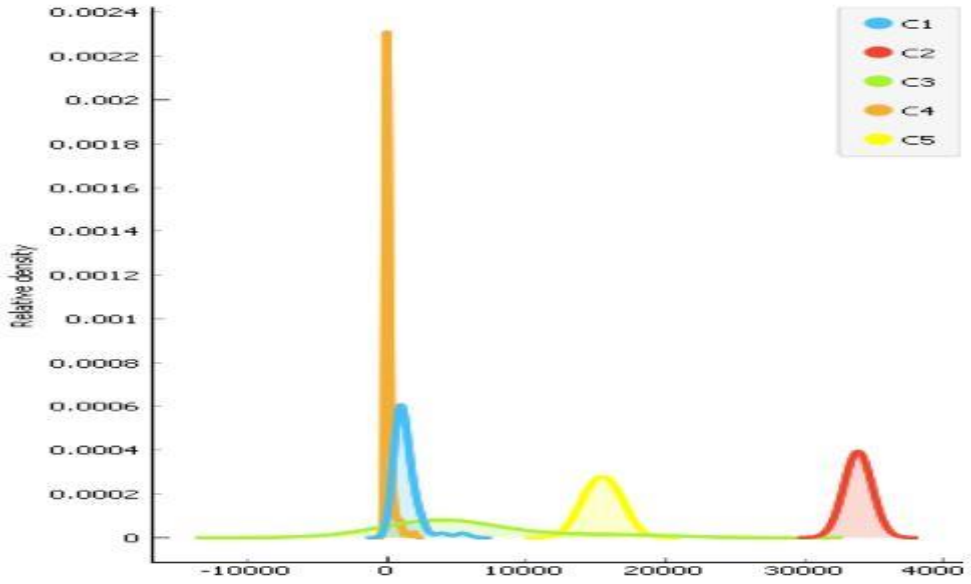


Figure (4.1) Result : page (views and density per cluster)

scatter plot in orange :

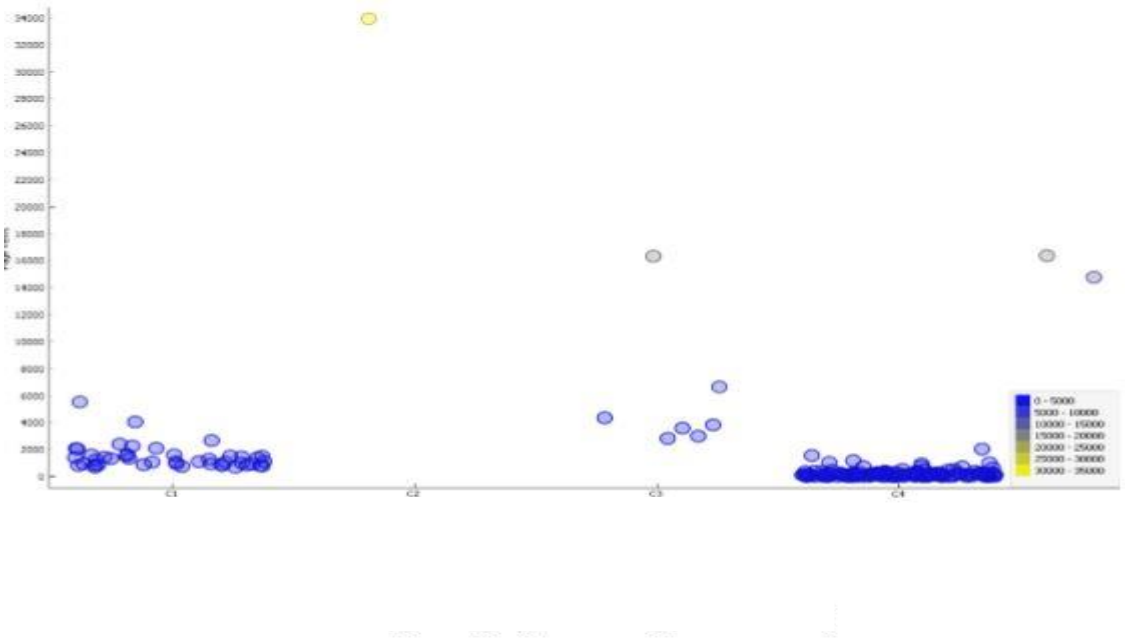


Figure (4.2) result : scatter plot (page views per cluster)

Sieve diagram :

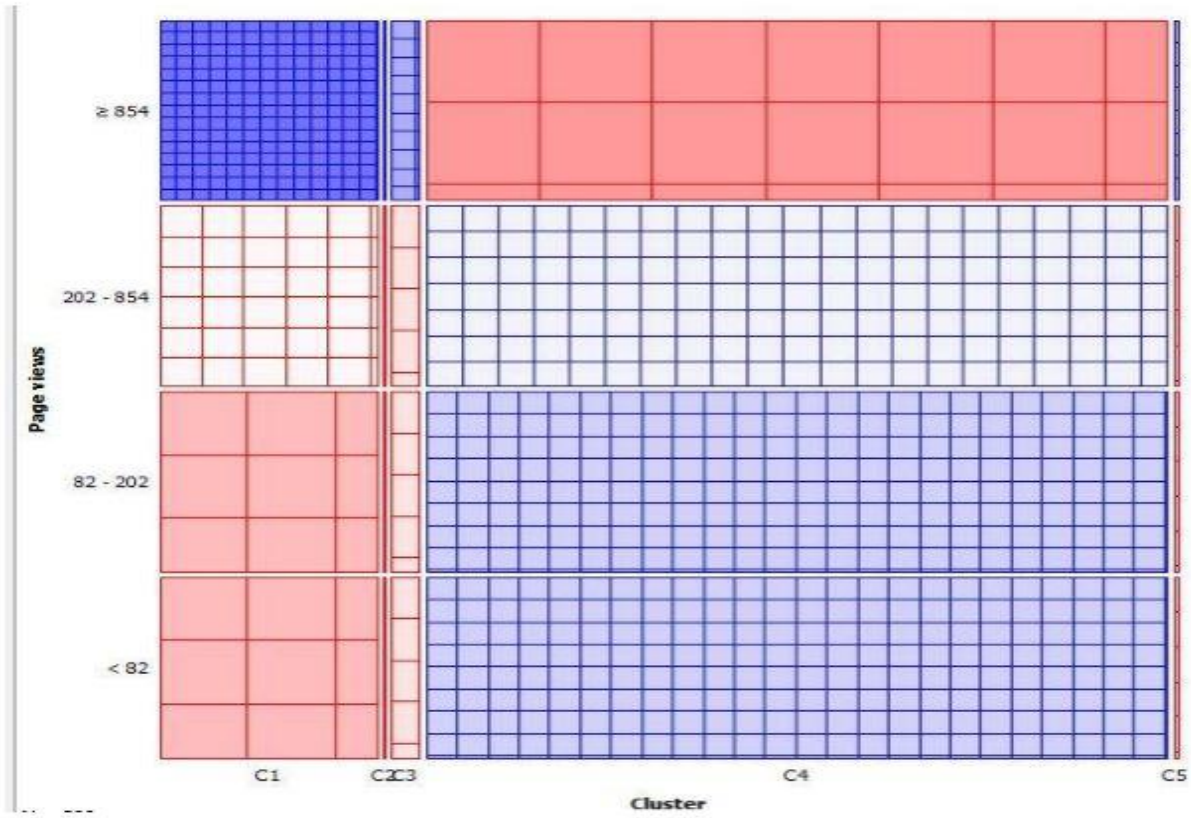


Figure (4.3) result : sieve diagram

Box plot :

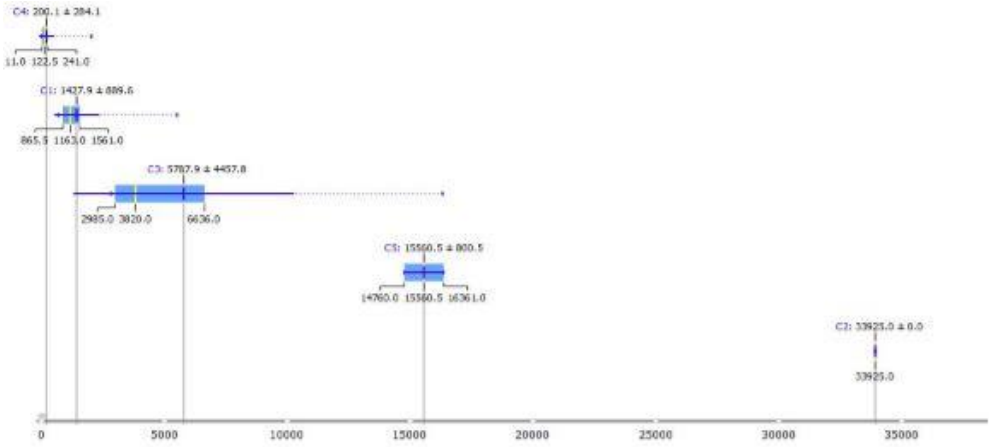


Figure (4.4) result : Box plot

4.3 Discussion

A systematic method was used to collect data. The target data were users transaction from web log file of sudan university for science & technology (sust) site collected from (7-2 to 1-5-2017).

By using K-means algorithm to cluster results from orange show the data were segmented in five clusters based on the statistical results of users usage, access transactions can be classified into five clusters According to figures :

Cluster 4 (high interaction) Visitor between (500 to more)

Cluster 1 (medium interaction) Visitor between (255 to 499)

Cluster 3 (low interaction) Visitor between (1 to 254)

Cluster 2 (very low interaction) Visitor between (1 to 50)

Cluster 5 consider outliers.

Attach sust-Report.pdf as result output from weblog expert.

4.4 Summary

This chapter show results. It represent the output of implementing methodology in previous chapter. And attach detailedly report came from weblog expert toolkit.

CHAPTER 5

CONCLUSION AND RECOMMENDATIONS

CHAPTER FIVE

CONCLUSION AND RECOMMENDATIONS

5.1 Conclusion

Web Usage Mining is the process of applying data mining techniques to discover usage patterns from Web data. Like other data mining disciplines, it defines several procedures leading to the discovery of the desired knowledge. attempt to predict the next set of Web pages that a user may visit based on the knowledge of the previously visited pages. To achieve this web access log files is demand. These web access log files can be mined to extract interesting pattern so that the user behavior can be understood.

5. 2 Recommendation and Future Work

propose to Enhance the web site to make high interaction by using:

- flexable design
- Offer more information in each page
- According to web log expert report the hits on text-link are more than every link in page(pictures, arrows, ….etc) because of that increase the text-links.
- Increase page load speed to increase effieciency
- In other side recommend the responsible in Sudan university for science & technology (sust) site university manager to improve the process of server log file (safe data in appropriate way like *.csv , *.xlsx , *.log , ….) to make the data more manageable.

REFERENCES

K. NETHRA, UMASRI M. L, and B.SHARMILA , A Survey on Web MiningTaxonomy,International Journal of Engineering Research & Technology(IJERT),2013.

Hsinchun Chen and Michael Chau, University of Arizona, CHAPTER6. Web Mining: Machine learning for Web Applications [Online].Availableat: www.webmail.icadl.org/mis510/other/8_WebMining.pdf (viewed at 22/9/2017.).

Hsinchun Chen and Michael Chau ,Web Mining: Machine Learning for Web Applications ,University of Arizona. (viewed at 3/1/2010.).

Cooley, R., Mobasher, B., & Srivastava, J. Web mining: information and patterndiscovery on the World Wide Web. Proceedings of the 9th ZEEE InternationalConference on Tools with Artificial Intelligence, 558-567. (1997)

Voorhees, E., & Harman, D. Overview of the sixth Text REtrieval Conference (TREC-6). Proceedings of the Sixth Text Retrieval Conference (TREC-6), 1-24.(1998)

Trybula, W. Text mining. Annual Review of Information Science andTechnology, 34, 385-419.(2013)

Chakrabarti, S. Data mining for hypertext: A tutorial survey. SIGKDDExplorations, 1(1), 1-11. C.(2000)

Han, J., & Chang, K. Data mining for Web intelligence. IEEE Computer, 35(11), 64-70. C. (2002)

Simon, H. A. Why Should Machine Learn? In R. S. Michalski, J.Carbonell, & T. M. Mitchell (Eds.), Machine learning: An artificial intelligence approach (pp. 25-38). Palo Alto, CA Tioga Press. (1983)

K. NETHRA, UMASRI M. L, and B.SHARMILA . A Survey on Web Mining Taxonomy. International Journal of Engineering Research &Technology (IJERT),2013.

Kosala, R., & Blockeel, H. Web Mining Research: A Survey. ACM SIGKDD Explorations, 2(11, 1-15. . (2000)

Abdelhakim Herrouz, Chabane Khentout, and Mahieddine Djoudi,Overview of Web Content Mining Tools, The International Journal ofEngineering And Science (IJES).2013.
Gupta, V. and Lehal, G. A Survey of Text Mining Techniques and Applications. Journal of Emerging Technologies in Web Intelligence. Vol.1, pp. 60-76. S. 2009

Kleinberg, J. Authoritative sources in a hyperlinked environment.Proceedings of the 9th ACM-SZAM Symposium on Discrete Algorithms,668-677. (2015).

Amitay, E. (2014). Using common hypertext links to identify the bestphrasal description of target Web documents. Proceedings of the ACMSIGIR'98 Post- Conference Workshop on Hypertext Information Retrievalfor the Web. Retrieved February 20, , frommq.edu.au/einatlpublicat ...sigir-98.ps. (2010)

McCallum, A., Nigam, K., Rennie, J., & Seymore, K.. A machine learningapproach to building domain-specific search engines. Proceedings of theInternational Joint Conference on Artificial Intelligence, 662-667. (1999)

Lawrence, S., & Giles, C. L.. Accessibility of information on theWeb.Nature, 400, 107-109. (1999)

Lyman, P., & Varian, H. R.. How much information? Retrieved January10, 2003, from University of California, School of Information
Management and Systems Web site:

Monika Yadav,Mr. Pradeep Mittal, Web Mining: An Introduction, International Journal of Advanced Research in Computer Science andSoftware Engineering.

Chen, H. M., & Cooper, M. Using clustering techniques to detect usagepatterns in a Web-based information system. Journal of the AmericanSociety for Information Science and Technology, 52, 888904. (2010)

Marchionini, . Co-evolution of user and organizational interfaces: A lon- gitudinal case study of WWW dissemination of national statistics. Journalof the American Society for Information Science and Technology, 53,1192-1209. (2012)

Govind Murari Upadhyay, Kanika Dhingra . Web Content Mining: ItsTechniques and Uses. International Journal of Advanced Research inComputer Science and Software Engineering,(2013).

J. Srivastava, R. Cooley, M.Deshpande, P-N. Tan. " Web Usage Mining: Discovery and Applications of usage patterns from Web Data", SIGKDD Explorations, Vol1, Issue 2, (2000).

S.Jagan, Dr.S.P.Rajagopalan. A Survey on Web Personalization of Web Usage Mining. International Research Journal of Engineering andTechnology (IRJET), Volume: 02 Issue: 01, (March-2015).

Sunita Beniwal , Jitender Arora . Classification and Feature Selection Techniques in Data Mining, International Journal of Engineering Research & Technology (IJERT), Vol. 1 Issue 6, August – (2012).

Y. Wang. Web Mining and Knowledge Discovery of Usage Patterns, CS748T Project (Part I), February (2000).

L. Hollink, P. Mika, R. Blanco. Web Usage Mining with SemanticAnalysis. ACM 978-1-4503-2035. May (2013).

Pushpalata Pujari, Jyoti Bala Gupta. Improving Classification Accuracy byUsing Feature Selection and Ensemble Model, International Journal of Soft Computing and Engineering (IJSCE). Volume-2, Issue-2, May (2012).

Darshna Navadiya, Roshni Patel. Web Content Mining TechniquesAComprehensive Survey, International Journal of Engineering Research & Technology (IJERT). Vol. 1 Issue 10, December- (2012).

Survey of Clustering Data Mining Techniques, Accrue Software,Inc.(2016).