###### Emerging Pattern-Based Clustering of Web Users Utilizing a Simple Page-Linked Graph

###### A PROJECT REPORT (Project Phase - I)

###### SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE AWARD OF BACHELOR OF TECHNOLOGY

###### IN

###### COMPUTER SCIENCE AND ENGINEERING

###### BY

###### K. RADHIKA 2210313732

###### J. SRI CHARAN 2210313725

###### M. NEERAJ 2210313737

###### B. RAHUL 2210313209

###### 

**UNDER THE GUIDANCE OF**

**Ms. G. RATHNAMMA**

**Asst.Professor**

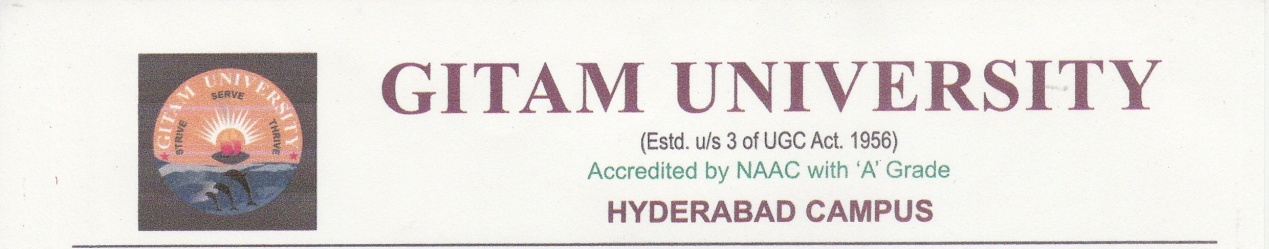
****

**DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING**

**SCHOOL OF TECHNOLOGY**

**GITAM UNIVERSITY**

**HYDERABAD**



**DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING**

**CERTIFICATE**

###### This is to certify that the project titled Emerging Pattern-Based Clustering of Web Users Utilizing a Simple Page-Linked Graph was presented satisfactorily at the Department of Computer Science and Engineering, GITAM School of Technology, GITAM University Hyderabad campus by Radhika, J. Sri Charan, N. Neeraj, B. Rahul in partial fulfillment of requirement for their project work phase-I carried out under our guidance and supervision.

**MS.G. RATHNAMMA**

Rajendra Prasad Babu Dr.S. PhaniKumar

Project Coordinator HOD-CSE

ACKNOLEDGEMENT

The satisfaction that accompanies that the successful completion of any task would be incomplete without the mention of people whose ceaseless cooperation made it possible, whose constant guidance and encouragement crown all efforts with success.

We are grateful to our project guide Ms. Rathnamma for the guidance, inspiration and constructive suggestions that helped us in the preparation of this project.

We thank our project coordinator Mr. Rajendra Prasad, Assistant Professor for his support and encouragement throughout the course of the project work.

At the outset, we thank our Head of the Department Dr. S. Phani Kumar and our Honorable Pro-Vice Chancellor Prof. N. Siva Prasad for the moral support and the excellent facilities provided. I would also like to thank all the teaching and non-teaching staff members of Computer Science department who have extended their full cooperation during course of my industrial training.

We wish to express our warm and grateful thanks to our project coordinator for the guidance and assistance he provided in completing our project successfully.

I thank all my friends who helped me sharing knowledge and by providing material to complete the project in time.

ABSTRACT

**Emerging patterns** are sets of items whose frequency changes significantly from one dataset to another. They are useful as a means of discovering distinctions inherently present amongst a collection datasets and have been shown to be a powerful method for constructing accurate classifiers. Internet has a wide use of data now-a-days. Clustering of webpages in structured manner helps in the efficient usage of internet. This can be possible by the implementation of clustering of webpages by using simple page linked graph (SPLG). In SPLG, the web pages are denoted as the nodes and the edges are drawn if two pages are accessed in one session.

This project includes an efficient way of generating large web pages and figuring out the emerging patterns in them. A user’s favorite web pages are found out and clustered them per the generated patterns. Then the labels are clustered by using Term Frequency-Inverse Document Frequency (TF-IDF). An approach is also used for the analyzing and clearing of data.

INDEX

1. INTRODUCTION 1

1.1 MOTIVATION...................................................................1

* 1. PROBLEM DEFINATION............................................................................1
  2. OBJECTIVE OF PROJECT........................................................1

2. LITERATURE SURVEY.................................................................................2

* 1. INTRODUCTION……………2

2.2 EXISTING SYSTEM 2

2.3 DISADVANTAGES OF EXISTING SYSTEM2

2.4 PROPOSED SYSTEM...................................................................................2

3. ANALYSIS......................................................................3

3.1INTRODUCTION.............................................................................3

* 1. SOFTWARE REQUIREMENT SPECIFICATION.................................3
     1. USER REQUIRMENT
     2. SOFTWARE REQUIRMENT
     3. HARDWARE REQUIRMENT
  2. CONTENT DIAGRAM OF PROJECT......................................................3

3.4 ALGORITHMS AND FLOWCHARTS...................................................3

4. DESIGN......................................................................................................................5

4.1INTRODUCTION......................................................................5

* 1. DFD DIAGRAM ..............................................................................5

4.3 MODULE DESIGN AND ORGANIZATION..........................5

4.3.1PREPROCESSING OF DATASET........................................5

4.3.2 GETTING LARGE WEB PAGES.................................6

4.3.3 GENERATING SPLG’S...............................................6

4.3.4EMERGING PATTERN MINING.........................6

4.3.5WEB USER CLUSTERING.......................................................6

4.3.6 CLUSTERS ANNOTATION..........................................7

5. IMPLEMENTATION AND RESULTS.................................................7

5.1 INTRODUCTION

5.2 EXPLANATION OF KEY FUNCTIONS

5.3 METHODS OF IMPLEMENTATION

5.3.1 FORMS

5.3.2 OUTPUT SCREENS

5.3.3 RESULT ANALYSIS

6. CONCLUSION

LIST OF FIGURES

1. Impleme.......................................................................................6

LIST OF ABBREVATIONS

1.SPLG – SIMPLE PAGE LINKED GRAPH

2.TF-IDF- TERM FREQUENCY-INVERSE DOCUMENT FREQUECNCY

3.LWP- LARGE WEB PAGES

4.WUM-WEB USA

1. INTRODUCTION
   1. MOTIVATION

Discovery of powerful distinguishing features between datasets is an important objective in data mining. Data Mining (sometimes called data or knowledge discovery) is the process of analyzing data from different perspectives and summarizing it into useful information. It allows users to analyze data from many different dimensions or angles, categorize it, and summarize the relationships identified. Technically, data mining is the process of finding correlations or patterns among dozens of fields in large relational databases.

An important class of patterns that can represent strong contrasts is known as emerging patterns. Emerging patterns are sets of items whose frequency changes significantly from one dataset to another. They are useful as a means of discovering distinctions inherently present amongst a collection datasets and have been shown to be a powerful method for constructing accurate classifiers. This application when applied to the large usage of web mining, the results are very accurate and useful. Web mining is the use of data mining techniques to automatically discover and extract information from Web documents and services.

In the current era, data is growing rapidly and there is lot of information to be gathered and stored. The main motto of the project is to take a web log file of some certain users, then clean the data. If a user visits some pages frequently, those are to be identified and support count is calculated. Based on the support count, there is a graph generated. From the graph the frequent pages are identified and clustered. By clustering them the user can access the pages in an easy way. Naming the clustering will be the final step of this project.

* 1. PROBLEM DEFINATION:

In the present day, there is a rapid growth of internet, studies on internet revealed some interesting topics such as social networks, web mining etc., focusing on web mining, Web mining is the integration of information gathered by traditional data mining methodologies and techniques with information gathered over the World Wide Web(WWW). It is further classified into 3 types web content mining (WCU), web structure mining (WSU), web usage mining (WMU).

Web content mining is used to examine data collected by search engines and Web spiders. Web structure mining is used to examine data related to the structure of a particular Web site and web usage mining is used to examine data related to a particular user's browser as well as data gathered by forms the user may have submitted during Web transactions. This is also called as web access. The information gathered through Web mining is evaluated by using traditional data mining parameters such as clustering and classification, association, and examination of sequential patterns. Due to the rapid growth of internet, there is a huge data generation and most of it is in hidden form. It is essential to learn about the favorite web pages of web users and to cluster web users in order to understand the structures that they use.

Many techniques in web usage mining have been proposed, and this field is still a hot topic for research in data mining. Most existing web mining techniques are performed based on association rule mining or frequent pattern mining, and these methods aim to find relationships among web pages or predict the behavior of web users.

Association rule mining is a procedure which is meant to find frequent patterns, correlations, associations, or causal structures from data sets found in various kinds of databases such as relational databases, transactional databases, and other forms of data repositories. Now, consider a web users log history. There might be some pages which are randomly visited by him in a frequent way. The access of such pages is to be made easy. This is possible by the concept of emerging patterns and clustering.

* 1. OBJECTIVE OF PROJECT

In the recent years there is a rapid growth of internet, the result of growth lead to the enormous formation of data. The data that is generated has to be stored for future use, to achieve that there are lot of methods and algorithms. One such method is data mining. It analyzes the data from the different perspectives and creates some useful information out of it.

Emerging patterns based clustering using a graph is a application that runs on the web server, and helps the analyzer to analyze the data of some particular users favorite web pages. Here, every user is having specific user ID using which he/she can be uniquely identified. In this project the source is web log file, which consists of the user ID, IP address and the URL. IP address can also be used to identify the user. The URL is the web pages user tried to access. This information is stored in the database. In the next step the database is cleaned. As the data is not perfect all the time, developer has to take care that the data he is using is clean and clear and do not have any kind of ambiguity. Once the cleaning is done the support count is to be identified.

Association rules are created by analyzing data for frequent if/then patterns and using the criteria support and confidence to identify the most important relationships. Support is an indication of how frequently the items appear in the database. Confidence indicates the number of times the if/then statements have been found to be true. So, by calculating the support count frequent patterns are observed. Once the frequent patterns are observed those are to clustered and named. Here each cluster denotes particular user with some set of frequent patterns.

The main motto of the project is to cluster the web pages of the user web log that are frequently visited by Web User. By clustering the web pages, one can understand the structural patterns and their behavior. The work of the user gets easy and he/she can access the web pages easily with just one click. There will be different tools and algorithms used in this project that are detailed in the further sections. Naming of clusters can be can by the Term Frequency- Inverse Document Frequency (TF-IDF).

1. LITERATURE SURVEY
   1. INTRODUCTION

Data mining, the extraction of hidden predictive information from large databases, is a powerful new technology with great potential to help companies focus on the most important information in their data warehouses. Data mining tools predict future trends and behaviors, allowing businesses to make proactive, knowledge-driven decisions. Data mining tools can answer business questions that traditionally were too time consuming to resolve. They scour databases for hidden patterns, finding predictive information that experts may miss because it lies outside their expectations. Most companies already collect and refine massive quantities of data.

Data mining techniques can be implemented rapidly on existing software and hardware platforms to enhance the value of existing information resources, and can be integrated with new products and systems as they are brought on-line. When implemented on high performance client/server or parallel processing computers, data mining tools can analyze massive databases. Data mining is highly useful in the following domains: Market Analysis and Management, Corporate Analysis & Risk Management, Fraud Detection.

Apart from these, data mining can also be used in the areas of production control, customer retention, science exploration, sports, astrology, and Internet Web Surf-Aid. So, this leads to web mining. Web mining describes the application of traditional data mining techniques onto the web resources and has facilitated the further development of these techniques to consider the specific structures of web data. The analyzed web resources contain (1) the actual web site (2) the hyperlinks connecting these sites and (3) the path that online users take on the web to reach a particular site.

Web usage mining then refers to the derivation of useful knowledge from these data inputs. In web mining, there are three categories: web content mining, web structure mining and web usage mining. In Web Usage Mining (WUM), also known as web access, web access pattern tracking can be deﬁned as the web page history; the mining task is a process of extracting interesting patterns from web access logs. Web usage mining is still a popular research area in data mining. With the rapid growth of the Internet, more and more useful information is hidden in web log data. It is essential to learn about the favorite web pages of web users and to cluster web users to understand the structures that they use.

* 1. EXISTING SYSTEM:

Web Usage Mining is the application of data mining techniques to discover interesting usage patterns from Web data in order to understand and better serve the needs of Web-based applications. Usage data captures the identity or origin of Web users along with their browsing behavior at a Web site.

Web usage mining itself can be classified further depending on the kind of usage data considered:

* **Web Server Data**: The user logs are collected by the Web server. Typical data includes IP address, page reference and access time.
* **Application Server Data**: Commercial application servers have significant features to enable e-commerce applications to be built on top of them with little effort. A key feature is the ability to track various kinds of business events and log them in application server logs.
* **Application Level Data**: New kinds of events can be defined in an application, and logging can be turned on for them thus generating histories of these specially defined events. It must be noted, however, that many end applications require a combination of one or more of the techniques applied in the categories above.

Many techniques in web usage mining have been proposed, and this ﬁeld is still a hot topic for research in data mining. Most existing web mining techniques are performed based on association rule mining or frequent pattern mining, and these methods aim to ﬁnd relationships among web pages or predict the behavior of web users.

Association rule mining is a procedure which is meant to find frequent patterns, correlations, associations, or causal structures from data sets found in various kinds of databases such as relational databases, transactional databases, and other forms of data repositories. Given a set of transactions, association rule mining aims to find the rules which enable us to predict the occurrence of a specific item based on the occurrences of the other items in the transaction. The main applications of association rule mining:

* **Basket data analysis** - is to analyze the association of purchased items in a single basket or single purchase as per the examples given above.
* **Cross marketing** - is to work with other businesses that complement your own, not competitors. For example, vehicle dealerships and manufacturers have cross marketing campaigns with oil and gas companies for obvious reasons.
* **Catalog design** - the selection of items in a business’ catalog are often designed to complement each other so that buying one item will lead to buying of another. So these items are often complements or very related.

Pattern mining consists of using/developing data mining algorithms to discover interesting, unexpected and useful patterns in databases. Pattern mining algorithms can be applied on various types of data such as transaction databases, sequence databases, streams, strings, spatial data, graphs, etc.

Pattern mining algorithms can be designed to discover various types of patterns: sub graphs, associations, indirect associations, trends, periodic patterns, sequential rules, lattices, sequential patterns, high-utility patterns, etc. some researchers define an interesting pattern as a pattern that appears frequently in a database. Other researchers want to discover rare patterns, patterns with a high confidence, the top patterns, etc. The most popular algorithm for pattern mining is without a doubt Apriori (1993).

It is designed to be applied on a transaction database to discover patterns in transactions made by customers in stores. But it can also be applied in several other applications. A transaction is defined a set of distinct items (symbols). Apriori takes as input (1) a minsup threshold set by the user and (2) a transaction database containing a set of transactions. Apriori outputs all frequent item sets, i.e. groups of items shared by no less than minsup transactions in the input database.

The Apriori algorithm has given rise to multiple algorithms that address the same problem or variations of this problem such as to (1) incrementally discover frequent item sets and associations , (2) to discover frequent sub graphs from a set of graphs, (3) to discover subsequences common to several sequences, etc.

A sequence database is defined as a set of sequences. A sequence is a list of transactions. For example in the left part of the following figure a sequence database containing four sequences is shown. The first sequence contains item a and b followed by c, followed by f, followed by g, followed by e. A sequential rule has the form X –> Y where X and Y are two distinct non empty sets of items. The meaning of a rule is that if the items of X appear in a sequence in any order, they will be followed by the items of Y in any order.

The support of a rule is the number of sequence containing the rule divided by the total number of sequences. The confidence of a rule is the number of sequence containing the rule divided by the number of sequences containing its antecedent. The goal of sequential rule mining is to discover all sequential rules having a support and confidence no less than two thresholds given by the user named “minsup” and “minconf” .For example, on the right part of the following figure some sequential rules are shown for minsup=0.5 and minconf=0.5.

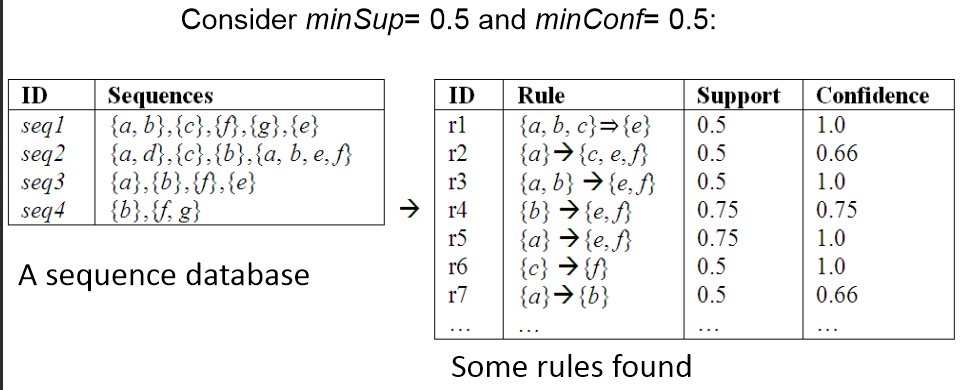


Fig 1: Example for generating rules.

Ref: <http://data-mining.philippe-fournier-viger.com/introduction-frequent-pattern-mining/>

* 1. DISADVANTAGES OF EXISTING SYSTEM:

In the present system, the following are the disadvantages:

* **Data/information extraction**: Our focus will be on extraction of structured data from Web pages, such as products and search results. Extracting such data allows one to provide services. Two main types of techniques, machine learning and automatic extraction are covered.
* **Web information integration and schema matching:** Although the Web contains a huge amount of data, each web site (or even page) represents similar information differently. How to identify or match semantically similar data is a very important problem with many practical applications. Some existing techniques and problems are examined.
* **Opinion extraction from online sources**: There are many online opinion sources, e.g., customer reviews of products, forums, blogs and chat rooms. Mining opinions (especially consumer opinions) is of great importance for marketing intelligence and product benchmarking. We will introduce a few tasks and techniques to mine such sources.
* **Knowledge synthesis**: Concept hierarchies or ontology are useful in many applications. However, generating them manually is very time consuming. A few existing methods that explores the information redundancy of the Web will be presented. The main application is to synthesize and organize the pieces of information on the Web to give the user a coherent picture of the topic domain.
* **Segmenting Web pages and detecting noise**: In many Web applications, one only wants the main content of the Web page without advertisements, navigation links, copyright notices. Automatically segmenting Web page.to extract the main content of the pages is interesting problem. A number of interesting techniques have been proposed in the past few years.

Web mining technology when used on data of personal nature might cause concerns. Privacy is considered lost, when information concerning an individual is obtained, used or disseminated, especially if this occurs without their knowledge or consent. Also, the companies collecting the data for a specific purpose might use the data for a totally different purpose, and this essentially violates the user’s interests.

It is difﬁcult to ﬁnd certain groups of web users with similar favorite web pages. In generating frequent patterns, based on a user-speciﬁed minimum support threshold, the process can obtain frequent web pages for all web users.

This means that if some web pages are frequently accessed by one web user, then they are accessed by other web users with a high probability. These kinds of frequently visited web pages are without discrimination in clustering web users; they are like noise pages in clustering.

* 1. PROPOSED SYSTEM:

Briefly, the large web pages are generated from web log data. The large web pages are scanned and parsed by using Simple page Linked Graph (SPLG). From there emerging patterns are obtained. The obtained patterns are clustered and named. This approach can be using the following diagram:

Fig 2: work flow

* + 1. PRE PROCESSING OF DATA:

Data preprocessing is a data mining technique that involves transforming raw data into an understandable format. Real-world data is often incomplete, inconsistent, and/or lacking in certain behaviors or trends, and is likely to contain many errors. Data preprocessing is a proven method of resolving such issues. Data preprocessing prepares raw data for further processing. Data goes through a series of steps during preprocessing:

* **Data Cleaning**: Data is cleansed through processes such as filling in missing values, smoothing the noisy data, or resolving the inconsistencies in the data.
* **Data Integration**: Data with different representations are put together and conflicts within the data are resolved.
* **Data Transformation**: Data is normalized, aggregated and generalized.
* **Data Reduction**: This step aims to present a reduced representation of the data in a data warehouse.
* **Data Discretization**: Involves the reduction of a number of values of a continuous attribute by dividing the range of attribute intervals.

In this project, Data Preprocessing is done considering the following four categories:

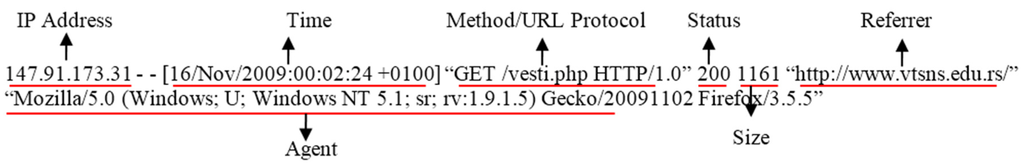
* **Removing Records with Missing Value Data**: Missing values are a common occurrence, and you need to have a strategy for treating them. A missing value can signify a number of different things in your data. Perhaps the data was not available or not applicable or the event did not happen. It could be that the person who entered the data did not know the right value, or missed filling in. Data mining methods vary in the way they treat missing values. Typically, they ignore the missing values, or exclude any records containing missing values, or replace missing values with the mean, or infer missing values from existing values. For example, if a click‐through to a web page was executed while the web server was shut down, then, in the log file, only the IP address, user ID, and access time will be recorded; the method, URL, referrer, and agent are lost. This kind of record cannot be used for our mining task, so these records must be deleted.
* **Removing Records with Exception Status Number**: when the data is being is preprocessed some situations occur like a bad client request. When a page is requested sometimes error occurs like 400 Bad Request the server cannot or will not process the request due to an apparent client error (e.g., malformed request syntax, too large size, invalid request message framing, or deceptive request routing).401 Unauthorized-similar to 403 Forbidden, but specifically for use when authentication is required and has failed or has not yet been provided. The response must include a WWW-Authenticate header field containing a challenge applicable to the requested resource. 401 semantically mean "unauthenticated", i.e. the user does not have the necessary credentials.403 Forbidden-the request was valid, but the server is refusing action. The user might not have the necessary permissions for a resource.404 Not Found-the requested resource could not be found but may be available in the future. Subsequent requests by the client are permissible. Sometimes, the error can also be like 500 Internal Server Error-A generic error message, given when an unexpected condition was encountered and no more specific message is suitable. 501 Not implemented-The server either does not recognize the request method, or it lacks the ability to fulfill the request. Usually this implies future availability (e.g., a new feature of a web-service API). 502 Bad Gateway-The server was acting as a gateway or proxy and received an invalid response from the upstream server. 503 Service Unavailable-The server is currently unavailable (because it is overloaded or down for maintenance). Generally, this is a temporary state. 504 Gateway Time-out-The server was acting as a gateway or proxy and did not receive a timely response from the upstream server. 505 HTTP Version Not Supported The server does not support the HTTP protocol version used in the request.
* **Removing Irrelevant Records with No Significant URLs:** When data is cleaned there are certain extensions which will not be useful. Those have to be removed. The extensions that cannot be used are .txt, .jpg, .gif, or .js. These will be automatically generated when a page is requested. These have to be removed to make data precise.
* **Selecting the Essential Attributes:** Consider a web log file, there will be different attributes in it. Out of which only some are useful to the project. Therefore, those should be considered and remaining attributes are deleted. For example, consider the below image. It has different attributes such as status, time agent etc., but we need only IP address and the Referrer (URL). So remaining attributes are deleted.

Fig 3: Example record in a Web log file.

Ref:<http://www.mdpi.com/sustainability/sustainability-08-00239/article_deploy/html/images/sustainability-08-00239-g002-1024.png>

* + 1. GENERATION OF LARGE WEB PAGES:

Web usage mining is the task of discovering the activities of the users while they are browsing and navigating through the Web. The aim of understanding the navigation preferences of the visitors is to enhance the quality of electronic commerce services (e-commerce), to personalize the Web portals or to improve the Web structure and Web server performance. In this case, the mined data are the log files which can be seen as the secondary data on the web where the documents accessible through the Web are understood as primary data.

There are three types of log files that can be used for Web usage mining. Log files are stored on the server side, on the client side and on the proxy servers. By having more than one place for storing the information of navigation patterns of the users makes the mining process more difficult. Really reliable results could be obtained only if one has data from all three types of log file.

The reason for this is that the server side does not contain records of those Web page accesses that are cached on the proxy servers or on the client side. Besides the log file on the server, that on the proxy server provides additional information. However, the page requests stored in the client side are missing. Yet, it is problematic to collect all the information from the client side. Thus, most of the algorithms work based only the server side data. Some commonly used data mining algorithms for Web usage mining are association rule mining, sequence mining and clustering.

A large page set is a set of frequent web pages. We define frequent web pages as those where support thresholds are greater than, or equal to, a user‐specified minimum support threshold. Large Web Pages (LWPs) denote the set of web pages that are accessed by web users with sufficient frequency over a period of time. A special period of time called a user session is an important definition for generating LWPs for web users.

In the view of our project, once the data preprocessing is done two attributes are considered IP address and the URL. IP address can be used to identify user as it is unique. The URL’s will be all the websites visited by the user. These two will be stored in the database. Here is a small example for it.

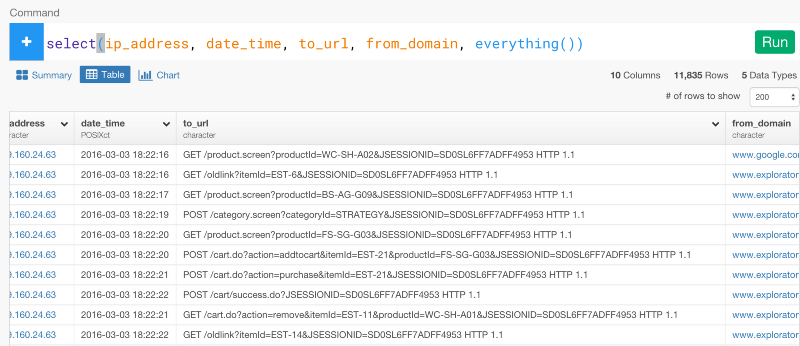
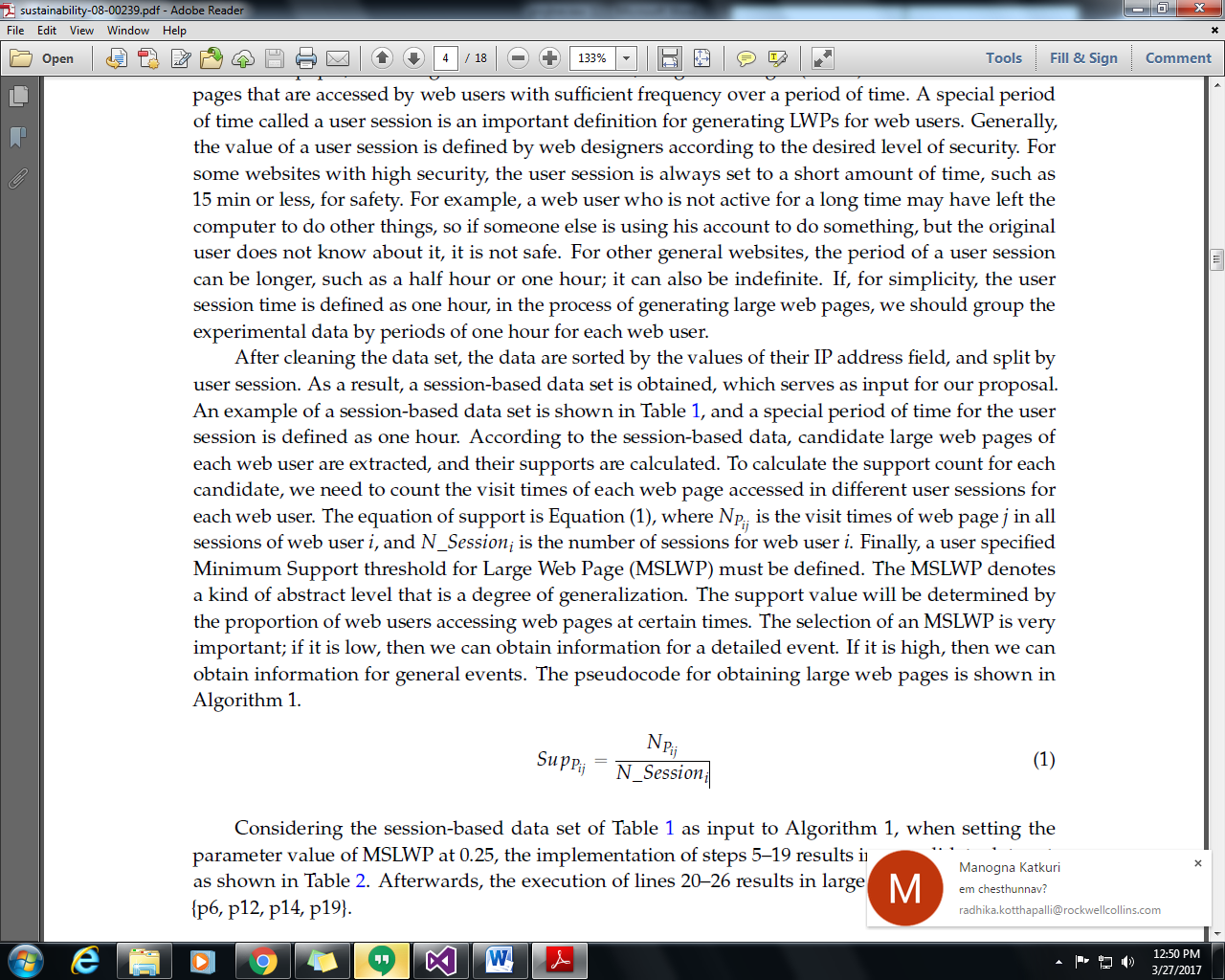


Fig 4: Example log file (real time)

Ref: [https://cdn-images-1.medium.com/max/800/1\*EEgC8FsjsYQZJsGpqFp0\_A.png](https://cdn-images-1.medium.com/max/800/1*EEgC8FsjsYQZJsGpqFp0_A.png)

For every user support value will be calculated according to the data. To calculate the support count for each candidate, we need to count the visit times of each web page accessed in different user sessions for each web user. The equation of support can be given as follows.

Fig 5: Support Formula

Ref: <http://www.mdpi.com/2071-1050/8/3/239/htm>

Here, NPij is the visit times of web page j in all sessions of web user i, and N\_Sessioni is the number of session for web user i. This will calculate the support value of web users. Support is an indication of how frequently the item set appears in the database.

So, in this project it indicates the no of times a web site has been visited. After calculating the support count the support count can be tabulated as follows:

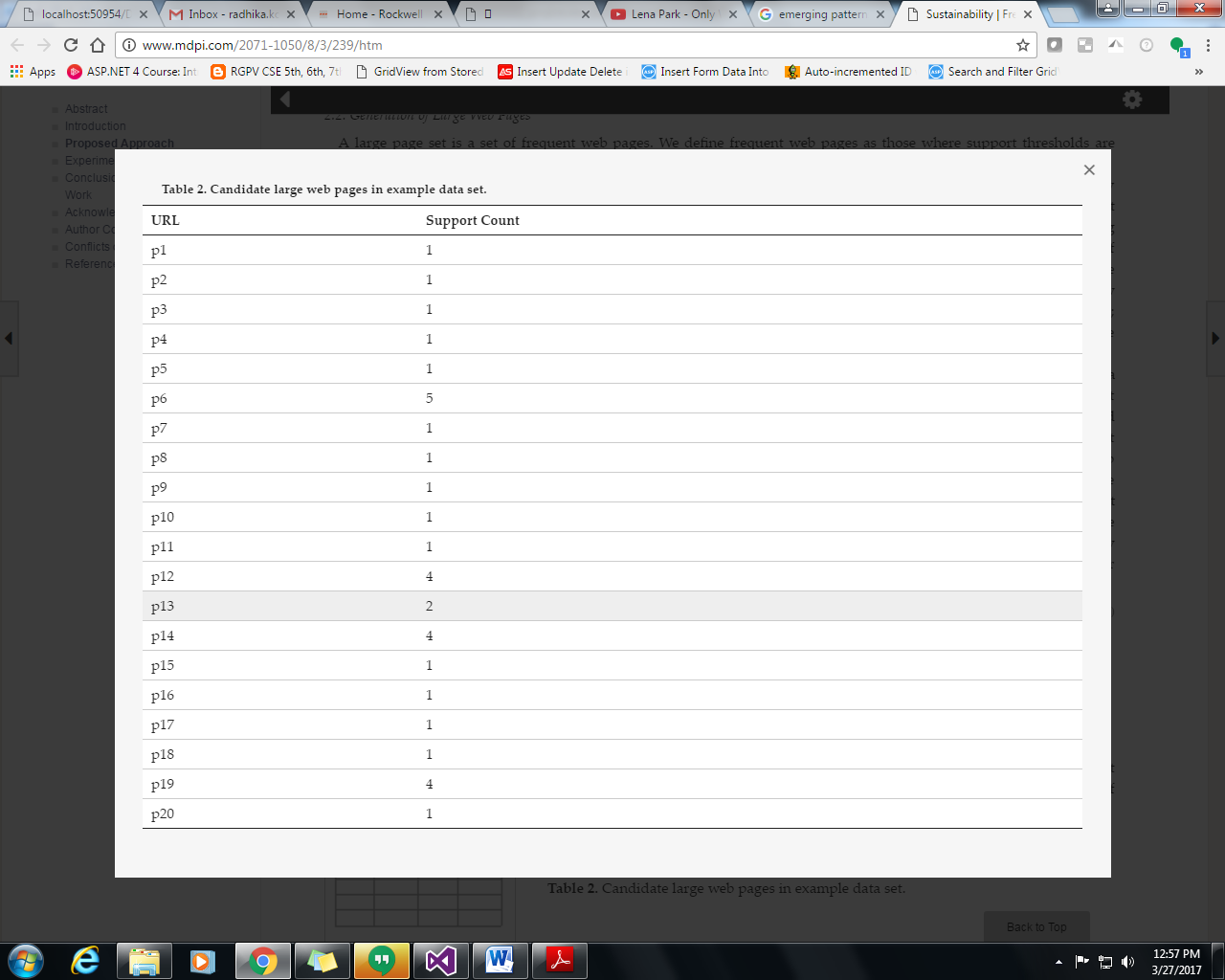


Fig 6: User 1 Support count

Ref: <http://www.mdpi.com/2071-1050/8/3/239/htm>

* + 1. GENERATING SPLG’S:

This is the third step the project. Once the support count is ready and the large web pages are generated SPLG’s are used. SPLG stand for Simple Page Linked Graph. Basically, a graph is a structure amounting to a set of objects in which some pairs of the objects are in some sense "related". The objects correspond to mathematical abstractions called vertices (also called nodes or points) and each of the related pairs of vertices is called an edge (also called an arc or line). In regular page‐linked graphs, each edge consists of every two web pages that are contained in one session.

Applying the concept of the SPLG to the structure of web page links can reduce large and complex regular page‐linked graphs to simple ones in order to reduce noise web pages. In the SPLG, links between each of the two large web pages should be checked. To check the link between every two vertices, the direction of link does not need to be considered, if the two vertices are visited by one user in one session, then they are connected. For every user, SPLG is created. From that the most frequently pages are separated.

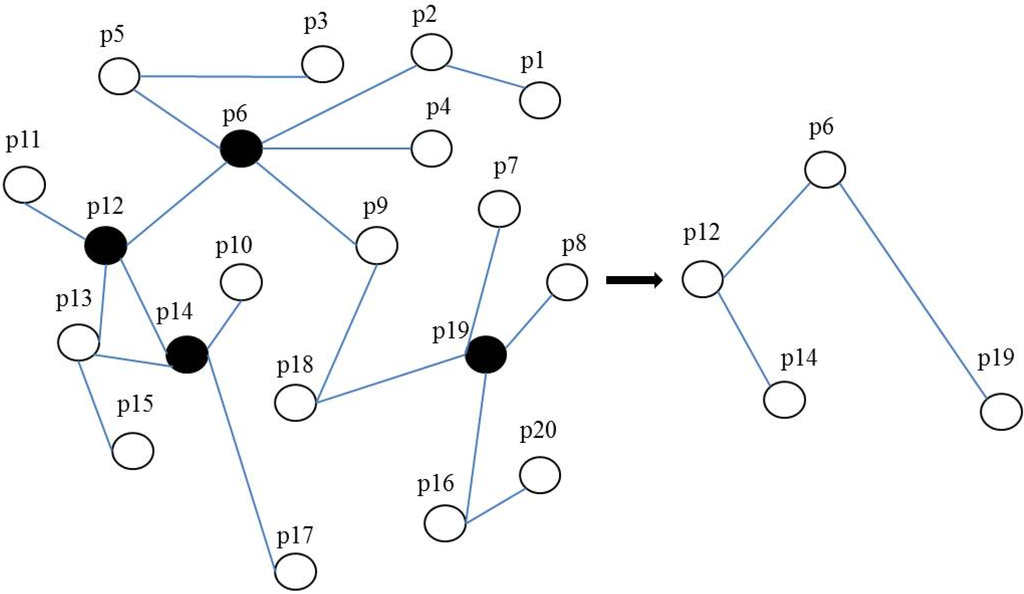
The figure 7 indicates example SPLG of a web user.

Fig 7: Example of SPLG.

Ref: <http://www.mdpi.com/sustainability/sustainability-08-00239/article_deploy/html/images/sustainability-08-00239-g003-1024.png>

* + 1. GENERATION OF EMERGING PATTERNS:

For every user SPLG’s are derived then for every user emerging patterns are observed. This can also be called as the Pattern analysis. When we decompose a complex problem we often find patterns among the smaller problems we create. The patterns are similarities or characteristics that some of the problems share. Pattern recognition is one of the four cornerstones of Computer Science. It involves finding the similarities or patterns among small, decomposed problems that can help us solve more complex problems more efficiently. Pattern mining algorithms can be applied on various types of data such as transaction databases, sequence databases, streams, strings, spatial data, graphs, etc. Pattern mining algorithms can be designed to discover various types of patterns: subgraphs, associations, indirect associations, trends, periodic patterns, sequential rules, lattices, sequential patterns, high-utility patterns, etc.

* + 1. CLUSTERING OF WEB USERS:

In this project, k-means algorithm is used. K-means is one of the simplest unsupervised learning algorithms that solve the well-known clustering problem. The procedure follows a simple and easy way to classify a given data set through a certain number of clusters (assume k clusters) fixed apriori. The main idea is to define k centers, one for each cluster. These centers should be placed in a cunning way because of different location causes different result. So, the better choice is to place them as much as possible far away from each other.

The next step is to take each point belonging to a given data set and associate it to the nearest center. When no point is pending, the first step is completed and an early group age is done. At this point we need to re-calculate k new centroids as barycenter of the clusters resulting from the previous step. After we have these k new centroids, a new binding has to be done between the same data set points and the nearest new center. A loop has been generated. As a result of this loop we may notice that the k centers change their location step by step until no more changes are done or in other words centers do not move any more.

In the view of our project, after getting emerging patterns they are clustered according to the web users. In this way clusters are obtained. Now, these clusters are considered as the frequent web pages. The main advantages of k-means algorithm are fast, robust and easier to understand. Gives best result when data set are distinct or well separated from each other.

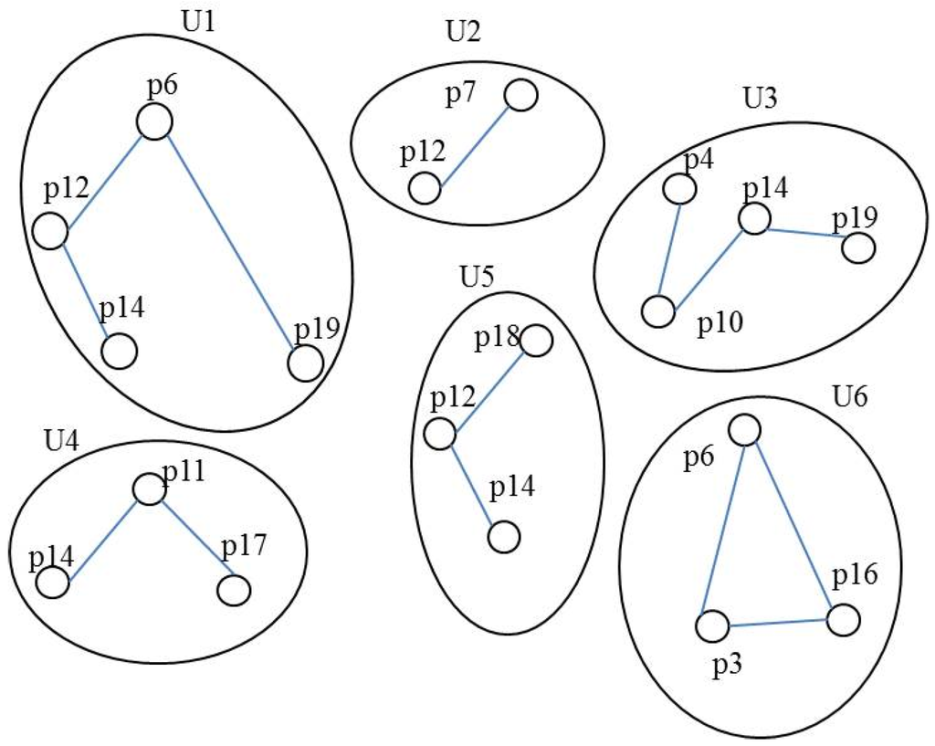


Fig 8: Clustering of web pages

Ref: <http://www.mdpi.com/2071-1050/8/3/239/htm>

* + 1. ANNOTATION OF CLUSTERS:

As the last step of this project, the obtained clusters are annotated or named. This annotation is based on the concept of Folksonomies. Folksonomies means a user-generated system of classifying and organizing online content into different categories by the use of metadata such as electronic tags.

A folksonomy is a system in which users apply public tags to online items, typically to aid them in re-finding those items. This can give rise to a classification system based on those tags and their frequencies, in contrast to a taxonomic classification specified by the owners of the content when it is published. This practice is also known as collaborative tagging, social classification, social indexing, and social tagging. However, these terms have slightly different meanings than folksonomy. Folksonomy was originally “the result of personal free tagging of information for one’s own retrieval.”

1. ANALYSIS

**3**.**1INTRODUCTION:**

As per the given approach, let’s brief the steps that were done. Data mining is the computational process of discovering patterns in large amount data sets involving methods at the intersection of artificial intelligence, machine learning of Data System. The World Wide Web is now a huge database with this growth there arises a need for analyzing the data. The proposed approach can be briefed as:

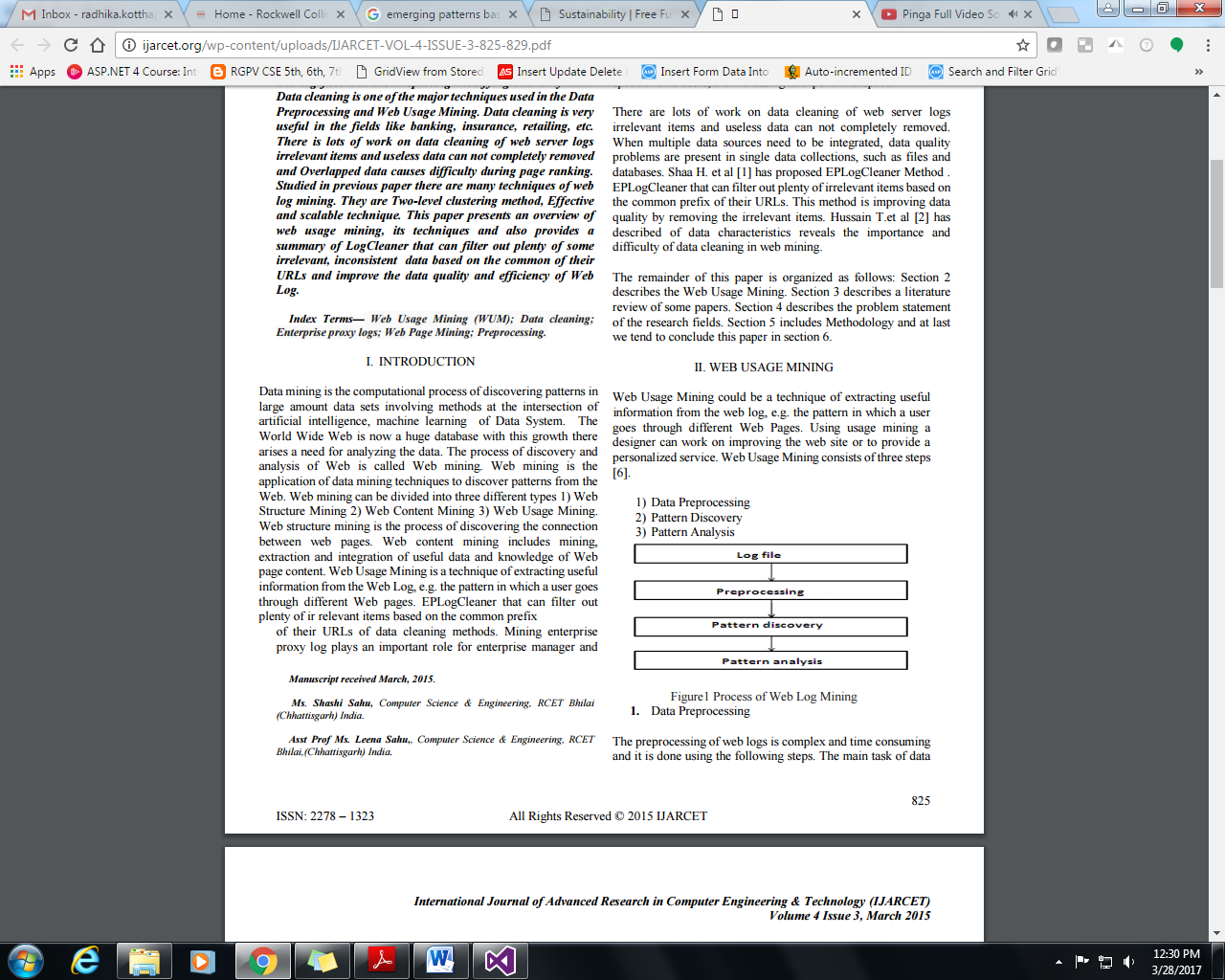


Fig 9: Briefing the Approach

Ref: <http://ijarcet.org/wp-content/uploads/IJARCET-VOL-4-ISSUE-3-825-829.pdf>

The process of discovery and analysis of Web is called Web mining. Web mining is the application of data mining techniques to discover patterns from the Web. Web mining can be divided into three different types 1) Web Structure Mining 2) Web Content Mining 3) Web Usage Mining. Web structure mining is the process of discovering the connection between web pages. Web content mining includes mining, extraction and integration of useful data and knowledge of Web page content. Web Usage Mining is a technique of extracting useful information from the Web Log, e.g. the pattern in which a user goes through different Web pages.

Web Log is generally noisy and ambiguous. Web applications are increasing at an enormous speed and its users, are increasing at exponential speed. This creates a need for identifying right data and the wrong data. Therefore, this leads to concept of Data Cleaning. Data cleaning is the process of detecting and correcting the irrelevant, incomplete data from the datasets and log files and then replacing modifying this dirty data. Data cleaning is one of the major techniques used in the Data Preprocessing and Web Usage Mining. Data cleaning is very useful in the fields like banking, insurance, retailing, etc. There is lots of work on data cleaning of web server logs irrelevant items and useless data can not completely removed and Overlapped data causes difficulty during page ranking. In our project data cleaning is done by considering four factors. They can be listed as follows

* Removing Records with Missing Value Data.
* Removing Records with Exception Status Numbers.

#### Removing Irrelevant Records with No Significant URLs.

* Selecting the Essential Attributes.

Then large web pages are to be derived from the cleaned data. A web log file denotes a data set; Large Web Pages (LWPs) denote the set of web pages that are accessed by web users with sufficient frequency over a period of time. After cleaning the data set, the data are sorted by the values of their IP address field, and split by user session. As a result, a session-based data set is obtained, which serves as input for our proposal. According to the session-based data, candidate large web pages of each web user are extracted, and their supports are calculated.

To calculate the support count for each candidate, we need to count the visit times of each web page accessed in different user sessions for each web. Support count is the important factor of our approach, based on those emerging patterns evolves. In the next step, frequent patterns are observed.

Classification is an important data mining problem. Given a training database of records, each tagged with a class label, the goal of classification is to build a concise model that can be used to predict the class label of future, unlabeled records. Many classification models have been proposed in the literature. Recently a new type of knowledge patterns, called emerging patterns has evolved. EPs are defined as multivariate features (i.e., item sets) whose supports (or frequencies) change significantly from one class to another. The concept of emerging patterns is very suitable for serving as a classification model.

By aggregating the differentiating power of EPs, the constructed classification systems are usually more accurate than other existing state-of-the-art classifiers. Then the obtained patterns are clustered based on the k-means algorithm.

Place K points into the space represented by the objects that are being clustered. These points represent initial group centroids. Assign each object to the group that has the closest centroid. When all objects have been assigned, recalculate the positions of the K centroids. Repeat Steps until the centroids no longer move. This produces a separation of the objects into groups from which the metric to be minimized can be calculated. After clustering, we label the clusters based on the concept of Folksonomies. Each cluster is defined as one user group, and the web pages in each cluster are defined as online items.

* 1. **SOFTWARE REQUIREMENT SPECIFICATION**:

This projected is implemented in .NET framework using c# and tool used for this is Microsoft Visual studio. Microsoft Visual Studio is an integrated development environment (IDE) from Microsoft. It is used to develop computer programs for Microsoft Windows, as well as web sites, web apps, web services and mobile apps. Visual Studio uses Microsoft software development platforms such as Windows API, Windows Forms, Windows Presentation Foundation, Windows Store and Microsoft Silverlight. It can produce both native code and managed code.

Visual Studio supports different programming languages and allows the code editor and debugger to support (to varying degrees) nearly any programming language, provided a language-specific service exists. Built-in languages include C, C++ and C++/CLI (via Visual C++), VB.NET (via Visual Basic .NET), C# (via Visual C#), and F# (as of Visual Studio 2010). Support for other languages such as Python, Ruby, Node.js, and M among others is available via language services installed separately. It also supports XML/XSLT, HTML/XHTML, JavaScript and CSS. Java (and J#) was supported in the past.

ASP.NET is a web development platform, which provides a programming model, a comprehensive software infrastructure and various services required to build up robust web applications for PC, as well as mobile devices. It works on top of the HTTP protocol, and uses the HTTP commands and policies to set a browser-to-server bilateral communication and cooperation. It is a part of Microsoft .Net platform. ASP.NET applications are compiled codes, written using the extensible and reusable components or objects present in .Net framework. These codes can use the entire hierarchy of classes in .Net framework. The ASP.NET application codes can be written in any of the following languages: C#, Visual Basic.Net, Jscript, J#

C# is a modern, general-purpose, object-oriented programming language developed by Microsoft and approved by European Computer Manufacturers Association (ECMA) and International Standards Organization (ISO). It is designed for Common Language Infrastructure (CLI), which consists of the executable code and runtime environment that allows use of various high-level languages on different computer platforms and architectures.

* + 1. SOFTWARE REQUIRMENT :
* Tools: Microsoft Visual Studio – 2013
* Operating System: Windows 7 and higher.
* Database: My SQL
* Client Side: HTML and JavaScript
* Server Side: ASP.NET and ADO.NET.
  + 1. HARDWARE REQUIREMENT
* CPU: Intel Core i5 (3rd Gen) 3317U / 1.7 GHz
* Max Turbo Speed: 2.6 GHz
* Number of Cores: Dual-Core
* Cache: 3 MB
* RAM: 4 GB
* Hard drive: 500 GB
  1. CONTENT DIAGRAM OF PROJECT:

We generate large web pages from processed web log data, then scan and transform the clean data set into simple page-linked graphs (SPLGs), and then, generate emerging patterns in the generated SPLGs. We cluster web users based on generated emerging patterns, and finally, label the clusters with typical web pages. Our work flow is shown in Figure

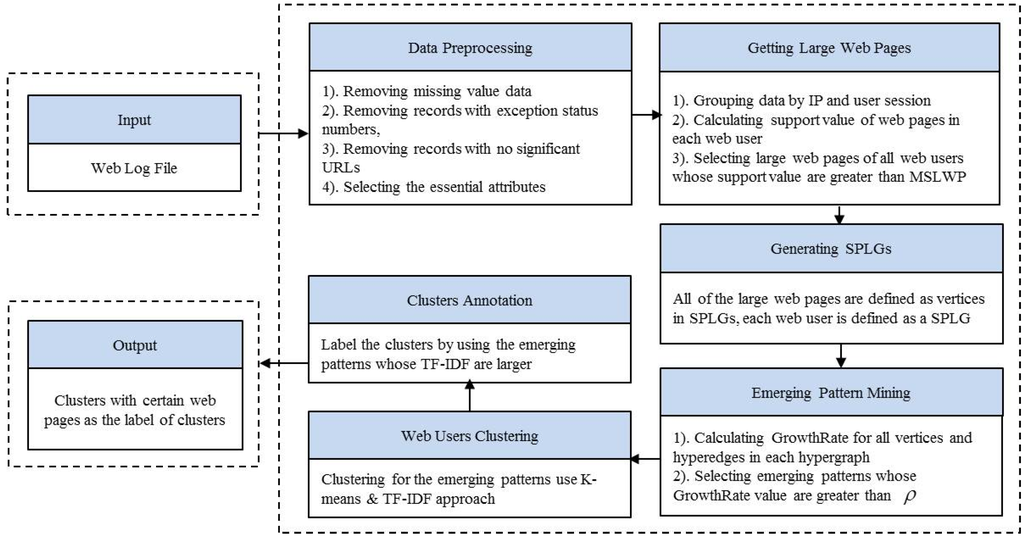
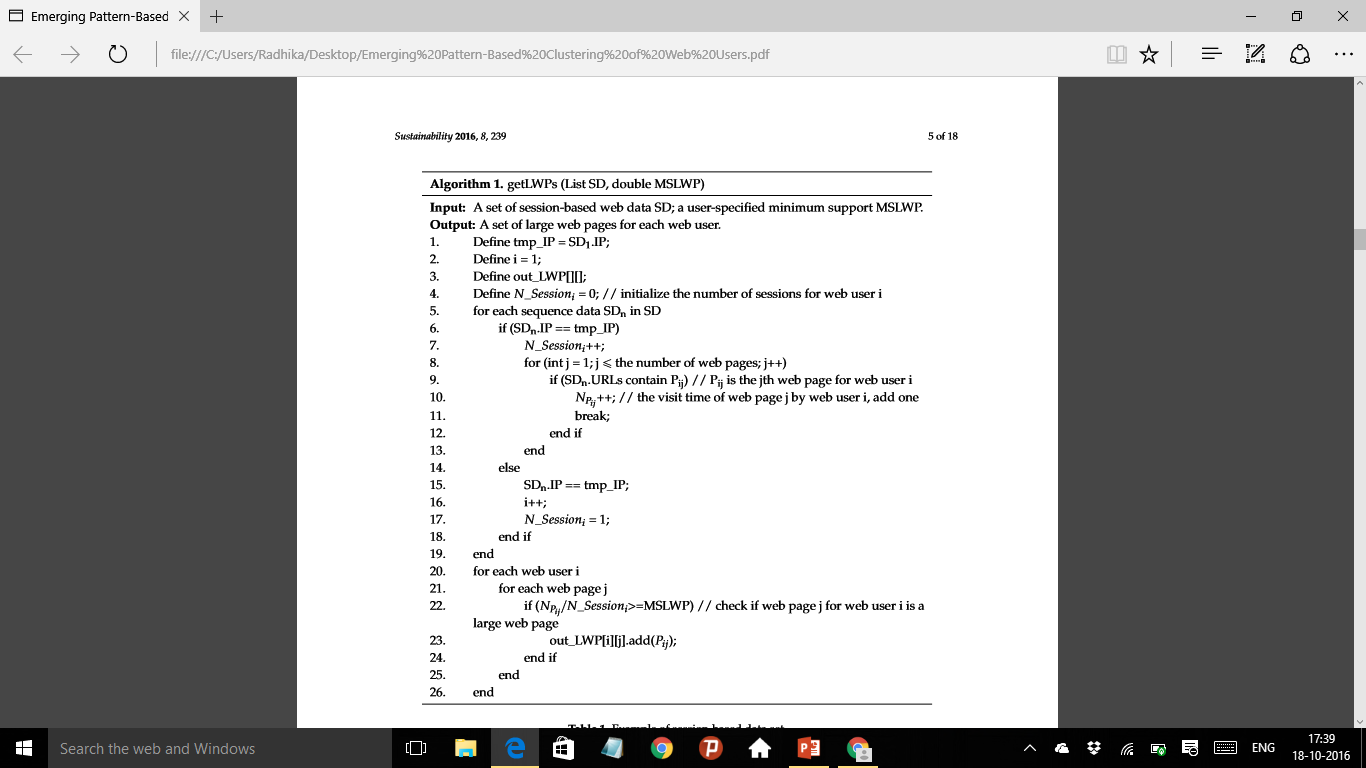


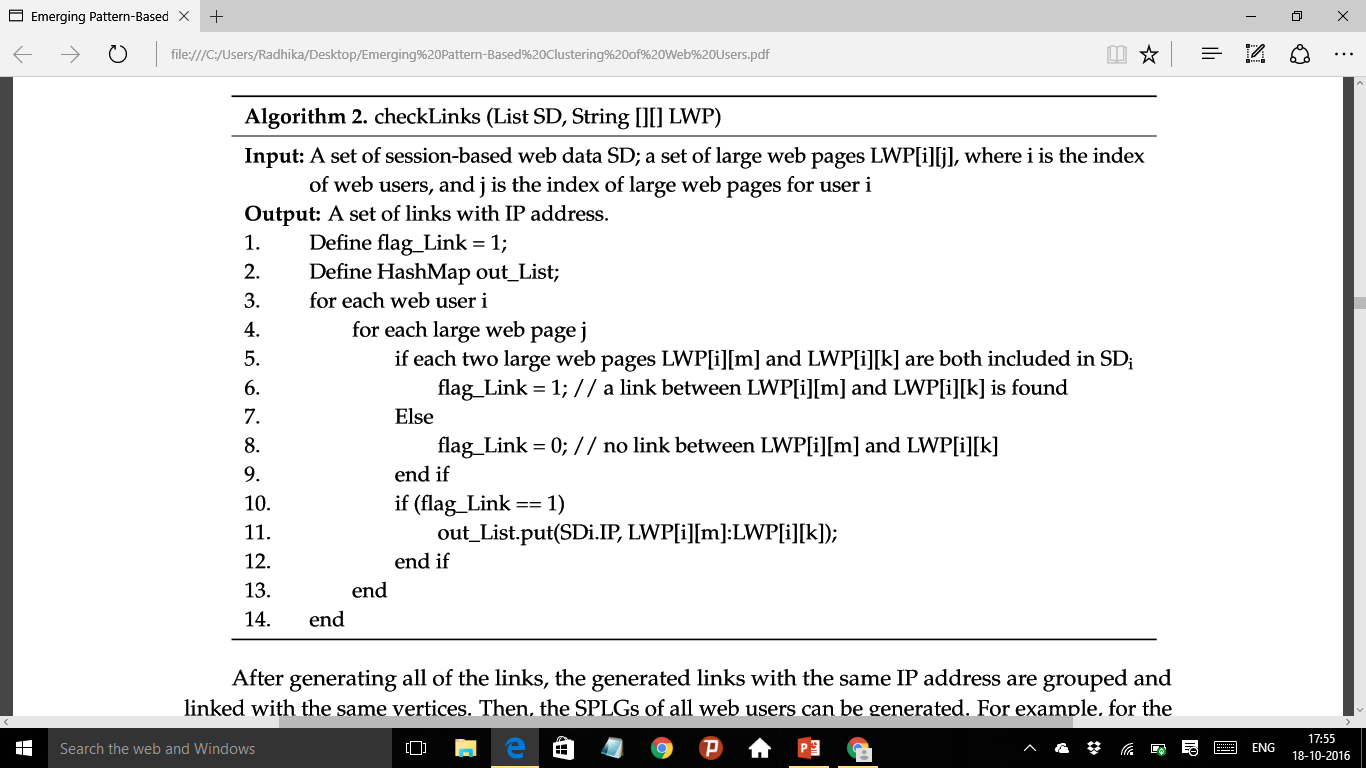
Fig 10: Content Diagram

Ref: <http://www.mdpi.com/sustainability/sustainability-08-00239/article_deploy/html/images/sustainability-08-00239-g001-1024.png>

The flow starts with input the web log file. Web log data is automatically recorded in web log files on web servers when web users access the web server through their browsers. Not all of the records sorted into the web log files have the right format or are necessary for the mining task, so before analyzing the web log data, a data cleaning phase needs to be implemented.

Data cleaning is the process of detecting and correcting (or removing) corrupt or inaccurate records from a record set, table, or database and refers to identifying incomplete, incorrect, inaccurate or irrelevant parts of the data and then replacing, modifying, or deleting the dirty or coarse data. Data cleansing may be performed interactively with data wrangling tools, or as batch processing through scripting. In this project

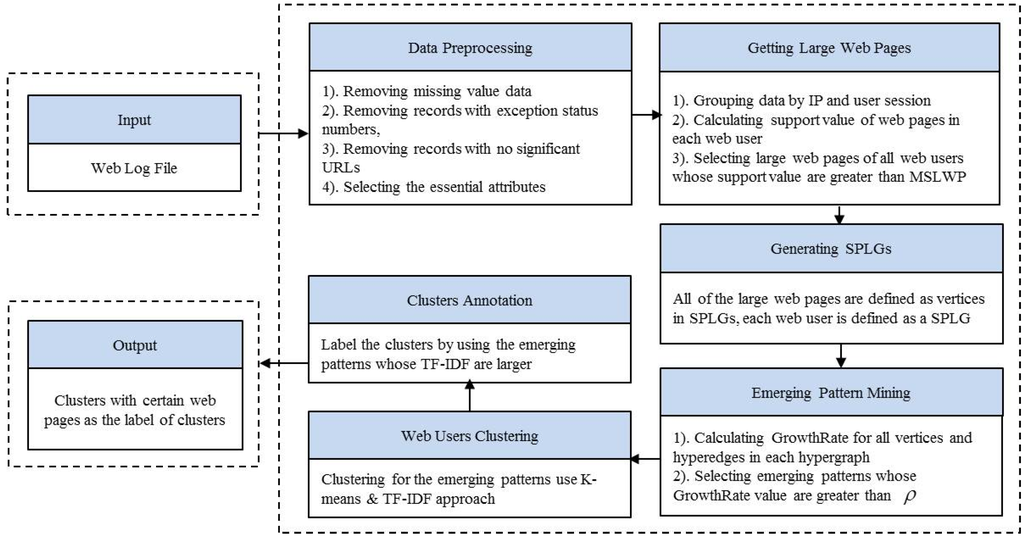




4.DESIGN

4.1INTRODUCTION:

In this section, we generate large web pages from processed web log data, then scan and transform the clean data set into simple page-linked graphs (SPLGs), and then, generate merging patterns in the generated SPLGs. We cluster web users based on generated emerging patterns, and ﬁnally, label the clusters with typical web pages

DIAGRAM: 

4.1 Implementation

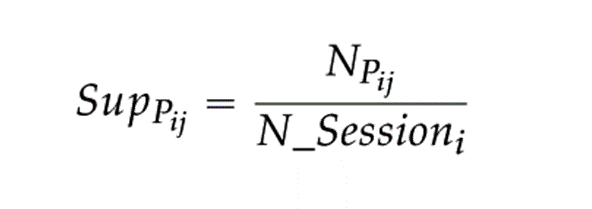
* 1. MODULE DESIGN AND ORGANIZATION:

4.3.1 PREPROCESSING OF DATASET: Web log data is automatically recorded in web log files on web servers when web users access the  web server through their browsers. Not all the records sorted into the web log files have the right format or are necessary for the mining task, so before analyzing the web log data, a data cleaning phase needs to be implemented.

* + - * + Removing Records with Missing Value Data
        + Removing Records with Exception Status Numbers
        + Removing Irrelevant Records with No Significant URLs
        + Selecting the Essential Attributes

4.3.2 GENERATION OF LARGE WEB PAGES:

Large web pages are the set of frequent item sets. From a Large Web Pages (LWP) set frequent item sets are observed with some minimum threshold. After cleaning the data, the frequent pages are observed by the support value. The support value can be calculated by the formula



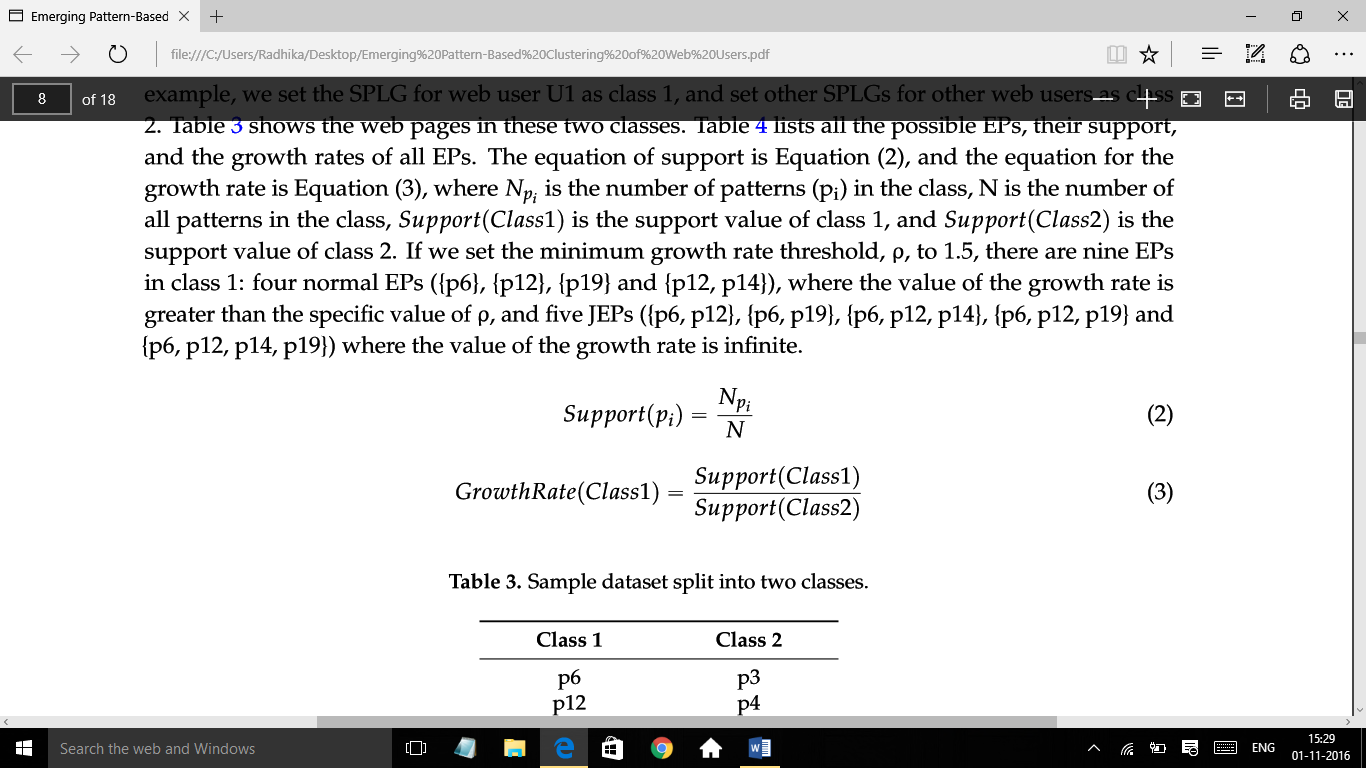
Where, Npij is the no of times a web page j is been visited in the web user i and N\_Sessioni is the no of sessions of web user i. Finally, a user specified Minimum Support threshold for Large Web Page (MSLWP)must be defined.  The MSLWP denotes a kind of abstract level that is a degree of generalization. The support value will be determined by the proportion of web users accessing web pages at certain times. The selection of an MSLWP is very important; if it is low, then we can obtain information for a detailed event.

4.3.3 GENERATION OF SIMPLE PAGE LINKED-GRAPH (SPLG):

After generating large web pages for each web user, all the large web pages are deﬁned as vertices in the SPLG. In regular page-linked graphs, each edge consists of every two web pages that are contained in one session. However, in a SPLG, each edge consists of every two-large web page of the web user. Applying the concept of the SPLG to the structure of web page links can reduce large and complex regular page-linked graphs to simple ones to reduce noise web pages. In the SPLG, links between each of the two large web pages should be checked. To check the link between every two vertices, the direction of link does not need to be considered, if the two vertices are visited by one user in one session, then they are connected

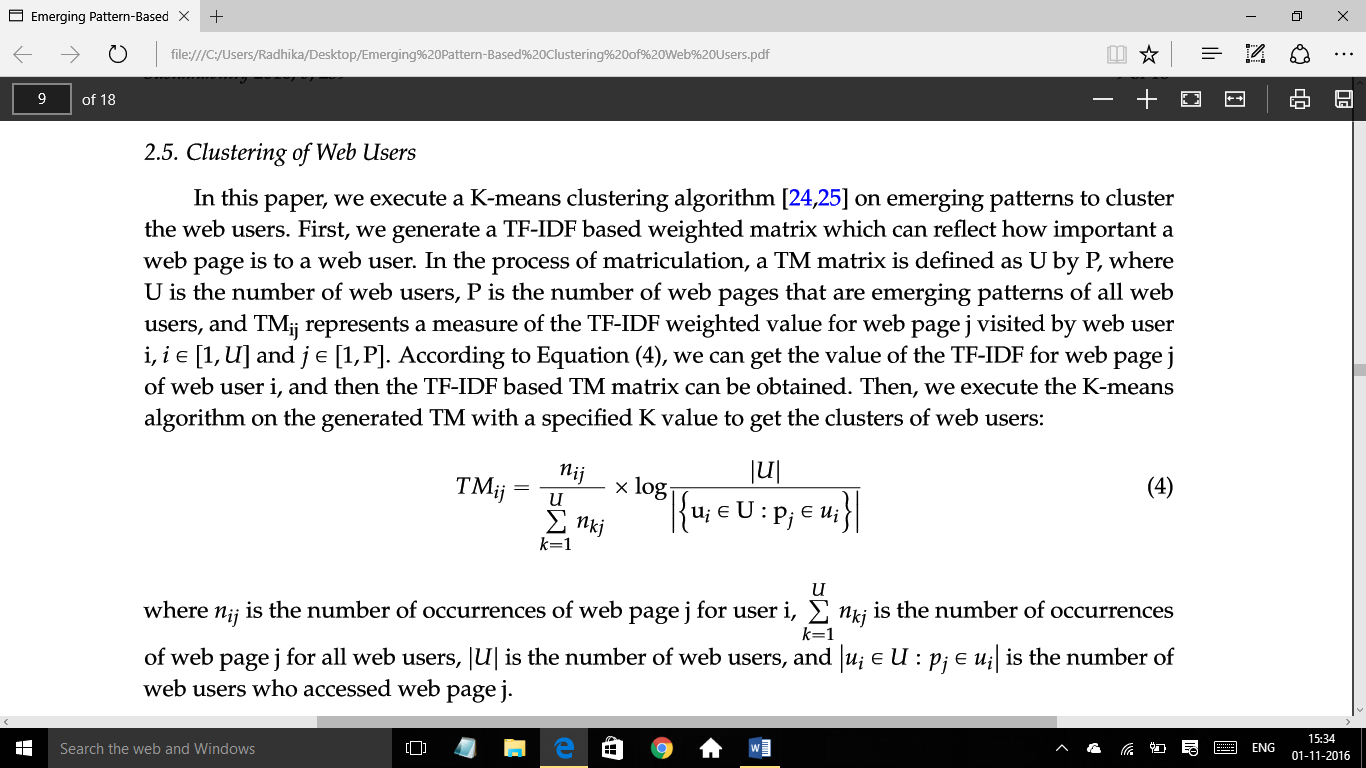
4.3.4 GENERATION OF EMERGING PATTERNS:

After generating SPLGs for all web users, we try to ﬁnd emerging patterns in these SPLGs. Examples of SPLGs for some web users. This can be possible by finding the growth and support value of the given item sets. They can be calculated as follows:



4.3.5 CLUSTERING OF WEB USERS:

we execute a K-means clustering algorithm on emerging patterns to cluster the web users. First, we generate a TF-IDF based weighted matrix which can reﬂect how important a web page is to a web user. we can get the value of the TF-IDF for web page j of web user i, and then the TF-IDF based TM matrix can be obtained. Then, we execute the K-means algorithm on the generated TM with a speciﬁed K value to get the clusters of web users.



4.3.6 ANNOTATION OF CLUSTERS:

After clustering, we label the clusters based on the concept of Folksonomies. Each cluster is deﬁned as one user group, and the web pages in each cluster are deﬁned as online items, we use TF-IDF to calculate the frequency of each web page in each cluster. we can calculate the TF-IDF value of each web page in each cluster, and then we can select some web pages where TF-IDF values are among the TopN (N can be the number chosen by a user with freedom, where N is smaller than the number of web pages in each cluster) and the largest in each cluster is the label of this cluster.

5.CONCLUSION

In this project, we tried to implement an emerging pattern from large web pages. These can be implemented from the we log data. The pages are implemented as the nodes in the SPLG’s. Then patterns are observed and clustered. The clustered patters are named per the concept of folksonomies. TF-IDF.

The main result of this study is to generate large web pages and emerging patterns to identify the personal favorite web pages of each user by eliminating noise due to overall popular pages.