###### 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**

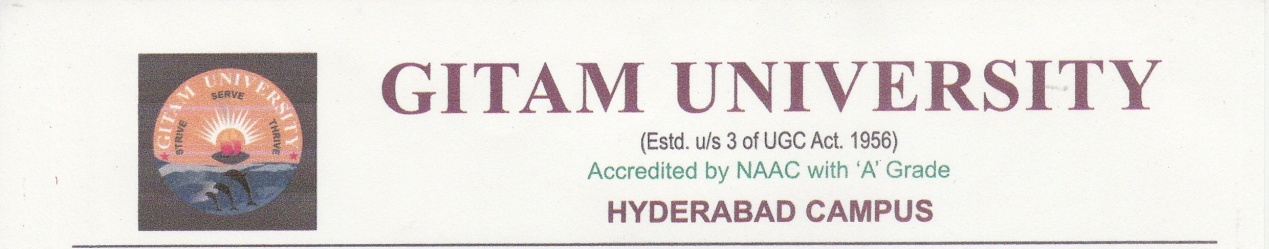
****

**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 SYSTEM....................................................................2

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

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

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

* 1. SOFTWARE REQUIREMENT SPECIFICATION............................................................3
  2. HARDWARE REQUIREMENT.........................................................................................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.CONCLUSION..................................................................................................................................7

LIST OF FIGURES

1. Implementation.........................................................................................................................................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 USAGE MINING

1.INTRODUCTION

* 1. MOTIVATION:

Emerging patterns are the frequent item sets that change significantly from one dataset to another dataset. This application when applied to the large usage of web mining, the results are very accurate and useful. In the growing data, there is lot of information to be gathered. Our main intention in the project is if a user visits some pages frequently, those are to be clustered and understand their behavior. By clustering them the user can access the pages in an easy way. Naming the clustering will be the final step in this project.

* 1. PROBLEM DEFINATION:

In the present day, there is a huge data usage and most of it is in an unstructured form. 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

Our project main motto is to cluster the web pages of the user web log that are frequently visited by him/her. By clustering the web pages, we 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. Naming of clusters can be can by the Term Frequency- Inverse Document Frequency(TF-IDF).

**2.**LITERATURE SURVEY

2.1 INTRODUCTION

With the rapid growth of the Internet, most research on the Internet has revealed some very hot topics, such as social networks web mining, and so on. 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:

Present, there is huge web log data the pages can be accessed from the log data.

2.3 DISADVANTAGES OF EXISTING SYSTEM:

In the present system, the following are the disadvantages

* there is lot of noise in the web log data
* the frequent patterns are not observed
* the behavior of pages is not observed

2.4 PROPOSED SYSTEM:

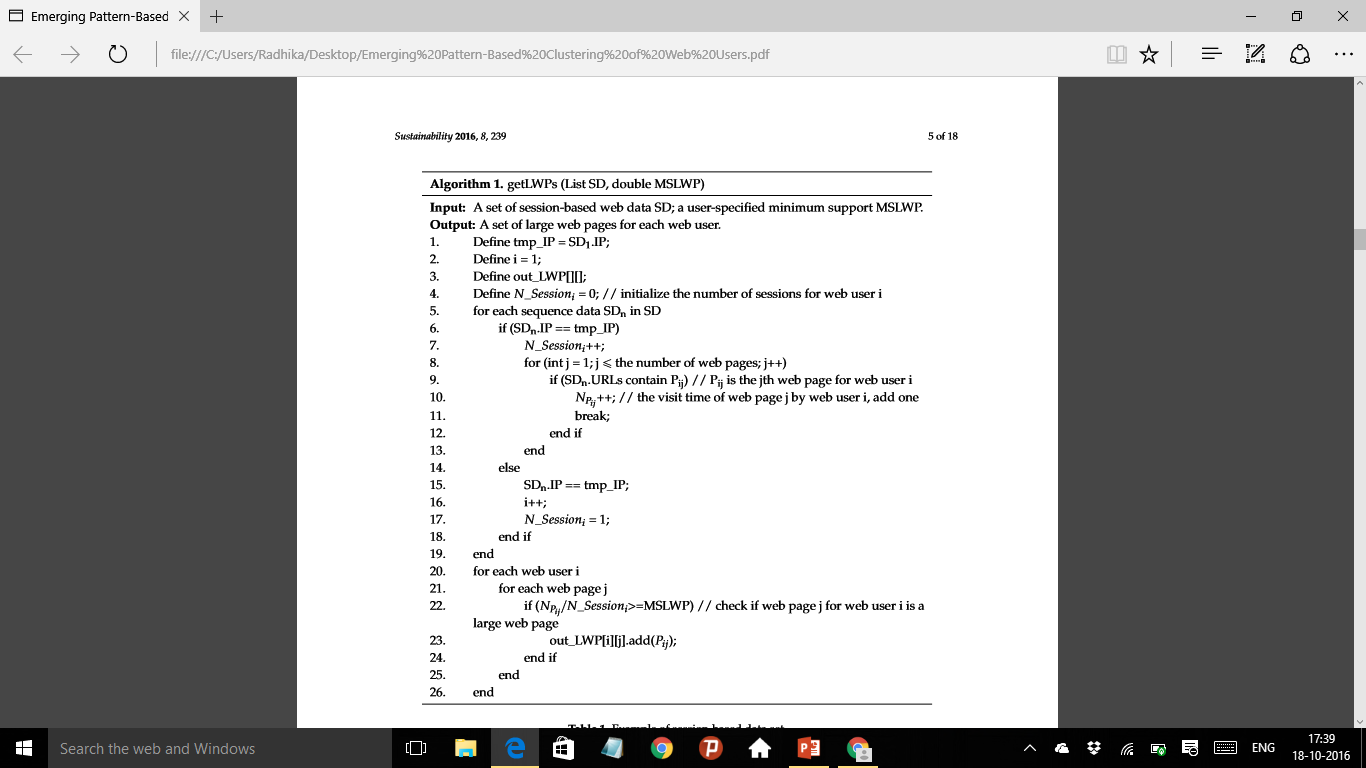
Considering the disadvantages in the proposed system, First the noise in the web log data will be removed. Then the frequent patterns are observed and then they are clustered based on the frequent patterns. The support value is calculated and then they are named based on the TF-IDF.

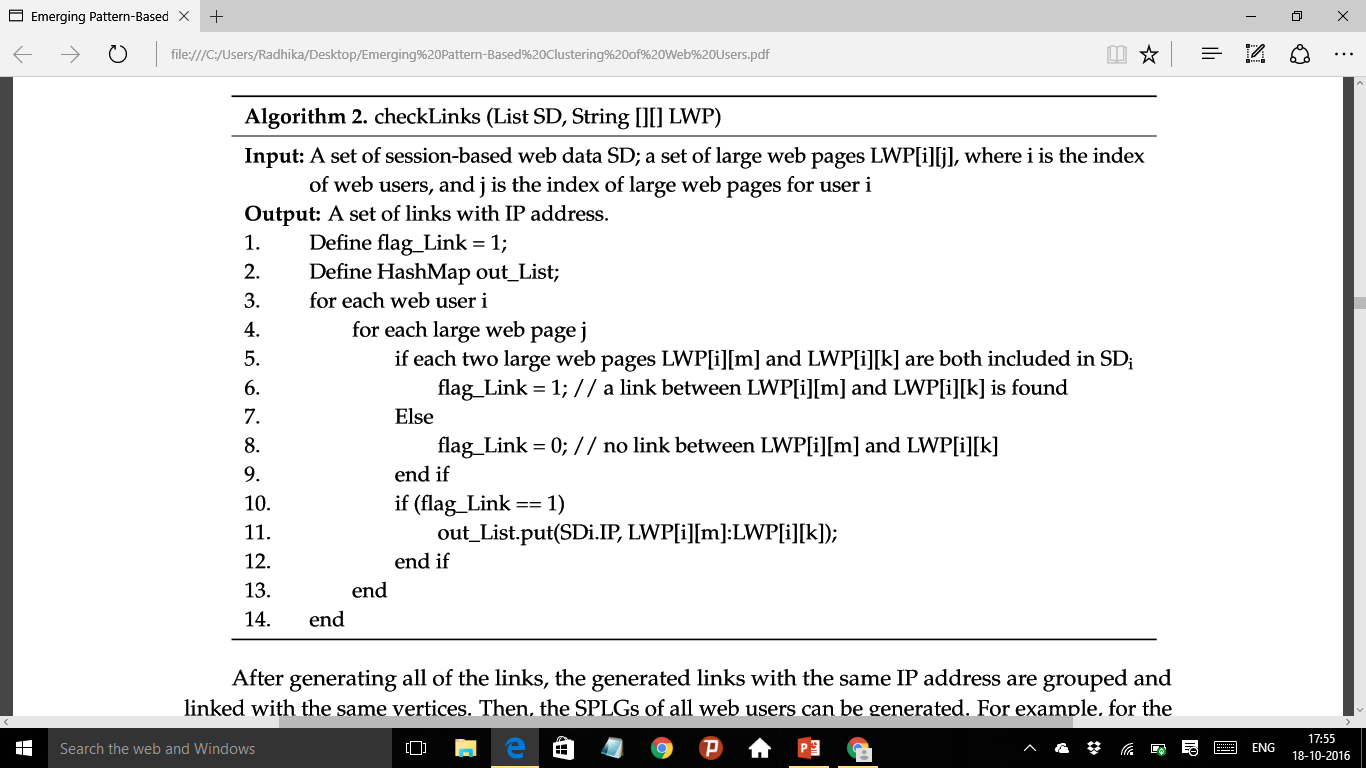
3.Analysis

3.1INTRODUCTION:

The project uses lot of software and hardware tools that are briefed in this section

* 1. SOFTWARE REQUIREMENT SPECIFICATION:
* WEKA software tool
* SMine algorithm
* Dataset
  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. ALGORITHMS AND FLOWCHARTS:

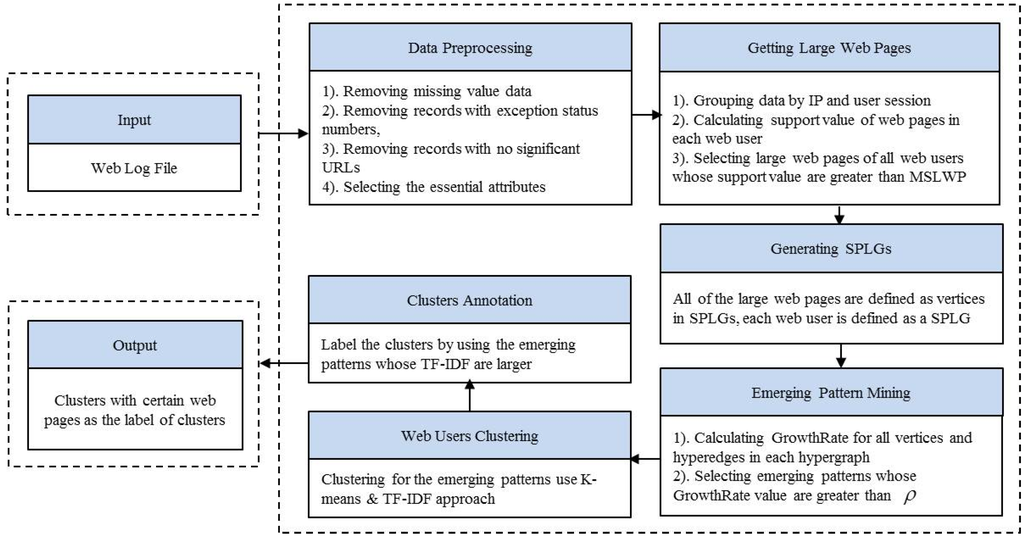




4.DESIGN

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

* 1. DFD DIAGRAM: 

4.1 Implementation

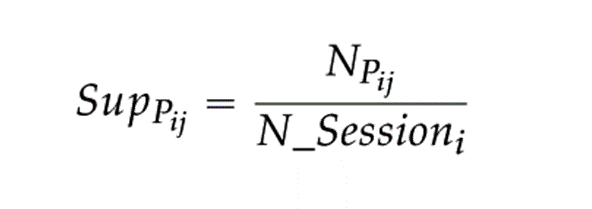
* 1. MODULE DESIGN AND ORGANIZATION:

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



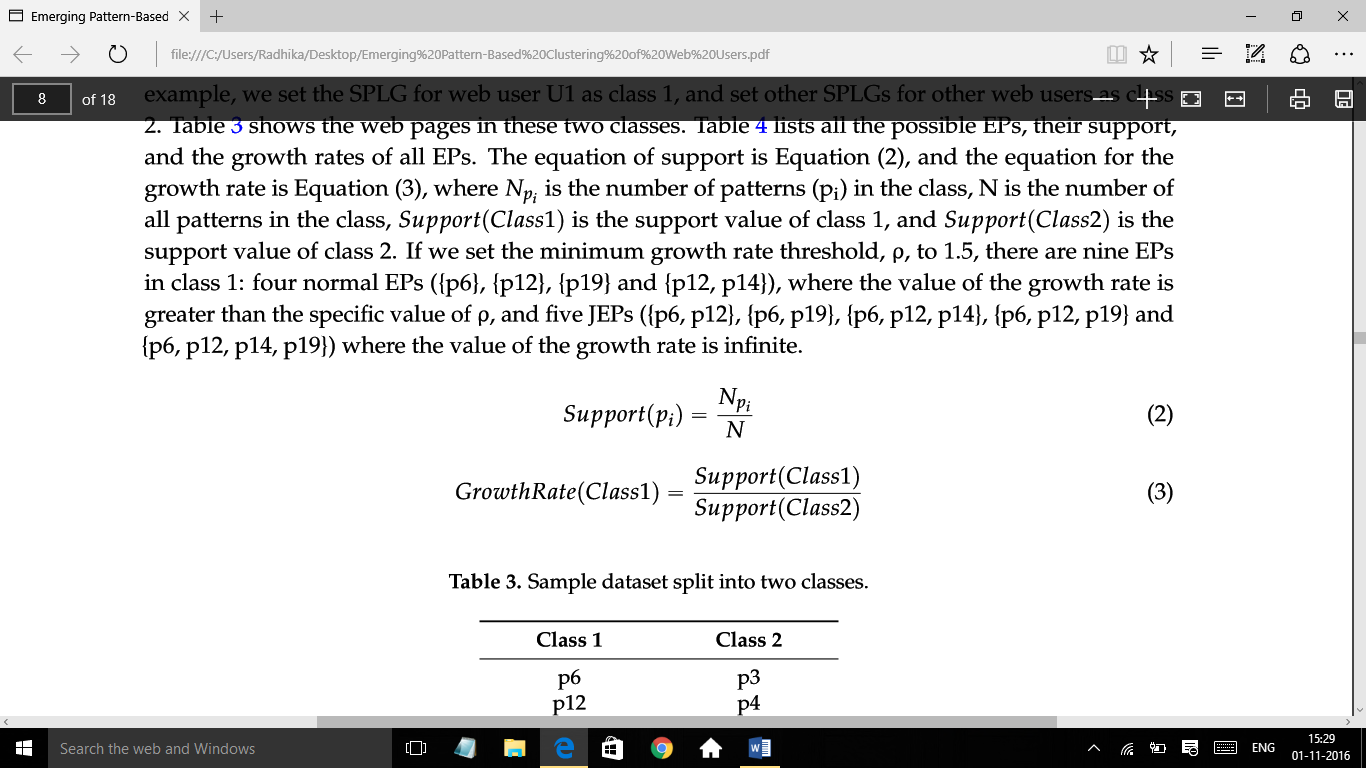
Where, Npij is the no of times a web page j is been visited in the web user i and N\_Sessioni is the no of sessions of web user i. Finally, a user specified Minimum Support threshold for Large Web Page (MSLWP)must be defined.  The MSLWP denotes a kind of abstract level that is a degree of generalization. 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

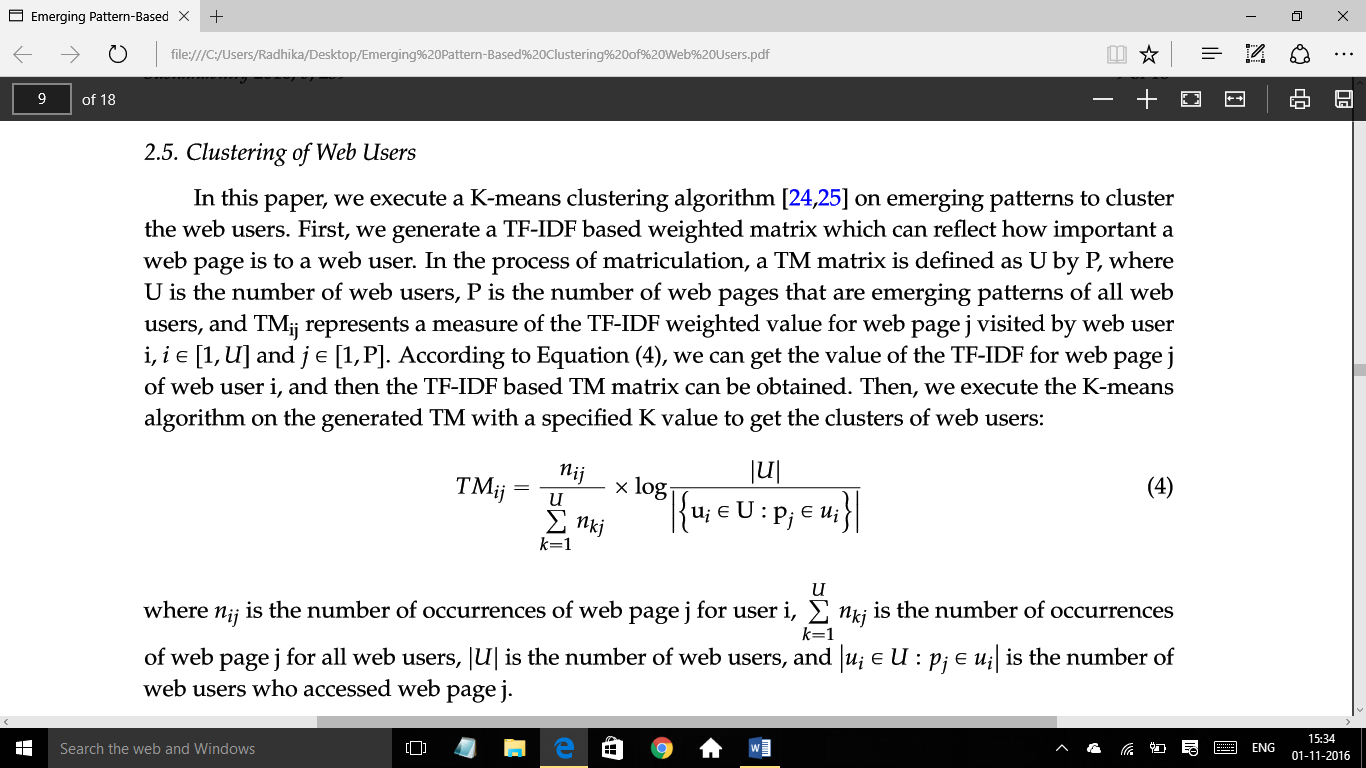
4.3.4 GENERATION OF EMERGING PATTERNS:

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.