Discovery of Patterns and evaluation of Clustering Algorithms in SocialNetwork Data (Face book 100 Universities) through Data Mining Techniques and Methods

Nancy.P[1], Dr.R. Geetha Ramani[2]

[1]Ph.D Research Scholar, Department of Computer Science and Engineering, Rajalakshmi Engineering College, Thandalam, Chennai, Tamilnadu, India.
nancysundar09@gmail.com
[2]Professor & Head, Department of Computer Science and Engineering, Rajalakshmi Engineering College, Thandalam, Chennai, Tamilnadu, India.
rgeetha@yahoo.com

ABSTRACT

Data mining involves the use of advanced data analysis tools to find out new, suitable patterns and project the relationship among the patterns which were not known prior. In data mining, association rule learning is a trendy and familiar method for ascertaining new relations between variables in large databases. One of the emerging research areas under Data mining is Social Networks. The objective of this paper focuses on the formulation of association rules using which decisions can be made for future Endeavour. This research applies Apriori Algorithm which is one of the classical algorithms for deriving association rules. The Algorithm is applied to Face book 100 university dataset which has originated from Adam D'Angelo of Face book. It contains self-defined characteristics of a person including variables like residence, year, and major, second major, gender, school. This paper to begin with the research uses only ten Universities and highlights the formation of association rules between the attributes or variables and explores the association rule between a course and gender, and discovers the influence of gender in studying a course. This paper attempts to cover the main algorithms used for clustering, with a brief and simple description of each.The previous research with this dataset has applied only regression models and this is the first time to apply association rules.

KEYWORDS

Data Mining, Social Networks, Face book, Association rules, Gender, Patterns.

1. INTRODUCTION

Data mining involves the use of sophisticated data analysis tools to discover previously unknown, valid patterns and relationships in large data sets [1]. Data mining allows finding the unprovoked from the voluminous data. Data mining can be performed on many different types of data like quantitative, textual, and even in multimedia forms. Data mining applications uses various ways of examination of the data which includes association, path analysis, classification, clustering, and forecasting. One of the emerging application areas in Data Mining is Social Networks.

Social Networks has become omnipresent in today's life. It gives way to share information between people anywhere and at anytime. Many Social Network sites (SNS) are now available like Orkut, Face Book, Twitter, Friendster, and MySpace. Face book is a social networking

service and website which was launched in February 2004, operated behind the closed doors of Face book, Inc. As of February 2012, Face book has more than 845 million active users [13].

This Research uses Face book 100 university dataset which defines various attributes like ID, Student/Faculty flag, Gender, Major, Second Major, Dorm /house/ Year and High school of 100 Universities. Our work focuses on Mining Association Patterns in only a subset of 100 Universities by randomly choosing ten Universities out of 100 and project the association between a major (course) and gender and to find out the higher percentage of gender participation in each of the courses offered by various Universities. The work also concentrates to evaluate the performance of different clustering algorithms on the 10 universities based on the accuracy in grouping the data in a gender specific way. This research work focuses on extracting patterns from the Face book 100 universities and projects gender specific association rules when applied to the dataset. It also evaluates the accuracy of various clustering algorithm in grouping the data based on gender.

The following section presents the past and current state of research in Mining Association rules and also the usage of Social Network.

2. RELATED WORK

Previous research on application of data mining techniques in Face book data is briefly summarized in the following paragraphs.

Traud (Traud et.al, 2010) examined the Face book network and identified the community structures of Face book networks within each of five American Universities [22].
Amanda (Amanda et.al, 2011) identified and compared the community structure of each network based on the given categorical data. They thereby compared and contrasted the organizations of the 100 different Face book networks [21].
Almahdi (Almahdi et.al, 2009) used a dataset which consist of data about Bachelor of Information Technology and Bachelor in Multimedia students of University Utara Malaysia from the year 2004 till 2005 and portrayed the process involved in extracting patterns using Apriori algorithm [3].
Agarwal (Agarwal et.al, 1994) implemented Apriori Algorithm to mine Single dimensional Boolean Association rules from transactional Databases [6].
Ma (Ma et.al, 2000) used association rules generated from Apriori Algorithm to find new patterns in hospital infection control [7].
Elena Zheleva (Elena et.al, 2009) discussed how an antagonist may foresee the undisclosed information of the users from an online social network [12].
Hetherly (Hetherly et.at, 2009) explained how deduction on attacks in social networking data to foresee undisclosed information of users through various techniques and projected the efficacy of the techniques through dataset obtained from the Dallas/Fort Worth, Texas network of the Face book social networking application [11].

2.1 Paper Organization

This paper is organized in the following manner. Section 3 portrays the proposed system design, clearly explaining each phase employed in the data mining process along with the dataset description. In Section 4, we discuss the Apriori algorithm applied for extraction of Association rules and some of the clustering algorithms while Section 5 portrays the Experimental Results and Section 6 concludes the paper.

3. PROPOSED DATAMINING FRAMEWORK

3.1 Overview

The proposed system design is diagrammatically presented in Figure 1. The original dataset involves the following process to be undergone.

1. Conversion of Mat file to Excel file
2. Discretization of Data
3. Selection of a subset which contains data of 10 Universities.
4. Application of Association Rule Algorithm(Apriori)
5. Extraction of Association Rules
6. Identification of Knowledge patterns.

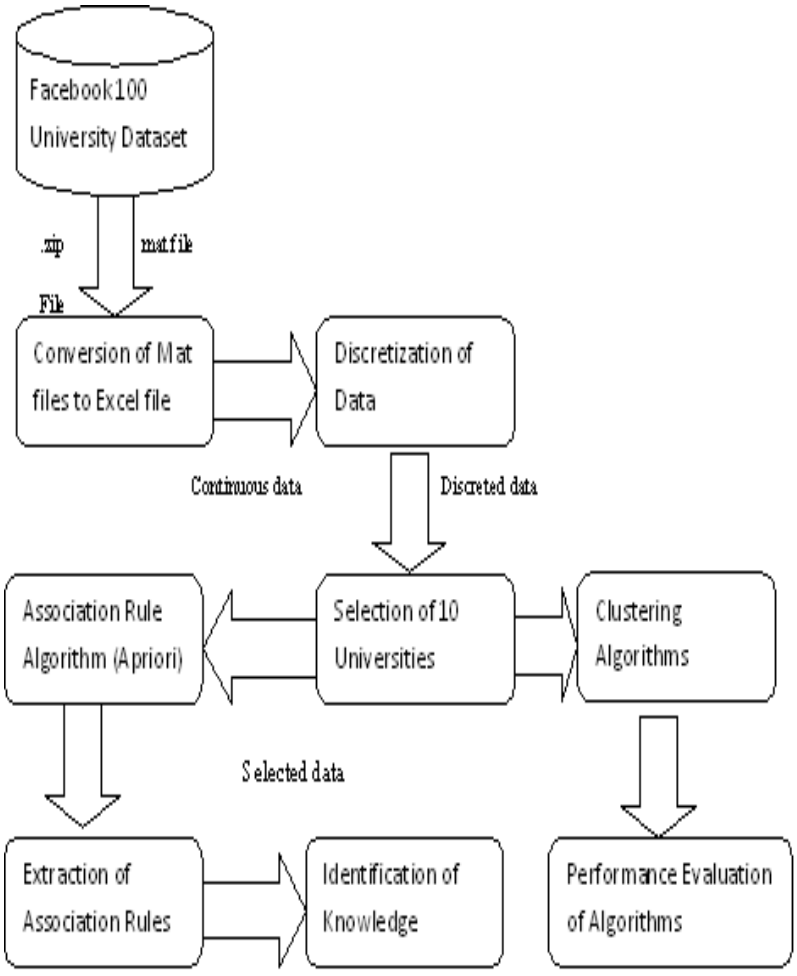


Figure 1. Proposed Data Mining Framework

3.2 Facebook 100 universities Data

The Original Data was a zip file. This .zip file contains the Face book networks (from a date in Sept. 2005) for 100 colleges and Universities. The following are the attributes in the file: ID, a student/faculty status flag, gender and major, second major/minor (if applicable), dorm/house, year, and high school. Missing data is coded 0. All the 100 University names in the original dataset are listed in Table 1.

Table 1. List of 100 Universities

Name of University	Name of University	Name of University	Name of University	Name of University
'Harvard'	'Michigan'	'Middlebury'	'Mach'	'Rutgers'
'Columbia'	'MSU'	'Hamilton'	'UCSC'	'Howard'
'Stanford'	'North western'	'Bowdoin'	'Indiana'	'Conn'
'Yale'	'UCLA'	'Vanderbilt'	'Vermont'	'UMass'
'Cornell'	'Emory'	'Carnegie'	'Auburn'	'Baylor'
'Dartmouth'	'UNC'	'UGA'	'USFCA'	'Penn'
'Upend'	'Tulane'	'USF'	'Wake'	'Tennessee'
'MIT'	'Chicago'	'UCF'	'Santa'	'Lehigh'
'NYU'	'Rice'	'FSU'	'American'	'Oklahoma'
'BU'	'Wash'	'GWU'	'Haverford'	'Reed'
'Brown'	'UC'	'Johns Hopkins'	'William'	'Brandeis'
'Princeton'	'UCSD'	'Syracuse'	'MU'	'Trinity'
'Berkeley'	'USC'	'Notre Dame'	'JMU'	'Wellesley'
'Duke'	'Caltech'	'Maryland'	'Texas'	'Oberlin'
'Georgetown'	'UCSB'	'Maine'	'Simmons'	'Mississippi'
'UVA'	'Rochester'	'Smith'	'Bingham'	'Colgate'
'BC'	'Becknell'	'UC'	'Temple'	
'Tufts'	'Williams'	'Villanova'	'Texas'	
'North eastern'	'Amherst'	'Virginia'	'Vassar'	
'Illinois'	'Swarthmore'	'UC'	'Pepperdine'	
'UF'	'Wesleyan'	'Cal'	'Wisconsin'	

The subset of ten Universities chosen for the research work is shown in Table 2. The selection was made in a random basis.

Table 2. List of ten Universities selected.

Name of University	
'American'	'Bingham'
'Amherst'	'Brandeis'
'Auburn'	'Brown'
'Baylor'	'Caltech'
'Berkeley'	'Columbia'

The data stored in .mat file is transferred to excel format. A sample excel sheet of one University named America is given in Table 3. The original data of the university contains 6386 records of which only fourteen is shown as a sample.

Table 3. Sample Data of America University

ID	Student/ Faculty	Gender	Major	Second Major	Dorm/ House	Year	High School
7400002	2	1	255	0	0	2005	17139
7400003	5	1	295	312	0	2006	16427
7400004	1	2	248	293	0	2006	17715
7400005	1	2	240	249	0	2006	17340
7400006	1	2	293	265	0	2007	5896
7400008	1	2	293	312	0	2006	51995
7400009	1	2	293	0	0	2007	50011
7400010	1	2	235	246	0	2005	26753
7400012	1	1	293	308	90	2008	17402
7400014	1	2	293	265	0	2007	21418
7400015	1	1	243	245	0	2006	9283
7400016	1	1	244	0	0	2007	9586
7400017	1	1	313	247	0	2006	16043
7400018	1	2	269	302	103	2007	52555

The data stored in .mat file is transferred to excel format. A sample excel sheet of one University named America is given in Table 3. The original data of the university contains 6386 records of which only fourteen is shown as a sample.

4. ALGORITHMS

4.1 Apriori Algorithm

The entire work uses the classical Apriority Algorithm for extracting the association rules. The problem of association rule mining is defined as: Let I= $\{i_1, i2, \ldots..i_n\}$ be a set of n binary attributes called items. Let D= $\{t_1, t2, \ldots., tm\}$ be a set of transactions called the database. Each

transaction in D has a unique transaction ID and contains a subset of the items in I. A rule is defined as an implication of the form X=>Y where X, Y $\subseteq$ I and X$\cap$ Y=ø. The sets of items X and Y are called antecedent (left hand side) and consequent (Right hand side) of the rule. To select interesting rules from the set of all possible rules, constraints on various measures of significance and interest can be used. The best-known constraints are minimum thresholds on **support and confidence** [19].

- The *support* supp(X) of an item set *X* is defined as the proportion of transactions in the data set which contain the item set
- The *confidence* of a rule is defined conf(X=>Y)=supp(X U Y) /supp(X)
- The *lift* of a rule is defined as lift(X=>Y)=supp(X U Y) / supp(Y) × supp(X) or the ratio of the observed support to that expected if X and Y were independent (Agarwal et.al)

The Pseudo code of Apriori Algorithm is given in Figure 2.

```
Ck: Candidate item set of size k
Lk : frequent item set of size k
L1 = {frequent items};
for (k = 1; Lk !=∅; k++) do begin
    Ck+1 = candidates generated from Lk;
   for each transaction t in database do
      increment the count of all candidates in Ck+1 that are contained in t
     Lk+1 = candidates in Ck+1 with min_support
   end
return ∪k Lk;
```

Figure 2**.** Pseudo code of Apriori Algorithm

Generation of Candidate

Suppose the items in L_{k-1} are listed in an order
Step 1: self-joining L_{k-1}
Insert into C_k
Select $p.item_1$, $p.item_2$, …, $p.item_{k-1}$, $q.item_{k-1}$
From L_{k-1} p, L_{k-1} q
Where $p.item_1=q.item_1$, …, $p.item_{k-2}=q.item_{k-2}$, $p.item_{k-1} < q.item_{k-1}$
Step 2: pruning
For all item sets c in C_k do
For all (k-1)-subsets s of c do
If *(s is not in L_{k-1})* then delete *c* from C_k

4.2 Clustering Algorithm

Cluster analysis or clustering is the task of assigning a set of objects into groups (called clusters) so that the objects in the same cluster are more similar (in some sense or another) to each other than to those in other clusters. Clustering is a main task of explorative data mining, and a common technique for statistical data analysis used in many fields, including machine learning, pattern recognition, image analysis, information retrieval, and bioinformatics [20].

Clustering Algorithms are broadly classified as follows

Connectivity Based Clustering (Hierarchical)
Density Based Clustering
Centroid Based Clustering (K-Means)
Distribution Based Clustering (EM)
Some of the above algorithm that is applied to the dataset is briefly narrated in this section.

4.2.1 K-Means Algorithm

In centroid-based clustering, clusters are represented by a central vector, which may not necessarily be a member of the data set. When the number of clusters is fixed to k, *k*-means clustering gives a formal definition as an optimization problem: find the k cluster centers and assign the objects to the nearest cluster center, such that the squared distances from the cluster are minimized [18]. The Pseudo code of the algorithm is shown in Figure 3.

X: a set of N data vectors	Data set
C_I: initialized k cluster centroids	Number of clusters,
C:the cluster centroids of k-clustering	random initial centroids

$P = \{p(i) \mid i = 1, \ldots, N\}$ is the cluster label of X

KMEANS(X, C_1) → (C, P)

REPEAT

$C_{previous} \leftarrow C_I$;

FOR all $i \in [1, N]$ DO Generate new optimal paritions

$p(i) \leftarrow \arg\min d(x_i, c_j)$;

$l \leq j \leq k$

FOR all $j \in [1, k]$ DO Generate optimal centroids

$c_j \leftarrow$ Average of x_i, whose $p(i) = j$;

UNTIL $C = C_{previous}$ [18]

Figure 3. Pseudocode of k-means Algorithm

4.2.2 Expectation Maximization (EM) Algorithm

EM algorithm is used to classify each point into the most likely Gaussian and estimate the parameters of each distribution. The Pesudo code is shown in Figure 4.

Set initial partition r_{ei} randomly.
repeat
Set the weight parameter w_{ei}
Call the mining algorithm to obtain F
Estimate θ_{ek} and θ_{ok} only for k € F (M-step)
Update the posterior r_{ei} (E-step)
Until the convergence [17]

Figure 4. Pesudocode of EM Algorithm

5. EXPERIMENTS AND RESULTS

The entire work is carried out using the Data Mining tool Tanagra1.4. All the data related to the selected 10 Universities are applied to Apriori Algorithm. The association rules are obtained for individual Universities and a sample with confidence 1 is shown in Figure 5.

Number of rules : 3

N°	Antecedent	Consequent	Lift	Support (%)	Confidence (%)
1	"maj=*339"	"gen=*m"	2.14149	0.010	100.000
2	"maj=*325"	"gen=*f"	1.87605	0.005	100.000
3	"maj=*394"	"gen=*f"	1.87605	0.010	100.000

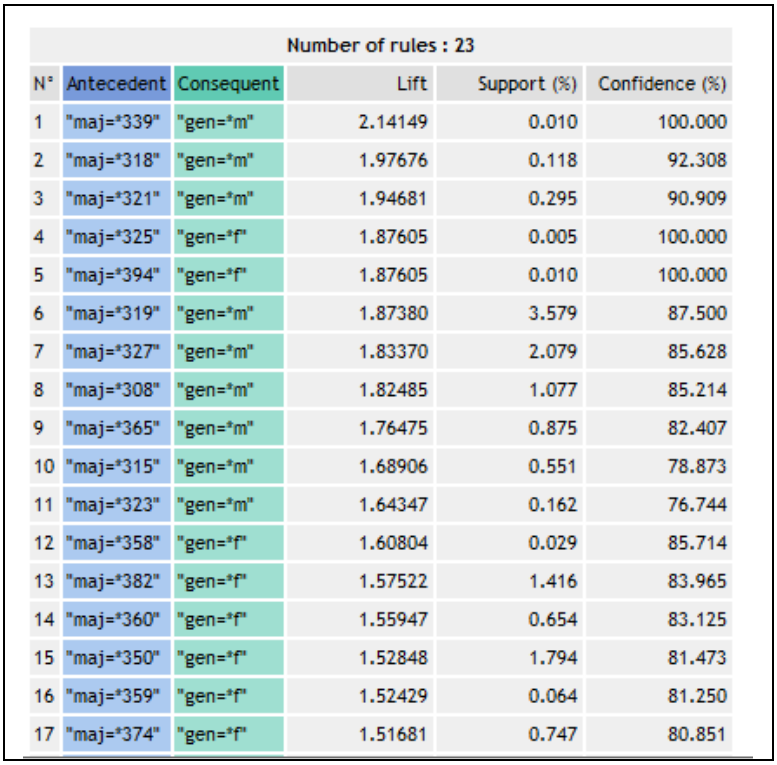
Figure 5. Association rules with confidence level 1 for Berkeley University

From the Figure 5 we can infer that majors (course) “325” and “394” are offered only to female and major “339” is offered to male. The value of support is calculated manually based on the number of records in each University and confidence level is set to 0.75 and 1.

The sample is with regard to Berkeley University. The confidence level was set to 1. Similarly when confidence was to 0.75 the number of rules increased and is shown in Figure 6. From the Dataset it is clear that Berkeley University is offering more than 100 courses, of which courses pertained to male and female can be found and also influence of gender in particular courses are also identified.

The patterns identified can be stated as

IF Major=”325” or Major=”394”
THEN Gender = “Female”
ELSEIF Major=”339”
THEN Gender= “Male”



Number of rules : 23

N°	Antecedent	Consequent	Lift	Support (%)	Confidence (%)
1	"maj=*339"	"gen=*m"	2.14149	0.010	100.000
2	"maj=*318"	"gen=*m"	1.97676	0.118	92.308
3	"maj=*321"	"gen=*m"	1.94681	0.295	90.909
4	"maj=*325"	"gen=*f"	1.87605	0.005	100.000
5	"maj=*394"	"gen=*f"	1.87605	0.010	100.000
6	"maj=*319"	"gen=*m"	1.87380	3.579	87.500
7	"maj=*327"	"gen=*m"	1.83370	2.079	85.628
8	"maj=*308"	"gen=*m"	1.82485	1.077	85.214
9	"maj=*365"	"gen=*m"	1.76475	0.875	82.407
10	"maj=*315"	"gen=*m"	1.68906	0.551	78.873
11	"maj=*323"	"gen=*m"	1.64347	0.162	76.744
12	"maj=*358"	"gen=*f"	1.60804	0.029	85.714
13	"maj=*382"	"gen=*f"	1.57522	1.416	83.965
14	"maj=*360"	"gen=*f"	1.55947	0.654	83.125
15	"maj=*350"	"gen=*f"	1.52848	1.794	81.473
16	"maj=*359"	"gen=*f"	1.52429	0.064	81.250
17	"maj=*374"	"gen=*f"	1.51681	0.747	80.851

Figure 6. Association rule with confidence level 0.75 for Berkeley University

From Figure 5, one can identify that more number of male candidates are studying the majors “318”,”321”,”319”,”327”,”308”,”365”,”315”,”323” and more number of female candidates are studying the majors “358”,”382”,”360”,”350”,”359”,”374” and for all other courses there is a complete blend of both male and female candidates. The results of applying Clustering algorithms to the Caltech University are shown in following figures.

Number of iterations: 6
Within cluster sum of squared errors: 266.8870209365126
Missing values globally replaced with mean/mode

Cluster centroids:

	Cluster#		
Attribute	Full Data	0	1
	(769)	(181)	(588)
stu/faculty	1.5059	2.3481	1.2466
major	188.2887	137.5304	203.9133
sec major	53.5007	28.8122	61.1003
dorm/house	131.078	8.3867	168.8452
year	1708.8205	953.337	1941.3759

Time taken to build model (full training data) : 0.02 seconds

Figure 7. Result of k-means algorithm

```
Class attribute: gender
Classes to Clusters:

  0    1    <-- assigned to cluster
117  358   | m
 51  177   | f
 60    6   | dk

Cluster 0 <-- dk
Cluster 1 <-- m

Incorrectly clustered instances :  351.0     45.6437 %
```

Figure 8. Result of Farthest Filter Algorithm

```
Class attribute: gender
Classes to Clusters:

  0    1    <-- assigned to cluster
 93  382   | m
 39  189   | f
 59    7   | dk

Cluster 0 <-- dk
Cluster 1 <-- m

Incorrectly clustered instances :  328.0     42.6528 %
```

Figure 9. Result of EM Algorithm

```
Class attribute: gender
Classes to Clusters:

  0    1    <-- assigned to cluster
 86  389   | m
 37  191   | f
 58  8 |     dk

Cluster 0 <-- dk
Cluster 1 <-- m

Incorrectly clustered instances :  322.0     41.8726 %
```

Figure 10. Result of Filtered Clustering Algorithm

A comparison of the accuracy rate of various Clustering algorithms of all the selected 10 universities are shown in Table 4.

Table 4. Accuracy rate of various Clustering Algorithms.

Name of the University	Name of the Clustering Algorithm applied				
	K-Means Alg	Make Density Based Alg	Farthest First Alg	EM Alg	Filtered Alg
'American'	59%	58%	66%	55%	59%
'Amherst'	53%	52%	62%	50%	54%
'Auburn'	56%	55%	61%	60%	57%
'Baylor'	64%	63%	58%	68%	66%
'Berkeley'	52%	50%	50%	58%	50%
'Bingham'	63%	60%	59%	69%	61%
'Brandeis'	58%	55%	59%	60%	56%
'Brown'	59%	54%	62%	64%	60%
'Caltech'	52%	57%	55%	58%	59%
'Columbia'	58%	51%	60%	60%	55%

6. CONCLUSION

In this paper Face book 100 dataset is considered that contains details of 100 Universities in Unites States, from which Association Rules are mined. As a preliminary stage, this research has considered only a subset with 10 Universities which was chosen randomly out of 1500 Universities. The Classical Algorithm for Mining Association Rule named Apriori is used in the dataset. The original Dataset undergoes various tasks like conversion to excel format, discretization of data to make it suitable for applying the data to Apriori Algorithm. Knowledge patterns regarding the association between the major (course) and gender were identified. The minimum threshold values for support is calculated manually based on the number of records in each of the University considered. The confidence value is set as 1 and 0.75 and numbers of rules were obtained. From the rules, one can clearly anticipate which major is offered only to male and female. Also one can foresee the affinity of gender influence in a particular course stating in which major a particular gender is dominating. Also various Clustering Algorithms were applied to the dataset and its performance was evaluated based on the accuracy level.

REFERENCES

[1] Two Crows Corporation, Introduction to Data Mining and Knowledge Discovery, Third Edition (Potomac, MD: Two Crows Corporation, 1999)

[2] Web Definition for Data Mining [http://searchcrm.techtarget.com/gDefinition/0, 294236, sid11_gci211901, 00.html].

[3] Almahdi Mohammed Ahmed , Norita Md Norwawi , Wan Hussain Wan Ishak Identifying Student and Organization Matching Pattern Using Apriori Algorithm for Practicum Placement 2009 International Conference on Electrical Engineering and Informatics 5-7 August 2009, Selangor, Malaysia

[4] Hipp, J., Guntzer, U., Gholamreza, N. (2000). Algorithm for Association Rule Mining: A general Survey and Comparison, ACM SIGKDD, volume 2 (Issue 1), p.58.

[5] Fayyad, U. M., Shapiro, G.P., Smyth, P., and Uthurusamy, R. (1996). Advances in Knowledge Discovery and Data Mining, Cambridge, AAAI/MIT press

[6] Agarwal, R., C.Faloutsos, and A.N.Swami (1994). Efficient similarity search in sequences databases in Proceedings of the 4th International Conference of Foundations of Data Organization and Algorithms (FODO), Chicago, Illinois pp, 9-84. Springer verlang

[7] Ma, Y., Liu, B., Wong, C.K., Yu, .S., &Lee, S.M. (2000) Targeting the Right Student Using Data Mining, ACM, PP. 457-463

[8] Ali, K., Manganaris, S. and Srikant, R. 1997. Partial classification using association rules. KDD-97,115-118

[9] Bayardo, R. J. 1997. Brute-force mining of high confidence classification rules. KDD-97, 123-126.

[10] Liu, B., Hsu, W., Ma, Y. (1998). Integrating Classification and Association Rule Mining, American Association for Artificial Intelligence

[11] Heatherly, R., Kantarcioglu, M., Thuraisingham, B., Lindamood, and J.: Preventing Private Information Inference Attacks on Social Networks. Tech. Rep.UTDCS-03-09, University of Texas at Dallas (2009)

[12] Zheleva, E., Getoor, L.: To join or not to join: the illusion of privacy in social networks with mixed public and private user profiles. In: WWW(2009)

[13] Facebook statistics, available at: http://www.Facebook.com/press/info.php?statistics

[14] Nancy.P, Dr.R.Geetha Ramani, Shomona Gracia Jacob, "Discovery of Gender Classification Rules for Social Network Data using Data Mining Algorithms",Proceedings of the IEEE International Conference on Computational Intelligence and Computing Research(ICCIC'2011), Kanyakumari, India, December 15-18, 2011, IEEE Catalog Number: CFP1120J-PRT, ISBN:978-1-61284-766-5, pp 808-812

[15] P. Nancy, Dr. R. Geetha Ramani: A Comparison on Performance of Data Mining Algorithms in Classification of Social Network Data", International Journal of Computer Applications (IJCA), 32(8): 47-54, October 2011. Published by Foundation of Computer Science, New York, USA.

[16] Shomona Gracia Jacob, Dr.R.Geetha Ramani, P.Nancy, "Feature Selection and Classification in Breast Cancer Datasets through Data Mining Algorithms", Proceedings of the IEEE International Conference on Computational Intelligence and Computing Research (ICCIC'2011), Kanyakumari, India, December 15-18, 2011, IEEE Catalog Number: CFP1120J-PRT, ISBN: 978-1-61284-766-5. Pp.661-667

[17] Koji Tsuda Taku Kudo, "Clustering Graphs by Weighted Substructure Mining" Proceeding ICML '06 Proceedings of the 23rd international conference on Machine learning Pages 953 – 960, 2006

[18] K-Means Clustering Algorithm,en.wikipedia.org/wiki/K-means_clustering.

[19] Association rule learning, available at http://en.wikipedia.org/wiki/Association_rule_learning

[20] Cluster Analysis available at http://en.wikipedia.org/wiki/Cluster_analysis

[21] Amanda. L.Traud,,Eric d.Kelsic, "Comparing Community Structure to Characteristics in Online Collegiate Social Networks in SIAW Review Vol 53.No 3 pp:536-543.

[22] Traud, ,Christina frost,Mason A.Porter,"visualization of communities in network" in CHOAS, 2009.