Cloud Based Big Data Application of FP-Growth Algorithm and K-Means Clustering Algorithm Based on MapReduce Hadoop

Than Htike Aung, Nang Saing Moon Kham
thanhtikeaung@ucsy.edu.mm
moonkhamucsy@ucsy.edu.mm

Abstract

In current time large volumes of data are being produced by various modern applications at an ever increasing rate. These applications range from wireless sensors networks to social networks. The automatic analysis of such huge data volume is a challenging task since a large amount of interesting knowledge can be extracted. Association rule mining is an exploratory data analysis method able to discover interesting and hidden correlations among data. Since this data mining process is characterized by computationally intensive tasks, efficient distributed approaches are needed to increase its scalability. This paper proposes a cloud-based service, named parallel FP-growth, to efficiently mine association rules on a distributed computing model. It consists of a series of distributed MapReduce jobs run in the cloud. Each job performs a different step in the association rule mining process, followed by cloud-based parallel k-means clustering algorithm to produce similar groups These outputs are verify and filter by three conditional levels which results is useful rules.

As a case study, the proposed approach has been applied to the educational data scenario.

Keywords- association rule mining, distributed computing model, cloud-based service, network data analysis.

1. Introduction

Cloud computing and Data Mining are both interesting hot topics nowadays. They bring convenience to diverse fields, including education. In general, clustering and prediction are two of the most remarkable features of data mining techniques. Unlike traditional analytical methods, data mining could offer more individual-oriented results. The application of data mining techniques in the education field enables numerous possibilities such as comprehensively analyzing the characteristics of each of students, predicting success in classes, pinpointing the gifted students and their learning paths [4] etc. In the higher education field, data mining applications have been highly suggested by many researchers such as C Romero and S Ventura in [5] and Luan in [9] to modify or design the curriculum to meet the different needs of students in terms of the learning abilities and knowledge construction. By these technologies, potentially valuable rules from educational data can be obtained for making decisions and strategies that can optimize the educational resource. In this project, the Jiaqu Yi et al.[8] proposed a cloud-based framework to generate the rules. Inside the framework, aparallel FP-growth association algorithm is adopted to generate useful rules among the students' grades, followed by reasonable analysis on the generated rules. All the analysis is based on learning skills identification for individual courses and a MapReduce -based K-means clustering algorithm on Hadoop. The experimental data come from the University of Computer Studies, Yangon in Myanmar (UCSY). The historical data can be increase as three Vs for educational schools. Thus the data are considered on the following conditions.

Volume: It is the most visible and major issue ofconcern referring to the fact that the amount ofgenerated data has increased tremendously in pastyears. The increase in internet users has increased inthe global data production. Organizations areoverflowed with data, unmanaged hundreds of terabytesand petabytes of information.

Variety: With the tremendous growth in datasources there are different types of data which needanalysis. Extends beyond structured data it includessemi-structured or unstructured data of all varieties, such as text, audio, video, web pages, log files andmore

Velocity: more and more data is generated and isprovided to the users immediately whenever required.This aspect captures the growing data rates. Millions ofconnected devices i.e. Smartphone, tablets etc.increases not only volume but velocity also. Data is therapid increase in rate of data transmission.

2. Related Work

There are many research papers about educational data mining, which are published in authoritative journals and conferences. For instance, Richard A. Huebner presents a survey of educational data mining research in [19]. In [19], Huebner describes how data mining can be utilized to analyze the data captured from course management systems. [21] Proposes a simple and efficient k-means clustering algorithm which requires a kd-tree as the only major data structure. Ramli, A.A adopts an Apriori algorithm to improve the content of learning portal [18]. Minaei-bidgoli, B Tan, P, Punch, W reveal interesting

ISBN 978-99971-0-381-9 © 2017 ICAIT

association rules among the attributes from students and problems in order to optimize online education systems [13]. Merceron, A, Yacef, K. utilize association rules to process learning data and find out whether students use resources to enhance grade and whether their use of such resources affects their grades [12].

3. Methodology

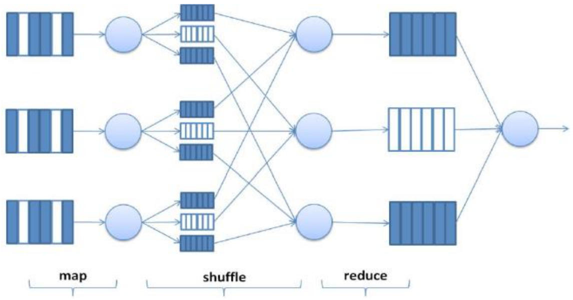


Figure 1.Typical composition of the MapReduce function.

MapReduce was originally designed and adopted by Google as a programming model for processing large data sets on a cluster with parallel processing over distributed storage.

The MapReduce paradigm now has become an industry standard and many platforms are internally built on this paradigm and support MapReduce implementation. Hadoop is an open source implementation that can be run either in-house or on cloud computing services with elastic MapReduce.

This has, at the core, the Map and Reduce functions that are capable of running in parallel across the nodes in the cluster. The Map function works on the distributed data and runs the required functionality in parallel, and the Reduce function runs a summary operation of the data.

In this section describe the data mining technique used algorithms which is MapReduce based parallel k-means algorithm and FP-Growth algorithm. These two algorithms used cloud-based framework. These two original algorithms can do in traditional data mining but also used for small and median data size.

3.1. ParallelFP-Growth on MapReduce

Nowadays, parallel distributed computing technology is quite mature, resulting in the emergence of cloud computing recent years. Cloud computing can take advantage of MapReduce and increase computing effectiveness by flexibly deploying online servers. However, the FP-Growth algorithm cannot be decomposed into multiple subtasks to facilitate distributed processes, and it also needs to scan databases twice in order to construct an FP-Tree. Xiaoting et al. [11] proposed a parallel version of FP-Growth called the Paralleled Incremental FP-Growth, (PIFP-Growth) algorithm, shown in Figure 2. In the proposed algorithm, MapReduce executes the FP-Growth algorithm each time new data arrive, which solves the problem of incremental database increases by dynamic threshold value comparison, and also avoids the double- computing problem. However the PIFP-Growth algorithm does not solve the problem faced by the FP-Growth algorithm of not being decomposable into multiple subtasks.

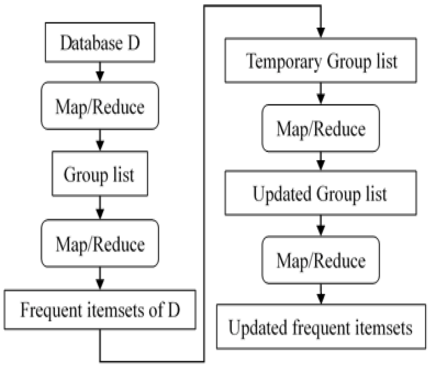
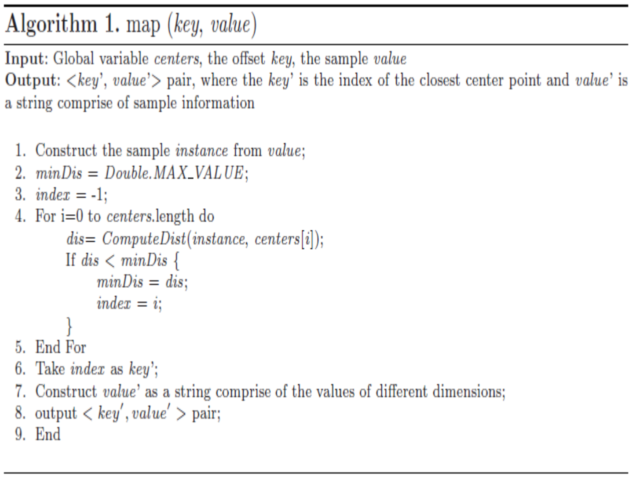


Figure 2.ThePIFP-Growthalgorithm

3.2. PKMeans Based on MapReduce

As the analysis above, PKMeans algorithm needs one kind of MapReduce job. The map function performs the procedure of assigning each sample to the closest center while the reduce function performs the procedure of updating the new centers. In order to decrease the cost of network communication, a combiner function is developed to deal with partial combination of the intermediate values with the same key within the same map task.

3.2.1. Map-function The input dataset is stored on HDFS[11] as a sequence file of <key, value> pairs, each of which represents a record in the dataset. The key is the offset in bytes of this record to the start point of the data file, and the value is a string of the content of this record. The dataset is split and globally broadcast to all mappers. Consequently, the distance computations are parallel executed. For each map task, PKMeans construct a global variant centers which is an array containing the information about centers of the clusters. Given the information, a mapper can compute the closest center point for each sample. The intermediate values are then composed of two parts: the index of the closest center point and the sample information. The pseudocode of map function is shown in Algorithm 1.



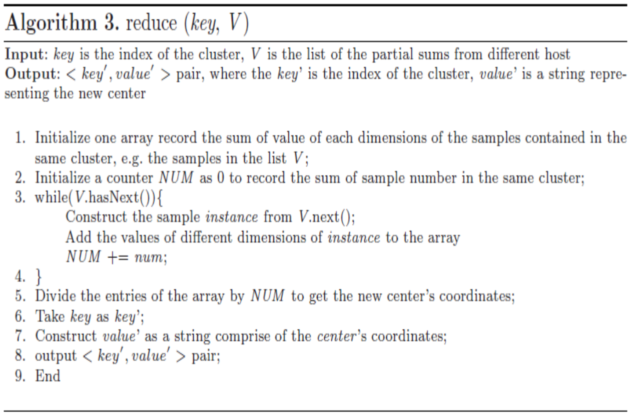
Algorithm 1. map (*key, value*)

Input: Global variable *centers*, the offset *key*, the sample *value*
Output: <*key*', *value*'> pair, where the *key*' is the index of the closest center point and *value*' is a string comprise of sample information

```
1. Construct the sample instance from value;
2. minDis = Double.MAX_VALUE;
3. index = -1;
4. For i=0 to centers.length do
       dis= ComputeDist(instance, centers[i]);
       If dis < minDis {
           minDis = dis;
           index = i;
       }
5. End For
6. Take index as key';
7. Construct value' as a string comprise of the values of different dimensions;
8. output < key', value' > pair;
9. End
```

Note that Step 2 and Step 3 initialize the auxiliary variable minDis and index ; Step 4 computes the closest center point from the sample, in which the function ComputeDist (instance, centers[i]) returns the distance between instance and the center point centers[i]; Step 8 outputs the intermediate data which is used in the subsequent procedures.

3.2.2. Combine-function. After each map task, we apply a combiner to combine the intermediate data of the same map task. Since the intermediate data is stored in local disk of the host, the procedure cannot consume the communication cost. In the combine function, we partially sum the values of the points assigned to the same cluster. In order to calculate the mean value of the objects for each cluster, we should record the number of samples in the same cluster in the same map task. The pseudocode for combine function is shown in Algorithm 2.

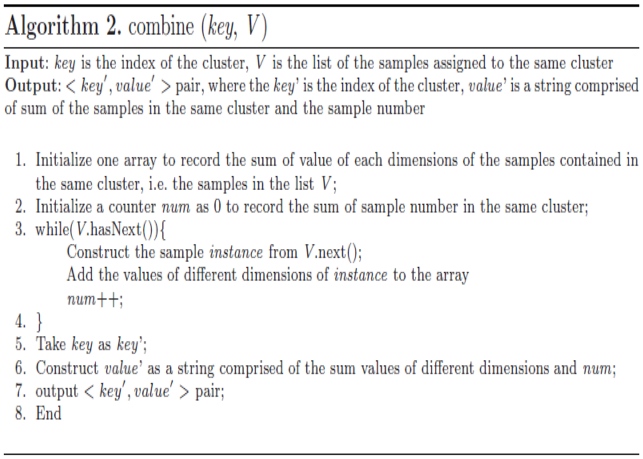


Algorithm 3. reduce (*key, V*)

Input: *key* is the index of the cluster, *V* is the list of the partial sums from different host
Output: < *key*', *value*' > pair, where the *key*' is the index of the cluster, *value*' is a string representing the new center

```
1. Initialize one array record the sum of value of each dimensions of the samples contained in the
   same cluster, e.g. the samples in the list V;
2. Initialize a counter NUM as 0 to record the sum of sample number in the same cluster;
3. while(V.hasNext()){
       Construct the sample instance from V.next();
       Add the values of different dimensions of instance to the array
       NUM += num;
4. }
5. Divide the entries of the array by NUM to get the new center's coordinates;
6. Take key as key';
7. Construct value' as a string comprise of the center's coordinates;
8. output < key', value' > pair;
9. End
```

3.2.3. Reduce-function. The input of the reduce function is the data obtained from the combine function of each host. As described in the combine function, the data includes partial sum of the samples in the same cluster and the sample number. In reduce function, we can sum all the samples and compute the total number of samples assigned to the same cluster. Therefore, we can get the new centers which are used for next iteration. The pseudocode for reduce function is shown in Algorithm3.



Algorithm 2. combine (*key, V*)

Input: *key* is the index of the cluster, *V* is the list of the samples assigned to the same cluster
Output: < *key*', *value*' > pair, where the *key*' is the index of the cluster, *value*' is a string comprised of sum of the samples in the same cluster and the sample number

```
1. Initialize one array to record the sum of value of each dimensions of the samples contained in
   the same cluster, i.e. the samples in the list V;
2. Initialize a counter num as 0 to record the sum of sample number in the same cluster;
3. while(V.hasNext()){
       Construct the sample instance from V.next();
       Add the values of different dimensions of instance to the array
       num++;
4. }
5. Take key as key';
6. Construct value' as a string comprised of the sum values of different dimensions and num;
7. output < key', value' > pair;
8. End
```

4. Learning Skill Test Metrics

Table 1 shows practical and tutorial scores in each subject. Three requirement levels indicate monthly test of score for students to be became skill full students in every academic subjects and courses. This is to measure the learning needs of the students based on the level of each subjects of practical /tutorial scores.

Table1. Details of practical and tutorial scores in each subjects

Requirement Level	Practical / Turorial Score
1	1-4
2	5-7
3	8-10

4.1. Experimental Result

Propose System design used cloud platform. Mining withParallel FP growth algorithm produced association rule from input data from cloud resources and then k-means clustering algorithm produced three clusters which is satisfy with learning skill test matrices in Table 1 . Finally rule verification and filtering on each cluster. The resulted data are useful rules shows in section 4.7.

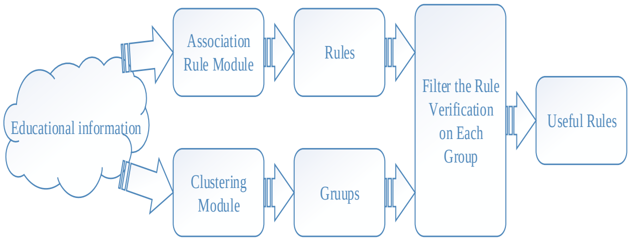
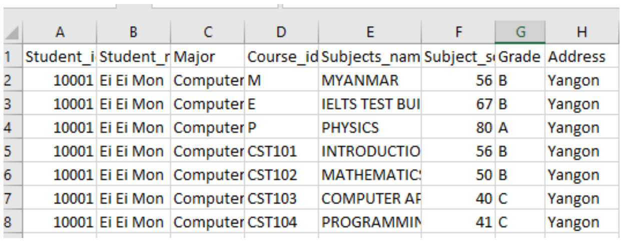


Figure 3.Research Framework

4.2. Raw Data

The raw data which contains students' grade from UCSY is shown in Figure 4. The data contains student ID, course and subject marks. All the observed students are from the same major. In total, there are 7000 lines of student information in the raw data, the five previous year data. All this data is stored in a CSV document.

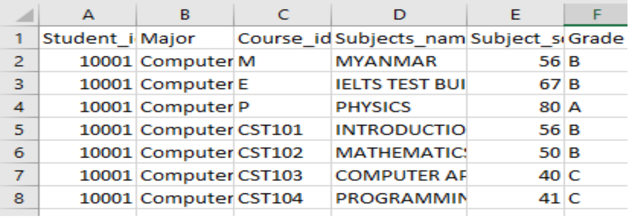


	A	B	C	D	E	F	G	H
1	Student_i	Student_r	Major	Course_id	Subjects_nam	Subject_s	Grade	Address
2	10001	Ei Ei Mon	Computer	M	MYANMAR	56	B	Yangon
3	10001	Ei Ei Mon	Computer	E	IELTS TEST BUI	67	B	Yangon
4	10001	Ei Ei Mon	Computer	P	PHYSICS	80	A	Yangon
5	10001	Ei Ei Mon	Computer	CST101	INTRODUCTIO	56	B	Yangon
6	10001	Ei Ei Mon	Computer	CST102	MATHEMATIC	50	B	Yangon
7	10001	Ei Ei Mon	Computer	CST103	COMPUTER AF	40	C	Yangon
8	10001	Ei Ei Mon	Computer	CST104	PROGRAMMIN	41	C	Yangon

Figure 4. The Raw Data

4.3. Data Pre-processing

In the raw data, some attributes such as gender and hometown are unnecessary, and some records are repetitive and redundant. Hence, the data needs to be cleaned up. And stored into MySQL After preprocessing, the clean data generated is shown in Figure 5.



	A	B	C	D	E	F
1	Student_i	Major	Course_id	Subjects_nam	Subject_s	Grade
2	10001	Computer	M	MYANMAR	56	B
3	10001	Computer	E	IELTS TEST BUI	67	B
4	10001	Computer	P	PHYSICS	80	A
5	10001	Computer	CST101	INTRODUCTIO	56	B
6	10001	Computer	CST102	MATHEMATIC	50	B
7	10001	Computer	CST103	COMPUTER AF	40	C
8	10001	Computer	CST104	PROGRAMMIN	41	C

Figure 5. The Clean Data

4.4.Data Mining

Figures stored all candidate 1-itemsets and record the support count of each 1-itemset. The minimum support condition is set to 0.1 which means 10% of the total amount (the total amount is more than 300). The minimum confidence condition is set to 0.7 which is the conditional probability between two frequent itemsets. A part of C1 and L1 generated by parallel FP-growth algorithm is displayed in Table 2. and Table 3. respectively. Those candidates whose support counts are lower than the support condition will be pruned.

Table 2. Candidate 1-Items C1

Itemset	Amount
CST103:ICS:A	0
CST103: ICS:B	24
CST103: ICS:C	72
CST103: ICS:D	75
CST103: ICS:E	8
CST104:PL:A	0
CST104:PL:B	8
CST104:PL:C	42
CST104:PL:D	38
CST104:PL:E	14
………..	…………..

Table 3. Frequent 1-Itemstes L1

Itemset	Amount
CST103: ICS:C	72
CST103: ICS:D	75
CST104:PL:C	42
CST104:PL:D	38
CST102:Mths:A	37
CST102:Mths:B	49
P:Phy:A	41
P:Phys:B	37
E:IELTS:B	59
E:IELTS:C	87
……..	………….

Table 4. Result of Association Rules

Rules	Confidence
CST102:Mths:C ⟹P:Phy:C	0.73
CS202:Mths:C⟹CS203:DS:C	0.8
CS202:Mths:C⟹203:DL:C	0.85
CST102:Mths:C⟹CST103:PL:C	0.73
CS206: SE:B⟹CS203:DS:C	0.81
CS204:SAD:B⟹CS201: JAVA:A	0.7
CS404:DBMS:B⟹CS405:UML:C	0.72
CS201:JAVA:A⟹CS203:DS:C	0.71
CS201:JAVA:A⟹CS202:Mths:C	0.71
CS304:UML:A⟹ CS305:CAT3:C	0.71
CS206:SE:B⟹CS204: SAD:B	0.76
CS304:DBMS:B:CS301:CO:C	0.7
..	………

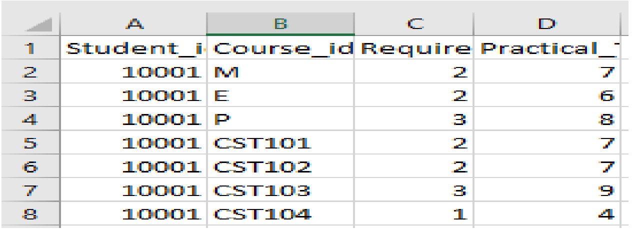
Significant trends can be seen from the results of FP-growth algorithm. For example, "CS204:SAD:B=> CS201:JAVA:A Confidence = 0.70 ", it can be said that if a student gets B level in SAD subject, he will excel in course CS201 , it means course CS204 may requires some similar skills as course CS201. In another word, students who are able to perform well in both course CS204 and CS201, certain parts of his/her learning skills are better than the rest of the students s' sample, vice versa. Another example, "304:DBMS:B =>301:APP:C Confidence=0.70", it tells us that a student gets C level in APP, he will probably get a bad level in course CS304:DBMS. That is very likely to say, there may be some similarity on the knowledge or skills requirement between course CS301 and course CS304. Students who cannot get a good level in both of the courses, he/she might be lacking of certain knowledge or learning skills, vice versa. According to above analysis, that can realize further.

4.5.Experimental Result of K-Means Clustering Algorithm

The learning skills test metrics presented in section 4 which are used by K-Means clustering algorithm. The courses are evaluated with this three-level test score based on the learning skill test matrices. Thus the output result produced three levels of test scores groups.

4.6. Experimental Raw Data and Results

An example of evaluated results is shown as Figure 6. The scores for the three levels range from 1 to 3. "1" indicates that this ability is highly desired, and "3" means that the corresponding ability is not required by this course. This data is input into "Clustering Module". Inside "Clustering Module" the courses are cluster via K-Means clustering algorithm. Courses which are similar according to the three level conditions will be grouped together. To demonstrate the results, the clustering result is shown visually as Figure 7. In Figure 7 is show that courses are divided into 3 groups/clusters by K-Means algorithm. In each cluster, courses are shown as a smallest circle with their course ID on it. The details in cluster 2 are shown in Figire.5. Among all these clusters, cluster 3 represents the courses which require all high level condition from those two kinds (2;3). Cluster 1 stands for the courses which not require high level condition (3).



	A	B	C	D
1	Student_i	Course_id	Require	Practical_
2	10001	M	2	7
3	10001	E	2	6
4	10001	P	3	8
5	10001	CST101	2	7
6	10001	CST102	2	7
7	10001	CST103	3	9
8	10001	CST104	1	4

Figure6. Example of Practical/Tutorial Test Score

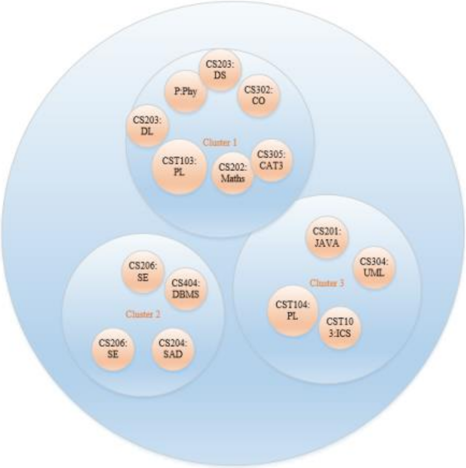


Figure7. Result of Clustering Module

In the following section that will analyze the results of "Rule Abstraction Module" by the results from "Clustering Module".

4.7. Rules Verification and Filtering

4.7. 1. Cluster 2 and Cluster 3

If a student got a B level in course from cluster 2, he/she was likely to get a A in most of the cluster 3 courses. Courses such as "JAVA", "SAD" etc. belong to cluster 3. Students who have level 2,3 condition he/she has good practice performance in their subjects such as JAVA and SAD. effected directly. Similar result can be found in the relationship between cluster 2 and cluster 3

4.7. 2. Cluster 1

If a student got a C in most of the cluster 1 courses, he/she also has level 1, at least who require strong level attitude level 2.

5. Conclusions

Data, classification and clusteringare an important task in parallel computing and distributed computing, when the processed dataset is large-scale, it becomes even more important. In this paperintroduce a parallel FP-Growth algorithm and K-Means clustering algorithms by using MapReduce and Hadoop, which improves the application of big data research in education field to produce more rules.The result also provides a good reference for education for students learning needs.

In particular, future works will aim at optimizing the MapReduce workflow and combiningthe workbench of the cluster and classification architecture.

5. References

[1] Abdullah Alshwaier , Ahmed Youssef and Ahmed Emam, “A Newtrend for E-Learning in Ksa Using Educationalclouds”, Advanced Computing : an International Journal, 2012,Vol.3(1), p.81

[2] Agrawal, R.; Imieli_ski, T.; Swami, A. (1993). "Mining association rules between sets of items in large databases". Proceedings of the 1993 ACM SIGMOD international conference on Management of data – SIGMOD '93. p. 207.

[3] American Bar Association Section of Legal Education and Admissions to the Bar, Legal Education and Professional Development –An Educational Continuum Report of The Task Force on Law Schools and the Profession: Narrowing the Gap, July, 1992.

[4] Bael, S., S. H. Hat, and S. C. Parka. "Identifying gifted students and their learning paths using data mining techniques." Data Mining in ELearning (Advances in Management Information) 4 (2006): 191-205.

[5] C Romero, S Ventura, Data mining in e-learning. WIT, 2006.

[6] David Chappell, (October 2008). " Introducing the Azure Services Platform An Early look at Windows Azure, .Net Services, SQL Services, And Live Services ". Chappell & Associates.157

[7] Fayyad, Usama; Piatetsky-Shapiro, Gregory; Smyth, Padhraic (1996)."From Data Mining to Knowledge Discovery in Databases". Retrieved 17 December 2008.

[8] Jiaqu Yi, Sizhe Li, Maomao, Wu, H.H. Au Yeung Wilton W.T Fok, Ying Wang, Fang Liu “Apriori algorithm and K-Means Clustering algorithm based on Students' Information”, 2014 IEEE Fourth International Conference on Big Data and Cloud Computing

[9] Luan, Jing. "Data mining and its applications in higher education." Newdirections for institutional research 2002.113 (2002): 17-36.

[10] MacKay, David (2003). "Chapter 20. An Example Inference Task: Clustering". Information Theory, Inference and Learning Algorithms. Cambridge University Press

[11] MacQueen, J. B. (1967). "Some Methods for classification and Analysis of Multivariate Observations". Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability 1. University of California Press.1-8.

[12] Merceron, A., Yacef, K. (2008). Interestingness Measures for Association Rules in Educational Data. In International Conference on Educational Data Mining, Montreal, Canada, 57-66.

[13] Minaei-bidgoli,B Tan, P., Punch, W. (2004). Mining interesting contrast rulesfor a web-based educational system. In International Conference on Machine Learning Applications, Los Angeles, USA.

[14] M.Lawanya Shri, Dr. S.Subha, “An Implementation of e-Learning System in Private Cloud”, International Journal of Engineering and Technology, 2013, Vol.5(3), p.3036

[15] Paul Pocatilu. “Cloud Computing Benefits for E-learning Solutions”.Oeconomics of Knowledge, 2010, Vol.2(1), p.9

[16] Peden, Elisabeth; Riley, Joellen, "Law Graduates Skills A Pilot Study into Employers Perspectives" [2005] LegEdRev 5; (2005) 15(1&2)Legal Education Review 87.

[17] Rakesh Agrawal and Ramakrishnan Srikant Fast algorithms for mining association rules in large databases. Proceedings of the 20th International Conference on Very Large Data Bases, VLDB, pages 487- 499, Santiago, Chile, September 1994]

[18] Ramli, A.A. (2005). Web usage mining using apriori algorithm: UUM learning care portal case. In International conference on knowledge management, Malaysia, 1-19.

[19] Richard A. Huebner, “A survey of educational data-mining research”,Research in Higher Education Journal, Retrieved 30 March 2014

[20] Tapas Kanungo, David M. Mount, Nathan S. Netanyahu, Christine D.Piatko, Ruth Silverman, Angela Y. Wu, “An Efficient k-MeansClustering Algorithm: Analysis and Implementation”, IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, VOL. 24, NO. 7, JULY 2002

[21] "The NIST Definition of Cloud Computing". National Institute ofStandards and Technology. Retrieved 24 July 2011.

[22] University of Computer Studies , Yangon in Myanmar http://www.ucsy.edu.mm

[23] http://en.wikipedia.org/wiki/Microsoft_Azure

[24] http://azure.microsoft.com/enus/documentation/articles/fundamentalsintroduction-to-azure/

[25] http://en.wikipedia.org/wiki/Platform_as_a_service

[26] http://en.wikipedia.org/wiki/IaaS#Infrastructure as_a_ service_.28IaaS.29