New Association Rule Mining Algorithms for Knowledge Discovery in MP Higher Education Datasets

Anil Rajput[1], Arun Sen[2]

Department of Mathematics and Computer Science, CSA Govt. PG Nodal College, Sehore[1].

Research Scholar, Barkatullah University, Bhopal[2]

dranilrajput@hotmail.com[1]

aruns1607@gmail.com[2]

Abstract -

The discovery of frequent article series is a key problem in important data mining applications, such as the discovery of association rules, schemas, relationships, etc. Apriori's algorithms work reasonably well when all frequent item sets are short, but tend to take much longer in large data sets. This paper reports a new implementation of a similar association mining algorithm. The new implementation can be programmed to discover patterns with different attribute as the point of interest. Thus we can have multiple analysis generating patterns around some specific parameter of analysis.

Keywords - Association Rule mining, Higher Education, Quality, GER, Education.

1.0 Introduction

The discovery of frequent article series is a key problem in important data mining applications, such as the discovery of association rules, schemas, relationships, etc. The main algorithms to solve this problem operate in a first search direction from the bottom up. The calculation starts with sets of frequent elements 1 (sets of frequent elements of minimum length) and continues until all the sets of maximum frequent elements (length) are found. During execution, each set of frequent elements is explicitly considered. Apriori's algorithms work reasonably well when all frequent item sets are short, but tend to take much longer in large data sets. Performance decreases theatrically when some of the maximum common element sets are relatively long. We have used a new algorithm that can create association rules around a specific attribute that we would like to focus on. The main mining direction continues to be bottom-up, but a limited search is also performed considering the presence of a specific attribute as the focal point of the model search. This search is only used to maintain and update the maximum set of frequent candidates containing the desired attribute. It is used to eliminate candidates in the bottom-up search. A very important feature of the algorithm is that it does not require an explicit examination of each set of frequent elements. Therefore, the algorithm works well even when some sets of frequent maximum elements are long. As a result, the algorithm produces the frequent maximum set, which is the set that contains all the sets of frequent maximum elements, which then immediately specifies all the frequent sets of elements.

Recently (Sen & Rajput 2018) in association rule mining and to identify the challenges in Higher Education Quality. This paper deals on The Association rules were mapped for understanding the patterns and perceptions of the eligible Higher Education prospects.

First Author (Rajput et. al. 2010) deals with analysis of Higher Education students' data for taking admission in various disciplines. For this purpose he used data from urban, semi - urban and rural area institutions.

New associations rule mining algorithm works similar to Apriori algorithm used world over for association rule mining. The key difference in this new implementation is that it can be programmed to discover patterns with different attribute as the point of interest. The user can decide the point of interest attribute and can even change it as and when required. Thus we can have multiple analysis generating patterns around some specific parameter of analysis. The rules generated are with respect to this parameter and present a new perspective of the same dataset. The algorithm builds association rules around that specific point of interest and iteratively condenses the minimum support while working for finding the required number of rules. These rules are worked up for the desired number of rules and the given minimum confidence which ever attains the threshold value earlier. This new algorithm is implemented as an addendum to the existing implementation of Apriori like algorithm implemented in WEKA Workbench.

Higher Education is the key factor in the growth strategy of any developing nation and has legitimately achieved an honorable place in society. (Planning Commission Govt of India, 2012) Higher Education is not only an instrument to improve efficiency, but also to improve the overall quality of the individual and society. (Mathew, 2014) A basic requirement for MHRD and Higher Education Institutions is the enhancement of Gross Enrolment Ratio (GER) by expanding access through all modes. The current rate of gross enrolment (GER) in higher education remains very low compared to the world average of 23.2% and is extremely low compared to the average of 54.6% for developed countries and 36.3% for developed countries. (McKinsey Global Institute , 2016)

The family background and occupational profile of the respondents was analyzed for better understanding the responses and association of various factors affecting enrolment. Financial position of the family and parents education was taken as the focal point of the current research paper for experimenting on the new algorithm that mine association rules through focused segmentation.

Data Analysis of research variables with Family- Financial- Position as the focus -variable shows interesting association of several factors on the family financial position and parental education level of the respondent.

2.0 Literature Review

2.1 Studies on Higher Education

Speaking of the role of higher education in the development of human resources, Odionye stated that tertiary education, being the agent at the center of attention in this context, has the function of guaranteeing the development of human resources.

The objectives of tertiary education as stipulated in the document are: (Odionye, 2014)

- Contribute to national development through the formation of high-level work.
- Develop people's intellectual ability to understand and appreciate their local and external capabilities environment.
- Acquire physical and intellectual abilities that enable the individual to be self-sufficient and useful members of the society.

Kassu Mehari and Jemal Ayalew in their paper Multilevel Analysis for Identifying Factors Influencing Academic Achievement of Students in Higher Education Institution: The Case of Wollo University identified several important factors related to higher studies. (Mehari & Ayalew, 2016)

Mothers and father educational level are the two most influential factors for academic achievement of students in Wollo University. Mothers and fathers educational level have positively and strongly associated with student's academic achievement. (Mehari & Ayalew, 2016) Parents provide higher levels of psychological support for their children through environments that encourage the development of skills necessary for success at school. (Esnault, 1992) Infrastructures are necessary to meet the growing number of children in primary schools. The provision of these facilities should not be just for number purposes, but quality should be considered. (Chand, 2013)

Modern learning environments are characterized by their independence from place and time, their integrated communication and presentation structures and their opportunities for the re-use of learning technologies in the form of learning objects. Many researchers say that the technological impulse will improve the quality of education. (Ray, 2007)

2.2 Studies on Association Rule mining

Progress in bar-code technology has made it possible for retail organizations to collect and store massive amounts of sales data, referred to as the basket data. A record in such data typically consists of the transaction date and the items bought in the transaction. Successful organizations view such databases as important pieces of the marketing infrastructure. (Han & Kamber, 2001) They are interested in instituting information-driven marketing processes, managed by database technology, that enable marketers to develop and implement customized marketing programs and strategies. Algorithms for discovering large itemsets make multiple passes over the data. In the first pass, we count the support of individual items and determine which of them are large, i.e. have minimum support. In each subsequent pass, we start with a seed set of itemsets found to be large in the previous pass. We use this seed set for generating new potentially large itemsets, called candidate itemsets, and count the actual support for these candidate itemsets during the pass over the data. At the end of the pass, we determine which of the candidate itemsets are actually large, and they become the seed for the next pass. This process continues until no new large itemsets are found. (Agrawal & Srikant, 1994)

3.0 Research Methodology

A survey of 2100+ youth in the age group 17 to 23 was conducted for getting an understanding of the factors affecting enrolment in Higher Education. As a research tool a research questionnaire was designed and administered. The purpose of this questionnaire is to understand the factors affecting enrolment in Higher Education. The objective was to identify the challenges in Higher Education Enrolment and the perceptions of the eligible Higher Education prospects. Understanding the groups' perceptions about importance, accessibility and affordability of higher education will contribute to higher education professionals in understanding how this population sees higher education. The Association rules were mapped for understanding the patterns and perceptions of the eligible Higher Education prospects. In the light of these perceptions we next designed an algorithm for Incremental mining of association rules to identify the patterns in Madhya Pradesh UG and PG dataset. The discoveries of this study are prized because they expand on the research on college access and enrolment which eventually contributes to GER. (Association of Indian Universities AIU, 2017)

The questionnaire designed for this purpose was administered through friends, relatives, personally and through online digital media. This data was tabulated, classified, analysed through visual tools, pivot charts, and extensive data mining was conducted through WEKA workbench. To help facilitate thematic coding and categorization of collected data, I used weka data mining software for implementing the new algorithm and discovering patterns and associations rules based analysis. For this survey a semi-structured research method was used. It allowed us to analyze the depth and feeling with respect to experiences and thoughts related to higher education.

4.0 Experiments Findings and Discussion

Association rules were mined on the collected research dataset. The initial control group experiment results are with base Apriori algorithm. Subsequently the new algorithm variation of Apriori was exercised and the experiment results were recorded with financial background of family as the focus variable. The algorithm generated new patterns and interesting association rules different from the previous ones. As the second experiment we chose the "parents education" of the respondent as the analysis viewpoint. The association mining results provided a new set of insights about the Madhya Pradesh Higher Education dataset.

4.1 Association Rules using simple Apriori Algorithm

Best rules found:

1. Finance=Important 1302 ==> ICT_Infra=Important 1302 conf:(1)
2. Infrastructure=very 1225 ==> Access=very 1225 conf:(1)
3. Syllabus=very 1225 ==> Access=very 1225 conf:(1)
4. Syllabus=very 1225 ==> Infrastructure=very 1225 conf:(1)
5. Infrastructure=very 1225 ==> Syllabus=very 1225 conf:(1)
6. Infrastructure=very Syllabus=very 1225 ==> Access=very 1225 conf:(1)
7. Access=very Syllabus=very 1225 ==> Infrastructure=very 1225 conf:(1)
8. Access=very Infrastructure=very 1225 ==> Syllabus=very 1225 conf:(1)
9. Syllabus=very 1225 ==> Access=very Infrastructure=very 1225 conf:(1)
10. Infrastructure=very 1225 ==> Access=very Syllabus=very 1225 conf:(1)
11. Rigid_sys=very 1218 ==> Family_support=very 1218 conf:(1)
12. Family_support=very 1218 ==> Rigid_sys=very 1218 conf:(1)
13. Emlpoyability=StrAgree 1197 ==> Gender_impact=StrAgree 1197 conf:(1)
14. Gender_impact=StrAgree 1197 ==> Emlpoyablity =StrAgree 1197 conf:(1)
15. Importance=very Rigid_sys=very 1155 ==> Family_support=very 1155 conf:(1)
16. Importance=very Family_support=very 1155 ==> Rigid_sys=very 1155 conf:(1)
17. ICT_Infra=Important 1369 ==> Finance=Important 1302 conf:(0.95)
18. Importance=very 1239 ==> Access=very 1176 conf:(0.95)
19. Family_support=very 1218 ==> Importance=very 1155 conf:(0.95)
20. Rigid_sys=very 1218 ==> Importance=very 1155 conf:(0.95)
21. Family_support=very Rigid_sys=very 1218 ==> Importance=very 1155 conf:(0.95)

Analysis of discovered Rules

The findings evident from this data mining using Apriori algorithm in WEKA workbench are:

- ✓ The importance of financial position of family in enrolment decision had a vivid association with availability of ICT infrastructure in the higher education institution. The insight delivered is people who value the financial structure of family and resources available with the family check and consider the ICT infrastructure as important.
- ✓ Infrastructure is demonstrating association with access variable. Candidates considering access as important consider infrastructure as an important decision variable.
- ✓ Not only is the infrastructure very important but also the syllabus is equally important in enrolment decision. A vice-versa association of the two variables is also true
- ✓ Infrastructure and syllabus combo is showing association with access
- ✓ 8 out of the first 10 best rules are revolving around infrastructure, access and syllabus

- ✓ The Impact of Rigid Systems, Inflexible Study Hours, Attendance Compulsion is associated to Family support vs negative pressure (Eg: In-Laws pressure on females) variable.
- ✓ Gender impact is strongly associated to employability variable
- ✓ Another strong association pattern was the perceived importance of Higher education and the Impact of Rigid Systems, Inflexible Study Hours, Attendance Compulsion is associated to Family support vs negative pressure (Eg: In-Laws pressure on females) variable. This is the segment which perceives higher education as important but could not avail enrolment due to rigid systems and family pressures.
- ✓ Respondents who perceived access as very important also perceived higher education as very important.
- ✓ The obvious insight visible in this set of rules is people perceive higher education as important but are constrained by Family finances, access, rigid systems and family support.

4.2 Association Rules with Financial status of the family as the focus variable.

The focus issue in this section of research was the perceived importance of financial status of the family in the enrolment decision. The best rules found were:

1. Importance=very ICT_Infra=Important 868 ==> Finance=Important 868 conf:(1)
2. Effectiveness=very ICT_Infra=Important 868 ==> Finance=Important 868 conf:(1)
3. Access=very ICT_Infra=Important 916 ==> Finance=Important 910 conf:(0.99)
4. Infrastructure=very ICT_Infra=Important 867 ==> Finance=Important 861 conf:(0.99)
5. ICT_Infra=Important Syllabus=very 867 ==> Finance=Important 861 conf:(0.99)
6. Access=very Infrastructure=very ICT_Infra=Important 867 ==> Finance=Important 861 conf:(0.99)
7. Access=very ICT_Infra=Important Syllabus=very 867 ==> Finance=Important 861 conf:(0.99)
8. Infrastructure=very ICT_Infra=Important Syllabus=very 867 ==> Finance=Important 861 conf:(0.99)
9. Access=very Infrastructure=very ICT_Infra=Important Syllabus=very 867 ==> Finance=Important 861 conf:(0.99)
10. Teaching_infra=Important ICT_Infra=Important Scholarship=Important 887 ==> Finance=Important 872 conf:(0.98)
11. ICT_Infra=Important Fees=Important Scholarship=Important 887 ==> Finance=Important 872 conf:(0.98)
12. Teaching_infra=Important ICT_Infra=Important Fees=Important Scholarship=Important 887 ==> Finance=Important 872 conf:(0.98)
13. Teaching_infra=Important ICT_Infra=Important 1060 ==> Finance=Important 1036 conf:(0.98)
14. ICT_Infra=Important Fees=Important 1060 ==> Finance=Important 1036 conf:(0.98)
15. Teaching_infra=Important ICT_Infra=Important Fees=Important 1060 ==> Finance=Important 1036 conf:(0.98)
16. Teaching_infra=Important Scholarship=Important 908 ==> Finance=Important 872 conf:(0.96)
17. Fees=Important Scholarship=Important 908 ==> Finance=Important 872 conf:(0.96)
18. Teaching_infra=Important Fees=Important Scholarship=Important 908 ==> Finance=Important 872 conf:(0.96)
19. ICT_Infra=Important 1369 ==> Finance=Important 1302 conf:(0.95)
20. ICT_Infra=Important Scholarship=Important 1010 ==> Finance=Important 955 conf:(0.95)

21. Teaching_infra=Important 1099 ==> Finance=Important 1036 conf:(0.94)

Analysis of discovered Rules

- ✓ The respondents who perceived financial status of the family as an important decision making variable in HE enrolment decision perceived Higher Education as 'very important' and ICT Infrastructure as important.
- ✓ The next association was Financial status of family perceived as important and Higher Education perceived as very effective combined with ICT infrastructure as important.
- ✓ Access was the next factor that was associated as important with ICT Infrastructure and Financial status but the confidence reduced to (0.99) for this rule
- ✓ The general Infrastructure of the college in terms of land, buildings, physical facilities, furniture, fixtures, lab equipment etc was the next consideration
- ✓ Next was Teaching infrastructure as Important ICT Infrastructure as Important Scholarship as Important in association with Finance as Important with a confidence level of (0.98)
- ✓ Next was the association of ICT infrastructure, Fees and scholarship as important along with finance as important. This interesting combination of fees and scholarship appeared somewhere in the middle after first ten rules. Some previous research papers have reported that the most "interesting" of the rules are visible in the "central section" of the rules as the earlier ones are "very obvious rules."
- ✓ Another important combination was, "Teaching infrastructure, ICT Infrastructure Fees and Scholarship" in association with Finance=Important. This presents a combination insight that can be used to improve HE enrolments
- ✓ The next 7 rules were some similar combinations of Teaching infrastructure with either ICT infrastructure or fees or scholarship.
- ✓ The analysis mentioned above also presents insights to sequence of thoughts in enrolment journey.
- ✓ An association of teaching infrastructure with Fees was reported with (0.96) confidence
- ✓ Fees with scholarship combo in association with Financial status as important was again with 0.96 confidence
- ✓ The subsequent rules connected the variables mentioned in analysis above but with reducing level of confidence
- ✓ In all this was a very new set of insights which as not visible with the previous set of experiment results.

4.3 Association rules with Parents Educational Level as focus

To check the functioning of the new algorithm the next focus variable was set as parent's educational level and its importance in enrolment decision.

Impact Factor	Count of Parent_Edu
Important	889
not	182
rather	126
somewhat	56
very	847
Grand Total	2100

Table 1: Impact of Parents Educational Level in Enrolment Decision

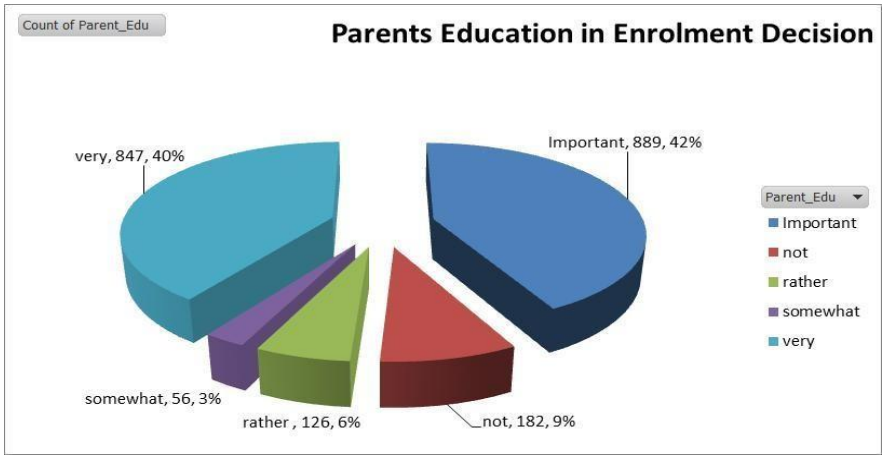

Figure 1: Impact of Parents Educational Level in Enrolment Decision

A vast majority of respondents perceived the role of parents education in enrolment decision as very important (42%) or important (40%) . The best association rules found using the new algorithm are:

1. Library_lab=Important 889 ==> Parent_Edu=Important 889 conf:(1)
2. Library_lab=very 847 ==> Parent_Edu=very 847 conf:(1)
3. Family_support=very Library_lab=very 693 ==> Parent_Edu=very 693 conf:(1)
4. Library_lab=very Rigid_sys=very 693 ==> Parent_Edu=very 693 conf:(1)
5. Family_support=very Library_lab=very Rigid_sys=very 693 ==> Parent_Edu=very 693 conf:(1)
6. Access=very Library_lab=very 686 ==> Parent_Edu=very 686 conf:(1)
7. Infrastructure=very Library_lab=very 665 ==> Parent_Edu=very 665 conf:(1)
8. Library_lab=very Syllabus=very 665 ==> Parent_Edu=very 665 conf:(1)
9. Access=very Infrastructure=very Library_lab=very 665 ==> Parent_Edu=very 665 conf:(1)
10. Access=very Library_lab=very Syllabus=very 665 ==> Parent_Edu=very 665 conf:(1)
11. Infrastructure=very Library_lab=very Syllabus=very 665 ==> Parent_Edu=very 665 conf:(1)
12. Access=very Infrastructure=very Library_lab=very Syllabus=very 665 ==> Parent_Edu=very 665 conf:(1)
13. Importance=very Library_lab=very 658 ==> Parent_Edu=very 658 conf:(1)
14. Importance=very Access=very Library_lab=very 658 ==> Parent_Edu=very 658 conf:(1)
15. Importance=very Family_support=very Library_lab=very 658 ==> Parent_Edu=very 658 conf:(1)
16. Importance=very Library_lab=very Rigid_sys=very 658 ==> Parent_Edu=very 658 conf:(1)
17. Access=very Family_support=very Library_lab=very 658 ==> Parent_Edu=very 658 conf:(1)
18. Access=very Library_lab=very Rigid_sys=very 658 ==> Parent_Edu=very 658 conf:(1)
19. Importance=very Access=very Family_support=very Library_lab=very 658 ==> Parent_Edu=very 658 conf:(1)
20. Importance=very Access=very Library_lab=very Rigid_sys=very 658 ==> Parent_Edu=very 658 conf:(1)

21. Importance=very Family_support=very Library_lab=very Rigid_sys=very 658 ==> Parent_Edu=very 658 conf:(1)

Analysis of discovered Rules

- ✓ Respondents who perceived parents education as important perceived library facilities and lab facilities as important
- ✓ All 847 respondents who reported parents education as very important also marked Library books, journals, lab equipment, Sports facilities etc as very important. (confidence level 1.0)
- ✓ The next association rule is Family support = very important, Library books, journals, lab equipment, Sports facilities etc as very important (693respondents) ==> Parent Education = very important (693 respondents in same transaction records, hence confidence level for the association rule is 1)
- ✓ The respondents next associated the subtle connection of parents education as important with family support, library and lab infrastructure, and Rigid Educational Systems, Inflexible study hours /compulsory attendance.
- ✓ The importance of access to higher education in combination with good library and lab facility was associated to parents education level.
- ✓ Good physical and civil infrastructure at college / university along with Syllabus / course content and evaluation methods was the next association.
- ✓ The respondents expressing parents education as very important acknowledged higher education as very important.
- ✓ An extension rule to this rule was visible in terms of family support, library lab etc, and System rigidity
- ✓ In all the rules were very different from the earlier set of experiments and proved the metal of the new algorithm.

5.0 Conclusions and Recommendations

Indian higher education machinery needs strong improvements in access to higher education and large-scale technical and professional education. There have been formidable national challenges for expanding and improving higher education, taking into account the inter-regional, rural-urban and male-female differentials. We need to make higher education more relevant to global needs and eliminating inequalities in access to education. No talent should be wasted on grounds of non-availability of right education. The youth from households with vulnerable sources of income cannot afford to go for higher education firstly because higher education is expensive and secondly because they join the work force to supplement their household income. The insight is that household income and cost of education are significant determinants of college enrolment. Sometimes less intelligent can secure admission to professional programs on the basis of financial strength of the family.

Also association rule mining revealed that the demand for higher education has a direct co-relation with parents education. It is observed that parental income is the main determinant of participation in higher education. Past studies have established that first-generation college students (students who do not have a parent who attended college) often dropout and do not seek enrolment in higher education. As such, first-generation students may not receive the right guidance for higher education. Parent counselling is as important as student counselling. Parents should be actively involved in this act of nation building. Talented student with uneducated parents need better handholding and support. Higher accomplishing parents nurture those things that are important in their children and prioritize academic success.

6.0 References

1. Agrawal, R., & Srikant, R. (1994). Fast Algorithms for Mining Association Rules. Proceedings of the 20th VLDB Conference. Santiago, Chile.
2. AISHE Dataset. (2017). AISHE Higher Education data 2015-2016. New Delhi: Government of India, Ministry of HRD.
3. Association of Indian Universities AIU. (2017). Retrieved from Association of Indian Universities AIU: www.aiuweb.org
4. Bhargava N, Rajput A., Shrivastava P. (2010). Mining higher educational students data to analyze students' admission in various discipline, BJDMN1.
5. Chand, P. a. (2013). Factors Affecting higher educational choices of senior secondary science students in Himachal Pradesh, India. International journal of social science Tomorrow (IISST), Vol 2, 32-41.
6. Esnault, E. (1992). From Higher Education to Employment. Organisation for Economic Cooperation and Development, 35-40.
7. Fischer, M. J. (2007). Settling into campus life: Differences by race/ethnicity in college involvement and outcomes. The Journal of Higher Education, Vol. 78, No.2, 125-161.
8. Ghara, T. K. (2016). Status of Indian Women in Higher Education. Journal of Education and Practice, Vol.7, No.34, 58-65.
9. Han, J., & Kamber, M. (2001). Data mining: Concepts and techniques. Morgan Kaufmann Publishers.
10. Harrington, P. E. (1987). The enrollment crisis that never happened: how the job market overcame demographics. The Chronicle of Higher Education, vol 33, 44-45.
11. Mathew, C. (2014). Introducing Key Performance Indicators: Perspective of Higher Education Performance Monitoring and Evaluation. Journal of Education and Practice, Vol.5, No.29,130-137.
12. McKinsey Global Institute . (2016). The Age of Analytics: Competing In A Data-Driven World. McKinsey Global Institute Series, 35-60.
13. Mehari, K., & Ayalew, J. (2016). Multilevel Analysis for Identifying Factors Influencing Academic Achievement of Students in Higher Education Institution. Journal of Education and Practice, Vol.7, No.23, 88-96.
14. Odionye, E. A. (2014). The Role of Tertiary Education in Human Resource Development. Journal of Education and Practice, Vol.5, No.35,191-196.
15. Planning Commission Govt of India. (2012). Twelfth Five Year Plan 2012-17 Chapter on Higher Education (draft). Retrieved from http://planningcommission.gov.in /plans/planrel/ 12thplan/welcome.html
16. Raveendra, M. K., & Bhat, B. J. (2016). Higher Education in India: Challenges and Opportunities. The International Journal Of Humanities & Social Studies, Vol. 4 No.7,1-5.
17. Ray, D. C. (2007). Degrees of choices; social class Race and Gender in higher education (review). Journal of college student Development, Vol.48 No.6, 733-735.
18. Sen A., Rajput A. , (2018), Association Mining for Quality Rules in MP Higher Education Universal Review Vol. 7, No. 8, 288-301
19. Titus, M. A. (2006). Understanding college degree completion of students with low socioeconomic status: The influence of the institutional finance context. Research in Higher Education , Vol. 47,No.4, 371-398.
20. Verma, J. J. (2004). Education, Sustainable Development and Human Rights Approach. NAAC Decennial Year Lecture Series, 1-183.