

Identifying and Resolving Higher Educational Problems using Data Mining Technique

Tripti Arjariya, Vijay K Chaudhari, Rakesh Shrivastava, Dinesh Varshney

Abstract— *The objective of the present paper is advancement of knowledge and theoretical understanding of the relations among variables for the study and development using data mining in higher education system and its solution for Madhya Pradesh state. New knowledge takes three main forms: Exploratory research: which structures and identifies new problems, Constructive research: develops solutions to a problem, Empirical research: tests the feasibility of a solution using empirical evidence. The research methodology includes two distinct methods of research either primary or secondary. This study is a survey based followed by the developmental study of data mining in higher education Madhya Pradesh state. The terms indicate that, through theory generation, basic research provides the foundation for the better exploration of data mining research. Based on the information analyzed, it is stressed that the data mining approaches and issues, the survey on different data sources regarding educational status of school and colleges, the various policies of government of scholarships are helpful for the development and the solutions of higher education in Madhya Pradesh state in India.*

Index Terms— *Knowledge development, data mining approaches and issues, higher education system.*

I. INTRODUCTION

The hardware and software approaches make changes in technological applications. This makes society more scientific day to day due to generation-to-generation changes in different areas. In the teaching learning process intervention Information Communication Technology (ICT) able to solve the problem of lecturer, learners and administrators with a systematic way. In India and abroad, no country is able to solve their basic educational problems even growth of literacy rate from elementary to higher education system. The developed and developing countries worldwide trying to focus on development of educational system of under developed countries. The education systems are also reformed by the interference of International organization like UNESCO. In higher education system it is a great challenge to take advantages of ICT applications in general and find out

Manuscript Received October 29, 2011.

* Corresponding Author

Tripti Arjariya, Department of Computer Science and Engineering, Madhya Pradesh Bhoj (Open) University, Bhopal (M.P.)-462021 India. (E-mail: tripti.arjariya@gmail.com).

Vijay K Chaudhari, Head & Associate Professor, Information Technology Department, Truba Institute of Engineering and Information Technology, Bhopal (M.P.)-462021, India. (E-mail: Vijay_Ashish@yahoo.co.in).

Dr. Rakesh Shrivastava, Professor, Department of Higher Education, Govt. of Madhya Pradesh, Bhopal (M.P.)- 462021, India. (E-mail: rakesh_geol@yahoo.co.in)

Dr. Dinesh Varshney, Professor, Multimedia Regional Center, Madhya Pradesh Bhoj (Open) University, Indore (M.P.) - 452001, India. (E-mail: vdinesh33@rediffmail.com).

root causes of problems, prospects of ennoblement of learners. Still below 20% of learners are able to get admission in higher education in Indian context. The IT applications in the form of data are mined through different areas of higher education studies.

II. CATEGORIZATION OF DATA MINING SYSTEMS:

There are many data mining [DM] systems [3] [5] available which are specialized systems dedicated to a given data source or confined to limited DM functionalities, that are versatile and comprehensive. Some DM systems are according to the type of data handled such as spatial data, multimedia data, time-series data, text data, World Wide Web etc., some DM systems [18] are based on the data model involved such as relational database, object-oriented database, data warehouse, transactional etc., Some DM systems are based on the kind of knowledge discovered or the functionalities such as characterization, discrimination, association, classification, clustering etc. Some systems tend to be comprehensive systems offering several DM functionalities together. DM systems employ and provide different techniques, classification according to the data analysis approach used such as machine learning, neural networks, genetic algorithms, statistics, visualization, data base or data warehouse-oriented etc. The classification can also take into account the degree of user interaction involved in the DM process such as query-driven systems, interactive exploratory systems, or autonomous systems.

III. ISSUES IN DATA MINING:

DM algorithms embodied techniques existed for many years, and applied as reliable and scalable tools to perform classical statistical methods. Various issues have to be addressed such as Security and social issue [3] [5] [9] which is important as: any data collection that is shared or intended to be used for strategic decision-making. DM could disclose new implicit knowledge about individuals or groups that could be against privacy policies. Issue arises from this concern is the appropriate use of DM. Because of the competitive environment some important information or knowledge has to be withheld while other information could be widely distributed and used without control. In User interface issues the knowledge discovered [15 -16] by DM tools is useful as long as it is interesting, and above all understandable by the user.

Identifying and Resolving Higher Educational Problems using Data Mining Technique

The Mining methodology issues pertain to the DM approaches applied and their limitations. Versatility of the mining approaches [19], diversity of data available, dimensionality of the domain, broad analysis needs, assessment of the knowledge discovered, exploitation of background knowledge and metadata, control and handling of noise in data etc. are all examples that can dictate mining methodology choices. Data pre-processing [3] [5] [9] [12] becomes vital and the most important phase in the knowledge discovery process. DM techniques should be able to handle noise in data or incomplete information. In the Performance issues many artificial intelligence and statistical methods exist for data analysis and interpretation which raises the issues of scalability and efficiency of the DM methods. The Data source issues [14] [17] related to the data sources, where some are practical such as the diversity of data types, while others are philosophical like the data glut problem? This issue is the central idea for this research work. The advent of DM is certainly encouraging more data harvesting, the concern is whether we are collecting the right data at the appropriate amount and about its working and significance. Regarding the practical issues related to data sources, there is the subject of heterogeneous databases, and relational databases and data ware-houses are focussed in the recent past.

IV. ANALYSIS OF DATA WITH MAJOR HEADS

A. Phase-I

A. Primary Sources of Data (Questionnaire)

Reactions of Learners with IT enabled courses of different colleges in Madhya Pradesh State:

In order to know the success of programme “reactions of the teachers and students towards the IT application and DM implementation in higher education questionnaire” [1] were developed. Collected information through questionnaire were analysed by the use of percentage and chi-square analysis techniques. Details about the question wise analysis of data are given below. The reactions scale development and the implementation wise collected information in the form of non-parametric statistics [7] test is used. Both contingency and liner chi-square analysis techniques were used. Reactions of learners about IT acquainted person in Higher Education were sought and they were asked to express their response in five point ordinal scale which includes Strongly Agree, Agree, Undecided, Disagree, Strongly Disagree. Both positive and negative statements were included [2]. Frequency, percentage and Chi-Square analysis were adopted to find the findings.

	Male and Female	Married and unmarried	SC,ST,OBC,Ph.D Handicapped and General	Age	Rural, Semi-urban and Urban	B.E./M.Tech and Ph.D
Valid	198	198	198	193	181	196
Missing	0	0	0	5	17	2

Table 1: Overall Data Profile

Table 1 shows about the over all-primary sources of data status with their distributions. Out of 200-sample size of the

study unit 198 are undertaken and missing numbers are within age, location and qualifications. In this case they are ignore or unable to express the real information.

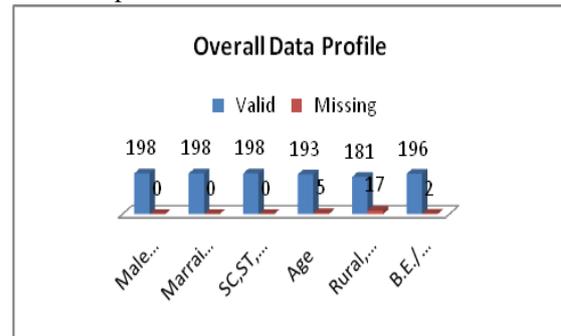


Fig. 1 overall data profile

Fig. 1 shows the overall data profile regarding various categories. The valid entries are under 198 out of 200 and others are missing with respect of their age, qualification and locations. This distribution shows the involvement of various categories in the profile.

Table 2: Sex-Wise Distribution Of Sample

	Frequency	Percent	Valid Percent	Cumulative Percent
Valid Male	140	70.7	70.7	70.7
Female	58	29.3	29.3	100.0
Total	198	100.0	100.0	

From Table 2 it can be noted that more than 70% of the IT acquainted DM sample are male where as less then 30% sample are female. As the national sample and the higher education status enrolment female respondents are less then male respondents but according to availability of the response through questionnaire they distributed as given in table.

Table 3: Marital Status Distribution Of Sample

	Frequency	Percent	Valid Percent	Cumulative Percent
Valid Unmarried	143	72.2	72.2	72.2
Married	55	27.8	27.8	100.0
Total	198	100.0	100.0	

In Table 3 more than 70% of the IT acquainted DM samples are unmarried where as less than 30% samples are married. According to availability of the response through questionnaire they are distributed as in table.

Table 4: Qualification Wise Distribution Of Sample

	Frequency	Percent	Valid Percent	Cumulative Percent
Valid BE	118	59.6	60.2	60.2
M. Tech	50	25.3	25.5	85.7
Ph.D./	28	14.1	14.3	100.0
Other				
Total	196	99.0	100.0	
Missing System	2	1.0		
Total	198	100		

Table 4 shows about the qualification wise distribution of sample. Near about 60% of sample are only BE qualification where as near about 40% are above then BE qualifications. In the questionnaire the parametric and non-parametric form of statistics required for analysis of asked questions both are in close ended and open ended are not able to analysis because of insufficient and qualitative form with unidirectional.

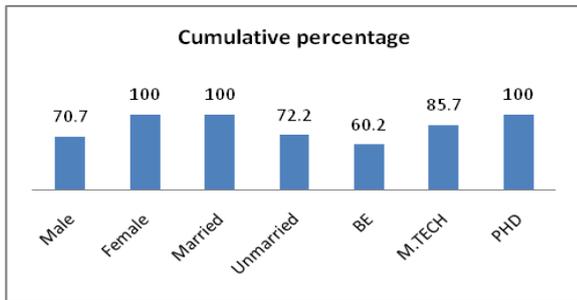


Fig. 2 Cumulative Percentage

Response	Strongly Agree	Agree	Un decided	Disagree	Strongly Disagree	Chi-Square
Frequency	17	31	4	3	4	46.38*
Percentage	28.81	52.54	6.78	5.06	6.78	

Based on the Questionnaire the distribution of various sample data shows the cumulative percentage of different kind of data which is classified as sex-wise data, marital status wise data, and Qualification wise data. The percentage value will help us to analyze the actual output data for further operations regarding higher educational problems.

Non-Parametric Form Of Data Analysis

Table 5: It Provides Better Understanding Of Students needs.

Response	Strongly Agree	Agree	Un decided	Disagree	Strongly Disagree	Chi-Square
Frequency	14	65	20	92	15	118.32*
Percentage	6.80	31.5	9.71	44.66	7.28	

Table-5 shows about the reactions of students against the statement/teachers, it provides better understanding of students and teachers need, chi-square value 118.32 with degree of freedom 1/208 is significant at 0.01 levels. Maximum students disagree with the statement. The Findings shows that there is IT applications which provide better understanding of students and teachers needs.

Table 6: Dm Provides Satisfaction To Teach The Challenged And Empower Them With Life Skills

Response	Strongly Agree	Agree	Un decided	Disagree	Strongly Disagree	Chi-Square
Frequency	31	98	53	14	11	120.44*
Percentage	14.98	47.34	25.60	6.76	5.31	

Table 7: The Technical Courses Develop Skills Required For An Effective Technical Education Assistant/Associate

Professor And Motivate To Perform As Expected In Present Day Job Situation.

Response	Strongly Agree	Agree	Un decided	Disagree	Strongly Disagree	Chi-Square
Frequency	11	17	20	55	103	140.94*
Percentage	5.34	8.25	9.71	26.69	50	

Table 7 shows that maximum learners strongly disagree with the statement that Technical courses develops skills required for an effective Technical education Assistant/Associate Professor and motivate to perform as expected in present day job situation. The chi-square value of the response is 140.94, which is significant at 0.01 levels. Thus for the learners of IT the Technical courses is not sufficient. The Findings shows that the Technical courses develops skills required for an effective Technical education Assistant/Associate Professor and motivate to perform as expected in present day job situation, all this needs more exercises and development.

Table 8: Technical Education Programme Develops Skills Through University Based Practical Activities.

Response	Strongly Agree	Agree	Un decided	Disagree	Strongly Disagree	Chi-Square
Frequency	17	31	4	3	4	46.38*
Percentage	28.81	52.54	6.78	5.06	6.78	

Table 8 shows reaction of teachers against the statement “The Technical Education Programme develops skills through University Based Practical Activities.” The chi-square value for the frequency distribution is 46.38, which is significant at 0.01 levels. It means the teachers are in favor of the statement. Thus the teachers accepted the statement. The Findings shows that Technical Education Programme develops skills through University Based Practical Activities.

Table 9: The Workshop-Based Practical Gives Exposure To The Equipments Related To The Children With Technical Needs.

Response	Strongly Agree	Agree	Un decided	Disagree	Strongly Disagree	Chi-Square
Frequency	17	31	4	3	4	46.38*
Percentage	28.81	52.54	6.78	5.06	6.78	

The frequency of responses of teachers shown in table 9 represents that maximum teachers strongly agree and agree with the statement “The workshop-based practical give exposure to the equipments related to the children with Technical needs”. The Chi-value for the statement is 14.04, is significant at 0.01 level. Thus teachers had accepted the statement. The Findings shows that the workshop-based practical give exposure to the equipments related to the children with Technical needs.

Table 10: Traditional And Distance Education Is Not Fulfilling Need Of Technical Education Requirement In India.

Response →	Strongly Agree	Agree	Un decided	Disagree	Strongly Disagree	Chi-Square
Frequency	11	33	7	4	4	46.92*
Percentage	18.64	55.93	11.86	6.78	6.78	

Table 10 shows that maximum teachers Traditional and Distance Education [13] are not fulfilling need of Technical education requirement in India. The Chi-square value of the statement is 46.92 which is significant at 0.01 level. The Findings shows that traditional and Distance Education is not fulfilling need of Technical education requirement in India.

Table 11: Open Distance Learning System Is More Valuable Through Data Mining Techniques.

Response →	Strongly Agree	Agree	Un decided	Disagree	Strongly Disagree	Chi-Square
Frequency	6	9	5	21	16	14.21*
Percentage	10.53	15.79	8.77	36.34	28.07	

Table 11 shows that maximum learners disagree with the statement “Open Distance Learning System is more valuable through DM techniques”, The chi-square value for the statement 14.21 is significant at 0.01 level. Thus the teachers rejected the statement. The Findings shows that the Open Distance Learning System is more valuable through DM techniques.

Table 12: Rural Technical Education Is Feasible Through Ict Data Mining Techniques.

Response: →	Strongly Agree	Agree	Un decided	Disagree	Strongly Disagree	Chi-Square
Frequency	17	31	4	3	4	46.38*
Percentage	28.81	52.54	6.78	5.06	6.78	

Table 12 shows the reaction of teachers against the statement “Rural technical education is feasible through ICT DM techniques”, The chi-square value for the frequency distribution is 46.38, which is significant at 0.01 level. It means the teachers are in favor of the statement. Thus the teachers accepted the statement. The findings show that rural technical education is feasible through ICT DM techniques.

Table 13: Dm Techniques Required For Administration And Facilitations Of Higher Education

Response: →	Strongly Agree	Agree	Undecide d	Disagree	Strongly Disagree	Chi-Square
Frequency	18	13	6	10	8	6.48**
Percentage	32.73	23.64	10.91	18.18	14.55	

Table 13 shows the reactions of teachers against the statement “DM techniques required for administration and facilitations of higher education”, is found to be significant at 0.05 level with chi-square value 6.48 with degree of freedom 1/58. It means teachers accept the statement. The Findings shows that DM techniques required for administration and facilitations of higher education.

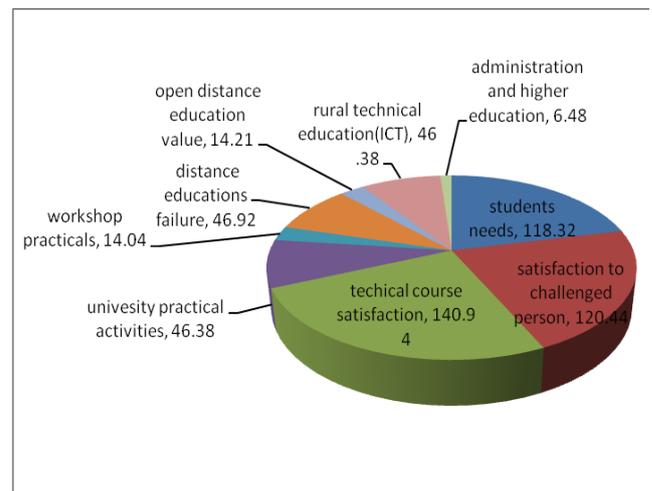


Fig. 3 Chi-square values of various statements

Fig 3 shows the analysis on the basis of chi-square values of various statements regarding their acceptance and rejections. The significance of chi-square value is at level 0.01 which shows particular statement is accepted otherwise it is rejected. The frequency and the percentage helps to find out the chi-square value on four fields i.e. Strongly Agree, Agree, Undecided, Disagree, and Strongly Disagree. On the basis of this data it is helpful to find out the appropriate results regarding the higher educational problems and at a great extent we can able to search out its solutions. Various parameters like the students needs, satisfaction to challenged person, technical course satisfaction, university practical activities, workshop practical, distance education failure and success, open distance education, rural technical education and the Administration and higher education are the major fields which are very necessary and helpful for the research in higher educational problem identification. The analytics and the calculations on these fields are able to get the final result.

Phase –II

B. Secondary Sources of Data (Government Initiations):

STATE POLICES [4] [10] [20] [21]:

Construction in govt. colleges: In the year 2009-10, construction of buildings in Govt. colleges [8], estimation is Rs.1000 Lakhs which is required under requirements no. 44 (general area) and in first rule, Rs. 2000 Lakhs is estimated, according to which, under construction of new buildings of college and extra construction of 85 colleges are in progress., As same in year 2009-2010, construction buildings of govt. colleges in requirement no. 41 (schedule tribe sub policies) Rs 485.32 Lakhs is required according to which 22 colleges are under construction.

To provide study material and stationary free of cost to students those belongs to SC/ST: The SC/ST students studying in Govt. colleges [10] are given books free of cost at the rate of Rs 600 only per student at UG level and Rs 800 only per student at PG level, and at both level rs 50 only per student free stationary is provided. In final year 2009-10 Rs 250 Lakhs for SC students and Rs 144 Lakhs for ST students are estimated. By this financial policy near about 55000 SC and 35000 ST students are profited.

Policy for daughter in villages: In this policy the girls who is living in village area and get maximum marks in merits list of class 12th get rupees 5000 yearly scholarship for her study is estimated [11]. In year 2008-09 estimated amount is Rs. 7500 per year to the girls student who study in engineering and medical colleges in year 2009-10 Rs. 1400.00 lakh is estimated, in this policy till now, 23958 girls are profited.

PRATIBHA kiran policy: To encourage the girls living below poverty line in urban areas for higher education in financial year 2008-09 PRATIBHA kiran policy has been started. Accordingly, girls living below poverty line in urban areas and who got first division in class 12th are given scholarship of Rs. 5000 yearly for regular studies in colleges. In this policy Rs 50.00 Lakhs is estimated for year 2009-2010. For the activation of this policy a budget is allotted to government colleges from which 2062 girl’s students are profitable.

VIKARAMDITYA policy: This Policy has been started to encourage students living below poverty line of general category, according to this policy those families whose annual income is below Rs. 42,000 will be profited. For the activation of this policy in year 2009-10 Rs. 40.00 lakh is allotted. Up to month Jan 2010 all the money for the policy are given according to the application came for college level to the government colleges.

Help to PhD scholar: The student belongs to ST/SC who are pursuing PhD will be given 8000 pm. In financial year 2009-10 rupees 25.00 lakh is allotted to ST students, and Rs.100.00 Lakhs is allotted to SC students.

Scholarship policy: A policy is running by higher education department for various regular students to give them scholarship according to their eligibility thus helping them financially.

Table 14: In Present Situation Various Scholarships And Their Quota, Which Are Activated

S. No.	Type of scholarship	Quota	Money allotted in Lakhs for session 2009-10
1	Single scholarship		17.60
	a. Research (PhD)	18	
	b. M. Phil	14	
	c. PG eligibility	73	
	d. CO- education PG eligibility	73	
	e. Sports and Games	22	
	f. UG eligibility	146	
	g. Co- education UG eligibility	146	
	h. Film & DD Doordarshan organization	02	
	i. National Academy of ARTS, new Delhi	02	
	j. National school of drama, new Delhi	02	
2	Sanskrit Scholarship	200	2.42
3	Scholarship given to the children of landless farmers of MP for professional education	100	1.00

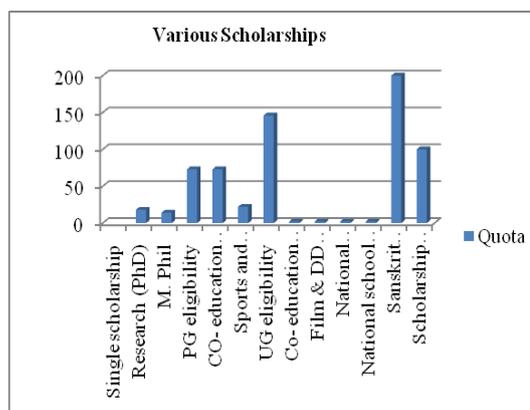


Fig. 3 various scholarships and their quota

Fig. 3 show about the scholarship schemes and the policies that are planned and executed by the government of India for higher education for the growth in educational status in Madhya Pradesh state. The policies are made on the considerations of Indian culture, its religious values, financial aspects and developmental criteria’s. These data is very much helpful to analyze and conclude the outcomes of educational problem identifications and there solution at a great extent.

PHASE –III

IMPLEMENTATION:

By knowledge discovery in databases, interesting knowledge, regularities, or high-level information can be extracted from the relevant sets of data in databases and be investigated from different angles, and large databases thereby serve as rich and reliable sources for knowledge generation and verification. Mining information and knowledge from large databases has been recognized by many researchers as a key research topic in database systems and machine learning and by many industrial companies as an important area with an opportunity of major revenues.

PROPOSED ALGORITHM

Algorithm 1

Input: Data set R, Attribute set A_i

Output: data set R'

$R' \rightarrow R$

For $I=1$ to n do

Max (A_i) = the deepest node in the attribute set A_i

If Max(A_i). Distance_to_max< I_i

Newnode=node.root_path_array[I_i -node.distance_t-_max]

Else

Newnode=max(A_i)

Endif

Replace node with new node

Endfor

Remove duplication from R'

End

Algorithm 2

Input : Primitive rules set R

Output Generalized rules set R'

$R' \leftarrow 0$

$N = |R|$

For $I=0$ to $N-1$ do

$r \leftarrow r_i$

$M \leftarrow |r|$

For $j=0$ to $M-1$ do

If r_i inconsistent with rule $r_n \in E$ then

Restore the dropped condition a_j

Endif

End for

Included in rule r

If rule r is not logically include in a rule $r' \in MRULE$ then

$MRULE \leftarrow r \cup MRULE$

Endif

End

R: Data set which is a any college web site because Ontology is used for specific domain.

R' : is the output

A_i is the attribute set

Max (A_i) is the function which finds the deepest node in the attribute set A_i

are very low but its applicability in different sectors of the society is highly in demand day to day. With specific to education sector it has great demand both teaching and learning prospects. The management aspects are highly interference by the Information Communication Technology and DM areas. Higher Education system in India, now a day's totally depended on DM majors. The demand and problem solving abilities within the framework of logical argument and accuracy of result need to explore through research and development procedure. Efforts are made by Government, NGOs and Independent bodies trying to make social problems solvable easily through the DM. the entire research work and the experimental results shows that the data mining techniques are very much helpful for the higher education development and its solutions. With the help of various findings on various illustrated parameters we are able to get the result on the higher educational problems and able to know that how to overcome it.

REFERENCES

1. Cave, M., Kogan, M. and Hanney, S. (1990), The scope and effects of performance measurement in British higher education, in F. J. R. C. Dochy, M. S. R. Segers and W. H. F. W. Wijnen (Eds.), Management Information and Performance Indicators in Higher Education, Van Gorcum and Comp, 48–49.
2. Fielden, J., and Abercromby, K. (2000), UNESCO Higher Education Indicators Study: Accountability and International Co-operation in the Renewal of Higher Education, Georgia Professional Standards. UNESCO, Paris.
3. Han, J. and Kamber, M. (2001), Data Mining: Concepts and Techniques, Simon Fraser University, Organ Kaufmann.
4. Johnstone, J.N. (1976), Indicators of the Performance of Educational Systems. UNESCO: International Institute for Educational Planning, Paris.
5. Luan, J. (2001), Data mining and knowledge management, a system analysis for establishing a tiered knowledge management model (TKMM), In Proceedings of Air Forum, Toronto, Canada.
6. Luan, J. (2002a), Data mining and knowledge management in higher education – potential applications, In Proceedings of AIR Forum, Toronto, Canada.
7. Luan, J. (2002b), Data Mining Application in Higher Education, SPSS Executive Report.
8. Luan, J., Zhao, C.M. and Hayek, J. (2004). Use data mining techniques to develop institutional typologies for NSSE, National Survey of Student Engagement.
9. Modern Data Warehousing, Mining and Visualization Core Concepts, George M. Marakas.
10. Wako, T. N. (2003), Basic Indicators of Educational System's Performance, National Educational Statistics Information Systems, UNESCO, Harare, Zimbabwe.
11. Yang, M., Goldstein, H., Rath, T. and Hill, N. (1999), The use of assessment data for school improvement purposes, Oxford Review of Education, **25**, 469–483.
12. Zhang Yofeng, Wu Jinhong, Wang Cuibo, Automatic Competitive Intelligence Collection Based On Semantic web mining IEEE 2007
13. Wang Jain, Li Zhuo, Research and Realization of Long Distance Education platform based on Web Mining, IEEE 2009
14. Chang-xin Song, Ke Ma ,Applications of Data Mining in the Education resource based on XML, International Conference on Advanced Computer Theory and Engineering, 2008
15. Intelligent Web Mining Model to Enhance Knowledge Discovery on the Web, Sunil Kr.Pandey1 , R.B.Mishra2, Proceedings of the Seventh International Conference on Parallel and Distributed Computing, Applications and Technologies (PDCAT'06) 0-7695-2736-1/06 © 2006 IEEE

V. CONCLUSION

DM is directly associated with use of technology for accessing data and to give result as required in a desired way. In Indian contest though computer literacy among the users

16. Learning Object Models from Semi structured Web Documents Shiren Ye and Tat-Seng Chua; IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 18, NO. 3, MARCH 2006
17. A Flexible Structured-based Representation for XML Document Mining Anne-Marie Vercoustre, Mounir Fegas, Saba Gul, and Yves Lechevallier; rXiv:cs/0607012v1 [cs.IR] 5 Jul 2006
18. Mao-sang Lin, Hui Zhang, Zhang Gua Yo, An ontology for supporting data mining process, IMACS CESA, 2006
19. Survey Paper on Top 10 algorithms in data mining; XindongWu · Vipin Kumar · J. Ross Quinlan · Joydeep Ghosh · Qiang Yang · Hiroshi Motoda · Geoffrey J. McLachlan · Angus Ng Bing Liu · Philip S. Yu · Zhi-Hua Zhou · Michael Steinbach · David J. Hand · Dan Steinberg Published online: 4 December 2007 © Springer-Verlag London Limited 2007
20. Amy M. Hightower, Hajime Mitani, Christopher B. Swanson, State Policies That Pay “ Survey of School Finance Policies and Outcome” , Copyright © 2010 by Editorial Projects in Education
21. Geeta Gandhi Kingdon, “The progress of school education in India “, GPRG-WPS-071, March 2007