INVESTIGATION OF EDUCATION AND EMPLOYMENT PATTERN CHANGES WITH BIG DATA ANALYTICS IN ACCORDANCE WITH A RURAL AREA OF COIMBATORE REGION USING APRIORI ALGORITHM

V. Kirubha[1], M. Hemalatha[2], T. Sumathi[3]

ABSTRACT

Education and employment are made an important and integral part of the national development efforts. The development process must be visualized in its broadest context if it is to meet the expectations of the citizenry for a more elevated standard of living. The important goals to be achieved are the national development, rapid growth of incomes, poverty reduction, education and employment facilities. To improve the education and employment facilities in a rural, it is important to analyze its present status in education and employment. It can be obtained by collecting and analyzing the education and employment data of a rural. This kind of data can be of large volume and that must be processed and transformed into useful results and hence data mining techniques are greatly used in this analysis task. Analyzing this kind of big data helps to find the education and employment pattern changes in a rural region. This can support government authorities to find fields that need to be improved in that rural region. This paper clearly presents the analysis made on education and employment pattern using apriori algorithm.

Keywords: Education, Employment, Rural, Association, Apriori, Big data.

[1]Department of Computer Science, Karpagam University, Coimbatore.
[2]Department of Computer Science, Karpagam University, Coimbatore.
[3]Department of Computer Science, Karpagam University, Coimbatore.

I. INTRODUCTION

The development of rural requires improvement in all fields like social and economic. In this education and employment are the major concerns. Employment and education analysis has very significant role in any country. Education and employment analysis is a task that finds the frequent patterns of the education and employment status of people living in a region. This high volume of datasets and complexity of relationships between these kinds of data have made education and employment analysis an appropriate field for applying data mining techniques. Government authorities everywhere have been handling a large amount of information and huge volume of records. This kind of big data has high volume and it is difficult to obtain useful results from this data. Analyzing education and employment data by applying data mining techniques provides useful information that helps to know about a rural education and employment. Data about a rural can be of very large. To analyze education and employment data an appropriate data mining approach has to be chosen. Association rule mining is an approach of data mining which finds frequent patterns. Mining data related to education provides useful information such as whether most people in a rural are studied or not studied. And mining data related to employment can give us the information like people in a

rural are employed or not. It also specifies some interesting pattern changes such as most preferred studies, most preferred occupation by people living in that rural. The results obtained from this analysis can help the government authorities to improve facilities for education and employment of a rural.

In this paper, Apriori algorithm has been used to find the frequent patterns in the education and employment dataset. Apriori algorithm on education and employment dataset of a rural is implemented using Matlab. Also for this analysis, the dataset is collected in real time. In this paper, education and employment pattern changes of a rural is being analyzed.

II. Related Works

Akash Rajak and Mahendra Kumar Gupta [1] presents various areas in which association rules are applied. Census makes a huge variety of general statistical information on society available to both researchers and the general public. The information related to population and economic census can be forecasted in planning public services(education, health, transport, funds) as well as in public business(for setup new factories, shopping malls or banks and even marketing particular products). The application of data mining techniques to census data and more generally to official data has great potential in supporting good public policy and in underpinning the effective functioning of a democratic society. Nisha Srivastava and Ravi Srivastava [2] analyzed the trends and patterns of women's employment in rural india. Large scale national surveys data are analyzed to examine the employment status of women and presented some suggestions for improving the position of women worker's in rural areas. Smitha and Suresh kumar [3] presents the applications of big data in data mining. Big data concept can be used to implement data mining concepts. Big data presents more opportunities for research and reference in the public sector as well in technical progress. The challenges in data analyzing can be overcomed by capturing the techniques in big data. Irina Tudor [4] described the association rule mining technique for discovering interesting association relationship among huge amounts of business transaction records and concepts of apriori algorithm. Wei Fan and Albert Bifet [5] presents some concepts on big data, available open source tools for handling Big data and also discusses the future challenges in working with big data. And they concluded that Big data is the new frontier for scientific data research and for business applications. Rohit Pitre and Vijay Kolekar [6] discuss on data mining with big data and features of big data are explained. Also they proposed the challenges for big data as mining platform, privacy and design of mining algorithms. Smitha T and V. Sundaram [7] presents the association rule mining using apriori concept with example. The itemset generations of frequent patterns are applied on example dataset and explained. And examined certain associations are exist between parameters of a database that can be used for prediction.

III. Proposed System

Development of a rural needs some prevailing information about that rural. The collected raw data about education and employment does not specify any information about the status of a rural. It is required that some pattern has to be extracted from the education and employment data. Analyzed results help us to determine the current situation and necessary improvement to be made to upgrade the rural region.

The following procedure is used in the proposed system:

- Collect the education and employment data from a rural and design a dataset.
- Extract required fields for processing.
- Apply Apriori algorithm concept to find the frequent patterns.
- Visualize the results.

The Education and Employment data collected from a rural area in Coimbatore is framed as a dataset and used in this analysis. The collected real time data has numerous attributes and hence big data is appropriate for analyzing this field. Apriori algorithm is used to find association among various attributes in this data. This paper focuses on finding frequent pattern changes on Education and Employment dataset using Apriori algorithm.

3.1 Association Rule Mining

[16]Association rule mining is a data mining technique which helps to find interesting association or correlation relationships among a large set of data items. With massive amounts of data continuously being collected and stored in databases, many companies are becoming interested in mining association rules from their databases to increase their profits. For example, the discovery of interesting association relationships among huge amounts of business transaction records can help them in catalog design, cross marketing, loss leader analysis, and other business decision making processes. A typical example of association rule mining is market basket analysis. This process analyzes customer buying habits by finding associations between the different items that customers place in their "shopping baskets". One of the most popular data mining techniques is association rule mining. The patterns discovered with this data mining technique can be represented in the form of association rules. Rule support and confidence are two measures which can be used to evaluate the data. Typically, association rules are considered interesting if they satisfy both a minimum support threshold and a minimum confidence threshold. Such thresholds can be set by users or domain experts.

Concepts

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.

Support

The support supp(X) of an itemset X is defined as the proportion of transactions in the data set which contain the itemset.

supp(X)= no. of transactions which contain the itemset X / total no. of transactions

Confidence

The confidence of a rule is defined:

conf(X'!Y)=supp(X U Y)/supp(X)

3.2 Apriori Algorithm

[9]Apriori is an influential algorithm for mining frequent item sets. The name of the algorithms is based on the fact that the algorithm uses prior knowledge of frequent item sets properties. Apriori employs an iterative approach known as a level-wise search.

To improve the efficiency of the level-wise generation of frequent item sets, an important property called the apriori property, **i.e.**" all nonempty subsets of a frequent item sets must also be frequent."

Apriori algorithms having a two-step process.

The join step: To find L_k , a set of candidate k item sets is generated by joining L_{k-1} with itself. This set of candidate is denoted C_k. The prune step: C_k is the superset of L_k, that is, its members may or may not be frequent, but all of the frequent k-itemsets are included in C_k. A scan of the databases to determine the count of each candidate in C_k would result in the determination of L_k. (i.e. all candidates having a count no less than the minimum support count are frequent by definition, and therefore belongs to L_k).

Pseudo-code:

Join Step: Ck is generated by joining L_k -1with itself

Prune Step: Any (k-1)-itemset that is not frequent cannot be a subset of a frequent k-itemset

C_k: Candidate itemset of size k

L_k: frequent itemset of size k

L1= {frequent items};

for(*k*= 1; L_k!="; *k*++) **do begin**

C_{k+1} = candidates generated from L_k;

for each transaction *t*in database do

increment the count of all candidates in L_k +1 that are contained in *t*

Lk+1= candidates in *Ck+1*with min_support

End

return*"*kLk*;

It makes multiple passes over the database. In the first pass, the algorithm simply counts item occurrences to determine the frequent 1-itemsets (itemsets with 1 item). A subsequent pass, say pass k, consists of two phases. First, the frequent itemsets L_{k-1} (the set of all frequent (k-1)-itemsets) found in the (k-1)th pass are used to generate the candidate itemsets C_k, using the apriori-gen() function. This function first joins L_{k-1} with L_{k-1}, the joining condition being that the lexicographically ordered

first k-2 items are the same. Next, it deletes all those itemsets from the join result that have some (k-1)-subset that is not in L_{k-1} yielding C_k.

The algorithm now scans the database. For each transaction, it determines which of the candidates in C_k are contained in the transaction using a hash-tree data structure and increments the count of those candidates. At the end of the pass, C_k is examined to determine which of the candidates are frequent, yielding L_k. The algorithm terminates when L_k becomes empty.

IV. Dateset Description

Education and employment dataset has around numerous attributes and records. This dataset is collected in Madukkarai a rural area in Coimbatore. This data is collected from the family head. It has 250 records and 53 attributes. Only relevant attributes needed for processing are selected and processed.

Attribute name	Description	Data type
st_name	Street name	nominal
fmlyhead	Head of family	nominal
Mem	Number of members	number
Hedu	Educational qualification of the head	polynominal
hedu_rsn	Reason for not continued education by head	nominal
Hsts	Occupation of the head	nominal
hfather_edu	Educational qualification of the head's father	polynominal
hmother_edu	Educational qualification of the head's mother	polynominal
prntedu_rsn	Reason for not continued education by head's parents	nominal
Prntocpn	Occupation of the head's parents	nominal
Ptredu	Educational qualification of the head's partner	polynominal
ptredu_rsn	Reason for not continued education by head's partner	nominal
Ptrocpn	Occupation of the head's partner	nominal
Nwards	Number of wards associated with the head	Number
w1edu	Educational qualification of ward1	polynominal
w1sts	Occupation of ward1	nominal
w2edu	Educational qualification of ward2	polynominal

w2sts	Occupation of ward2	nominal
w3edu	Educational qualification of ward3	polynominal
w3sts	Occupation of ward3	nominal
w4edu	Educational qualification of ward4	polynominal
w4sts	Occupation of ward4	nominal
w1ptredu	Ward1's partner's educational qualification	polynominal
w1ptredu_ rsn	Reason for not continued education by ward1's partner	nominal
w2ptredu	Ward2's partner's educational qualification	polynominal
w2ptredu_ rsn	Reason for not continued education by ward2's partner	nominal
w3ptredu	Ward3's partner's educational qualification	polynominal
w3ptredu_ rsn	Reason for not continued education by ward3's partner	nominal
w4ptredu	Ward4's partner's educational qualification	polynominal
w4ptredu_rsn	Reason for not continued education by ward4's partner	nominal
Talent	Whether known about talent exam	nominal
preparing	Whether preparing for	nominal
Readhbt	Whether having reading habit or not	nominal
w1ocpn	ward1's occupation	nominal
w1org	ward1's working organization	nominal
w2ocpn	Ward2's occupation	nominal
w2org	Ward2's working organization	nominal
w3ocpn	Ward3's occupation	nominal
w3org	Ward3's working organization	nominal
w4ocpn	Ward4's occupation	nominal
w4org	Ward4's working organization	nominal
w1ptrocpn	Ward1's partner's occupation	nominal

w1org	Ward1's partner's working organization	nominal
w2ptrocpn	Ward2's partner's occupation	nominal
w2org	Ward2's partner's working organization	nominal
w3ptrocpn	Ward3's partner's occupation	nominal
w3org	Ward3's partner's working organization	nominal
w4ptrocpn	Ward4's partner's occupation	nominal
w4org	Ward4's partner's	nominal

V. Results And Discussion

Head's and Partner's Educational Qualification

Figure 5.1 shows the frequent pattern of educational pattern of head and partner. From this result the maximum educational qualification of head and partner is below sslc.

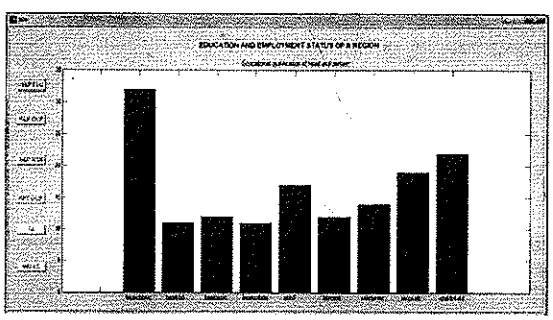

Fig.5.1 Educational pattern of head and partner.

Occupation of head and partner

Figure 5.2 shows the employment status of head and partner. Maximum heads in this region are private employees and their partners are house wives.

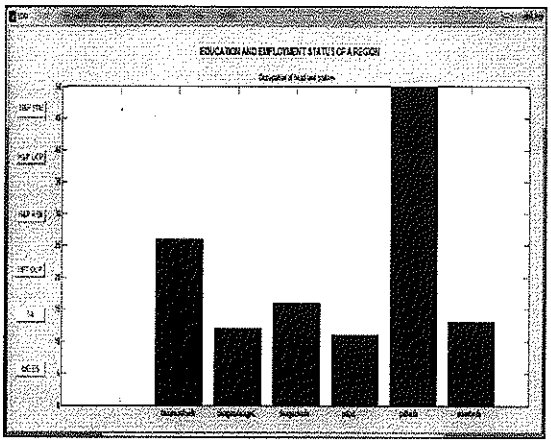

Fig.5.2 Occupation pattern of head and partner.

Reason for discontinued education

Figure 5.3 presents the reason for discontinued education of head and partner is not interested and not allowed respectively.

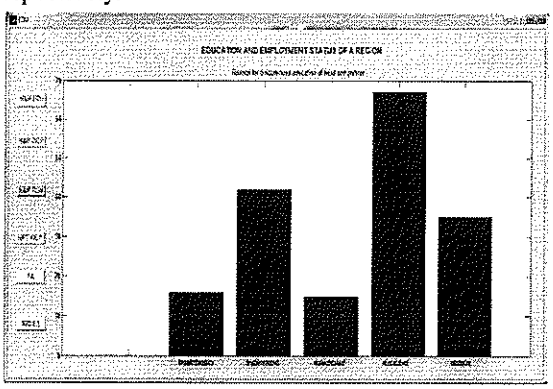

Fig. 5.3 Reason for not continued education

Occupation of head and parent

Figure 5.3 shows the frequent pattern of head's and their parent's occupation. In this two results are obtained i.e.

one is daily wages is the occupation of both head and parent and the another is private employees and daily wages are the occupation of both head and parent respectively.

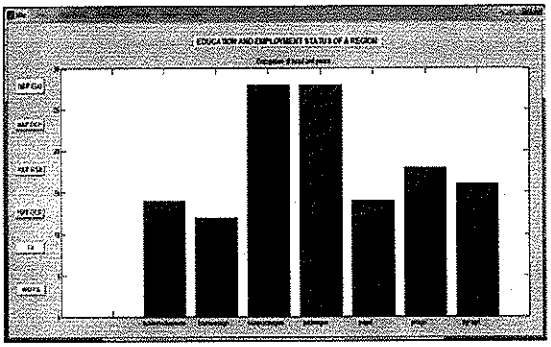

Fig. 5.4 Employment pattern of head and parent.

Awareness about Talent exam

Figure 5.5 shows that awareness about Talent exam among government school students in the region.

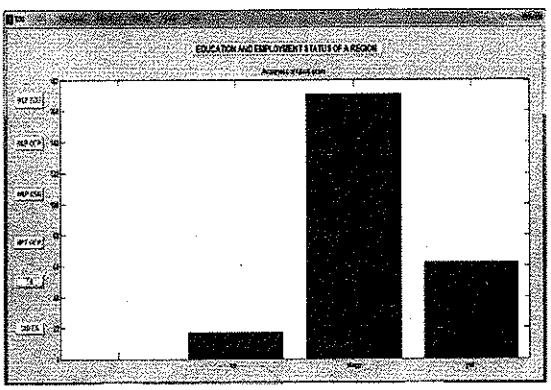

Fig. 5.5 Talent exam Awareness

Educational Qualification of wards

Figure 5.6 shows the result of educational and employment pattern of head's wards in the region. From the analyzed data, it shows that maximum people in this region are preferring private schools for their wards education.

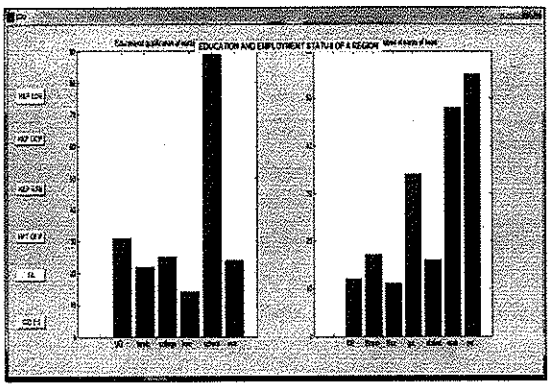

Fig. 5.6 Educational Qualification of wards

Discussion about the Results

From the dataset collected only east zone of Madukkarai is taken for analysis. Considering 250 records for analysis, only 12.8% of head and partner in this zone are studied below sslc. 20% of people are private employees. Out of 250 records 22% of people preferred private schools and only 14% of people preferred government schools. The result obtained proves that people in this zone have mostly preferred private schools for their children, even though they are daily wages and private employees. If the entire rural area is taken for analysis this results may be considered some improvement.

If the awareness about government school facilities is provide among people of this rural their mentality about government school may change.

CONCLUSION

This project focuses on finding frequent patterns on education and employment dataset of a rural. From the results the existing education and employment status of

a rural has been obtained. It may help the government authorities to improve the educational and employment opportunities in this rural area.

FUTURE ENHANCEMENT

The results have proven that data mining is an appropriate field for mining high dimensional data. This project finds the frequent patterns on a rural education and employment. It can also be extended to all rural of a Coimbatore district to know about the district's educational and employment status for development. This project used Apriori Algorithm to find the association rules. It can be extended as a model for analyzing education and employment data of all rural area as well as urban areas.

REFERENCES

[1] Akash Rajak and Mahendra Kumar Gupta, *"Association Rule Mining: Applications in Various Areas"*, International Conference on Data Management.

[2] Nisha Srivastava and Ravi Srivastava, *"Women, work, and employment outcomes in rural India"*, FAO-IFAD-ILO Workshop on Gaps, trends and current research in gender dimensions of agricultural and rural employment: differentiated pathways out of poverty, Rome, 31 March - 2 April 2009.

[3] Smitha T, V. Suresh Kumar, *"Application of Big Data in Data Mining"*, International Journal of Emerging Technology and Advanced Engineering, ISSN 2250-2459, Volume 3, Issue 7, July 2013.

[4] Irina Tudor, *"Association Rule Mining as a Data Mining Technique"*, BULETINUL, Vol. LX No. 1/ 2008, 49-56.

[5] Wei Fan and Albert Bifet, *"Mining Big Data: Current Status, and Forecast to the Future"*, SIGKDD Explorations Volume 14, Issue 2.

[6] Rohit Pitre and Vijay Kolekar *"A Survey Paper on Data Mining With Big Data"*, International Journal of Innovative Research in Advanced Engineering (IJIRAE), Volume 1 Issue 1, April 2014.

[7] Smitha. T, V. Sundaram, *"Association Models for Prediction with Apriori Concept"*, International Journal of Advances in Engineering & Technology, Vol. 5, Issue 1, pp. 354-360, ISSN: 2231-1963, Nov 2012.

[8] Jiawei Han and Micheline Kamber, *"Data Mining: Concepts and Techniques"*, Morgan Kaufmann Publishers, second Edition, 2006.

[9] Divya Bansal Lekha Bhambhu *"Execution of APRIORI Algorithm of Data Mining Directed towards Tumultuous Crimes Concerning Women"*, International Journal of Advanced Research in Computer Science and Software Engineering Volume 3, Issue 9, September 2013.

[10] Jiao Yabing, *" Research of an Improved Apriori Algorithm in Data Mining Association Rules"*, InternationalJournal of Computer and Communication Engineering, Vol. 2, No. 1, January 2013.

[11] T. Karthikeyan and N. Ravikumar, *"A Survey on Association Rule Mining"*, International Journal of Advanced Research in Computer and Communication Engineering, Vol. 3, Issue 1, January 2014.

[12] Rachna Somkunwar, *"A study on Various Data Mining Approaches of Association Rules"*, International Journal of Advanced Research in Computer Science and Software Engineering, Volume 2, Issue 9, September 2012.

[13] Ms Shweta Dr. Kanwal Garg," *Mining Efficient Association Rules Through Apriori Algorithm Using Attributes and Comparative Analysis of Various Association Rule Algorithms"*, International Journal of Advanced Research in Computer Science and Software Engineering, Volume 3, Issue 6, June 2013.

[14] Sotiris Kotsiantis, Dimitris Kanellopoulos, *"Association Rules Mining: A Recent Overview"*, GESTS International Transactions on Computer Science and Engineering, Vol.32 (1), 2006, pp. 71-82.

[15] Arvind Jaiswal, Gaurav Dubey, *"Identifying Best Association Rules and Their Optimization Using Genetic Algorithm"*, International Journal of Emerging Science and Engineering (IJESE) ISSN: 2319–6378, Volume-1, Issue-7, May 2013.

[16] Association rule mining- Wikipeia the free encyclopedia.

AUTHOR'S BIOGRAPHY

Dr. Sumadhi.T is an Associate Professor, Department of Computer Science in Karpagam Univesity, Coimbatore. She received her B.Sc.(CS), in 1996 and MCA in 1999 from Bharathiar University, Coimbatore. She obtained her M.Phil. in the area of Image Processing from Periyar University, Salem in 2006. Her research interest lies in the area of Image Processing and Data mining. . She obtained her Ph.D. in the area of Image Processing from Karpagam University, Coimbatore in 2013. She has published about 20 papers in international and national journals. She has attended about 15 international and national conference held at various places in Tamil Nadu.

Dr. M.Hemalatha completed M.Sc., M.C.A., M. Phil., Ph.D (Ph.D, Mother Terasa women's University, Kodaikanal). She is Professor & Head and guiding Ph.D Scholars in Department of Computer Science in Karpagam University, Coimbatore. Twelve years of experience in teaching and published more than hundred papers in International Journals and also presented more than eighty papers in various national and international conferences. Area of research is Data Mining, Software Engineering, Bioinformatics, and Neural Network. She is a Reviewer in several teaching and published more than hundred papers in International Journals and also presented more than eighty papers in various national and international conferences. Area of research is Data Mining, Software Engineering, Bioinformatics, and Neural Network. She is a Reviewer in several National and International Journals.