RESEARCH PAPER | Social Science | Volume : 4 | Issue : 9 | September 2014 | ISSN - 2249-555X

Knowledge Discovery from Birth Registration E-Governance Data using Data Mining

KEYWORDS Data Warehouse, Data Mining, Clustering and Association Rules Mining

Pushpal Desai

Assistant Professor, M.Sc. (I.T.) Programme, Veer Narmad South Gujarat University, Gujarat, India

ABSTRACT *This paper discusses about proposed methodology of knowledge discovery from Birth Registration e-governance data of the Surat city. The proposed methodology includes three important phases: Data Preprocessing, Data Warehouse Design and its implementation and Knowledge Discovery using Data Mining. The paper includes discussion about implementation of Clustering and Association Rules Mining algorithms on Birth Registration e-governance data. The results obtained from the practical implementations of Clustering and Association Rules Mining algorithms are also discussed. The results obtained from practical implementation suggest that hidden trends and new relationships obtained from Data Mining implementations can be utilized by the corporation for future planning and decision making.*

1. INTRODUCTION

Inmon defined Data Warehouse as a subject oriented, integrated, nonvolatile, and time variant collection of data in support of management decisions" [7] [8]. Hen and Kamber defined data mining as "Extracting or mining knowledge from large amount of data" [2]. In last few decades, private sector companies have implemented Data Warehouse and Data Mining for knowledge discovery from organizations' OLTP data. The Data Warehouse and Data Mining are widely adopted in Banking, Insurance, Telecom, Medical Science, Education, Weather Forecasting, Network Security, Stock Market, Retail, eCommerce and many other sectors. The companies are using Data Warehouse and Data Mining for knowledge discovery from their operational data. The discovered knowledge is used for organization's future planning and strategic decision making. In India, government bodies like Municipal Corporation have computerized their day-to-day transactions and this has resulted in collection of large amount of data. However, Government organizations in India have not adopted Data Warehouse and Data Mining technologies to uncover knowledge from e-governance data. In this research paper, implementation of Data Mining on Birth Registration e-governance data for the city of Surat is discussed.

2. METHODOLOGY

In the first phase, various data preprocessing tasks on source data are performed. In the second phase, multi dimensional schema for Data Warehouse implementation is designed. OLAP operations like Slice, Dice, Roll-Up, Drill etc…on Data Cube are performed on data analysis. In the third phase, Clustering algorithm is used to identify important clusters from e-governance data. The Association Rules Mining algorithm is applied on important Data Clusters obtained in the earlier step. In this paper, only third phase of proposed methodology is discussed.

2.1 DATA PREPROCESSING

In the Data Preprocessing, Extraction, Transformation and Loading tasks on Birth registration e-governance data are performed. The ETL tasks were implemented using Microsoft SQL Server Integration Services (SSIS). This step is important because data stored in heterogeneous environment is migrated to a common environment and OLTP data is converted into consistent format. Furthermore, error and noise is removed from the OLTP data [1].

2.2 MULTIDIMENSIONAL SCHEMA DESIGN

In this phase, various dimensions, measures and facts for multidimensional schema design were indentified. The multidimensional schema was designed considering analytical needs of the organization and several data cubes were created based on multidimensional schema design and performed various OLAP operations on it. The OLAP was implemented using Microsoft SQL Server Analysis Services (SSAS). The Star Schema design was proposed for Birth registration e-governance data because dimension data for Birth registration rarely changes and query response time is very high for Star Schema design. In earlier research papers, we discussed about various OLAP operations on the Birth registration Data Cube to support analytical needs of the organization [4].

2.3 DATA MINING

From Birth Data Cube, Clustering Model was created considering Registration Year, Registration Date, Birth Date, Religion ID, Birth Location Code, Sex, Mother Age, Marriage Year, Mother Age Year Birth, Delivery Attention ID, Delivery Method ID, Birth Weight, Father Education ID, Mother Education ID, and Zone Fields.

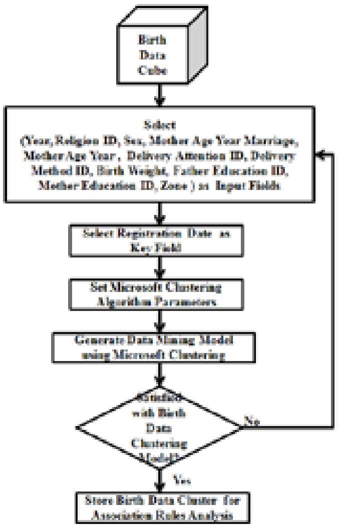
Birth Data Cube
Select (Year, Religion ID, Sex, Mother Age Year Marriage, Mother Age Year , Delivery Attention ID, Delivery Method ID, Birth Weight, Father Education ID, Mother Education ID, Zone) as Input Fields
Select Registration Date as Key Field
Set Microsoft Clustering Algorithm Parameters
Generate Data Mining Model using Microsoft Clustering
Satisfied with Birth Data Clustering Model?
No
Yes
Store Birth Data Cluster for Association Rules Analysis

Fig -1: The proposed Clustering Data Mining Model for Birth Data Cube

The Registration Date was used as the key column for the model. Based on these input fields and key column, we created Microsoft Clustering model for Birth Data Cube [5]. The proposed methodology is depicted in the Figure 1. In our earlier research work, we used Clustering and Association Rules algorithm separately [5][6]. In the proposed model, Association Rules models is implemented using key fields observed in the Clustering Model. From Clustering data mining model important fields such as Zone, Sex, Religion, Education Levels, Delivery Attention and Delivery Method were indentified. In the first experiment, Religion, Zone and Sex as input fields and Delivery Method as predict only attribute were selected. In the second experiment, Religion, Zone and Sex with Delivery Attention given at the time of birth attributes were selected. In the third experiment, Religion, Education, Delivery Method and Delivery Attention attributes were selected to find new relationships. The proposed method is depicted in the Figure 2.

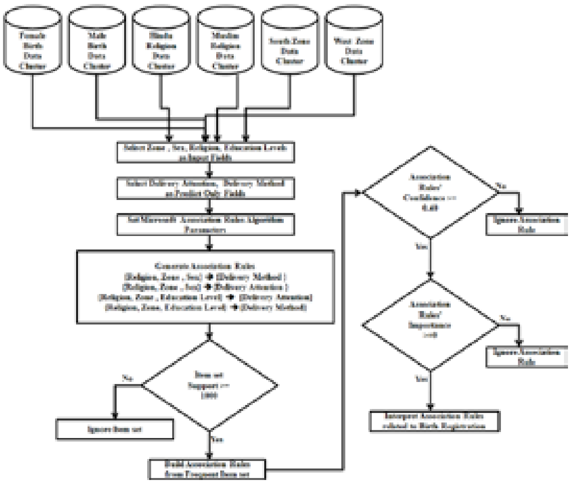
Female Birth Data Cluster
Male Birth Data Cluster
Hindu Religion Data Cluster
Muslim Religion Data Cluster
South Zone Data Cluster
West Zone Data Cluster
Select Zone , Sex, Religion, Education Levels as Input Fields
Select Delivery Attention, Delivery Method as Predict Only Fields
Set Microsoft Association Rules Algorithm Parameters
Generate Association Rules
[Religion, Zone , Sex] → [Delivery Method]
[Religion, Zone , Sex] → [Delivery Attention]
[Religion, Zone , Education Level] → [Delivery Attention]
[Religion, Zone, Education Level] → [Delivery Method]
Item set Support >= 100
No
Ignore Item set
Yes
Build Association Rules from Frequent Item set
Association Rules' Confidence >= 0.40
No
Ignore Association Rule
Yes
Association Rules' Importance >=0
No
Ignore Association Rule
Yes
Interpret Association Rules related to Birth Registration

Fig -2: The proposed model for generating Association Rules from Birth Data Clusters

3 RESULTS

The SQL Server Analysis services provide features to analyze Cluster data [3] . The Clustering Data Mining algorithm divides Birth data into different groups considering various fields and its values.

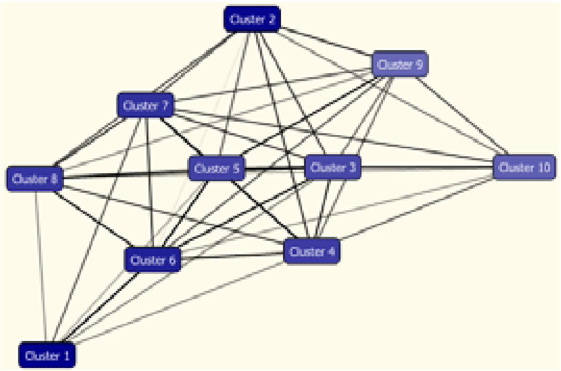

Fig -3: Cluster diagram for Birth e-governance data

The Figure 3 shows cluster diagram of 10 Clusters obtained by using Microsoft Clustering algorithm. The Microsoft Clustering algorithm also provides "Cluster Profile" to better understand and interpret the Clustering Model. The Figure 4 shows "Cluster Profile" for the Birth data. The Cluster profile indicates four different states for "Delivery Attention Name" variable.

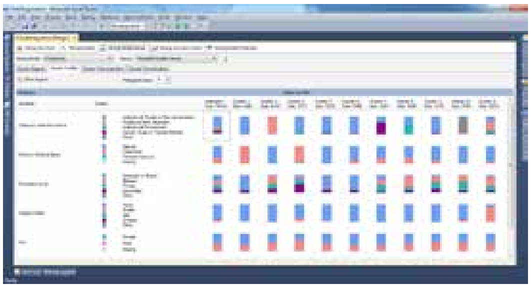

Fig -4: Cluster profile for Birth e-governance data

These states are "Institutional - Private or Non - Government", "Traditional Birth Attendant", "Institutional-Government", "Doctor, Nurse or Trained Midwife" and "Relatives or Other". The "Delivery Method Name" variable has three states "Natural", "Caesarean" and "Forceps/ Vaccum". The Education Level variable has four different states and these are "Graduate or Above", "Illiterate", "Primary" and "Secondary". In case of "Religion Name" field different states are "Hindu", "Muslim", "Jain", "Christian", "Parsi", "Shikh" and "Boddh". The Zone variable has "South East", "South", "East", "West", "Central", "North East", "North" and "South West" states. It was observed that cluster diagram provides too little information where as cluster profile gives us too much information. To resolve this problem and to answer very specific questions such as:

- In which clusters are citizens whose education level is "Graduate or Above" and what kind of "Delivery Attention" mother got at the time of delivery?
- In which clusters are citizens whose education level is "Illiterate" and what kind of "Delivery Attention" mother got at the time of delivery?
- In which clusters are citizens with "Muslim" religion and what is the "Delivery Method" at the time of child's birth?
- In which clusters are citizens with "Hindu" religion and what is the "Delivery Method" at the time of child's birth?

To answer such specific questions, relationship of different variable, its state and individual clusters characteristics were further explored. For example, for "Education Level" variable, "Graduate or Above" state and its result shows that "Cluster 1", "Cluster 7" and "Cluster 8" are having high density for it. The Figure 4 indicates that "Cluster 1" has highest population for "Graduate or Above" state. Further analysis of the "Cluster 1" shows that variable "Delivery Method Name" with value "Caesarean" has 99% probability and variable "Delivery Attention Name" with state "Institutional-Private or Non-Government" has 96% probability in the "Cluster 1".

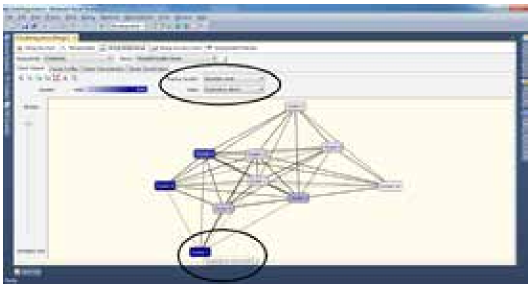

Fig -4: Cluster Diagram for Birth data (Education Level = Graduate or Above)

The probability indicates that presence of certain values in the cluster. The result of "Cluster 1" suggests that in case of educated citizens, mother is most likely to get "Delivery Attention" at "Private Institute" and "Delivery Method" is most likely to be "Caesarean". The "Cluster 1" characteristic is shown in the Figure 5.

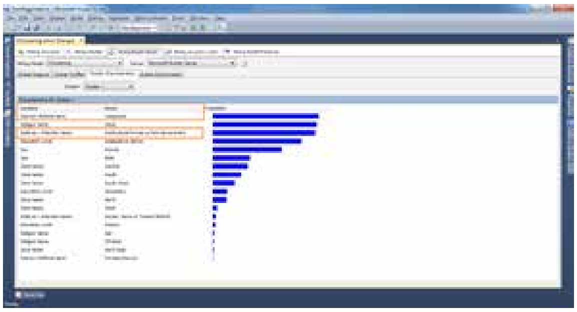

Fig -5: "Cluster 1" Characteristic for Birth data

The "Education Level" variable with value "Illiterate" state was explored. The result shows that, "Cluster 5" is having the highest density for the "Illiterate" state. The "Cluster 5" diagram is shown in the Figure 6.

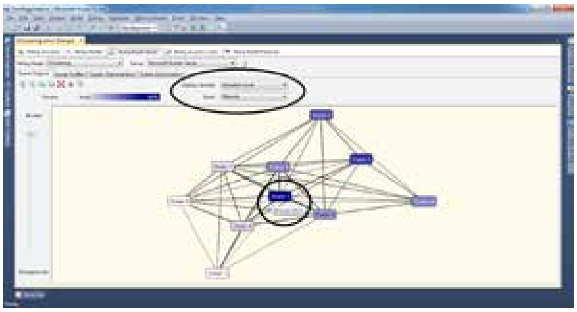

Fig -6: Cluster Diagram for Birth data (Education Level = Illiterate)

The analysis of "Cluster 5" characteristics show contrasting results compare to "Graduate or Above" state. In the "Cluster 5", shows that for illiterate parents, variable "Delivery Method" has value "Natural" with 63% probability and variable "Delivery Attention" has value "Institutional-Government" with 53% probability. The "Cluster 5" characteristics are shown in the Figure 7.

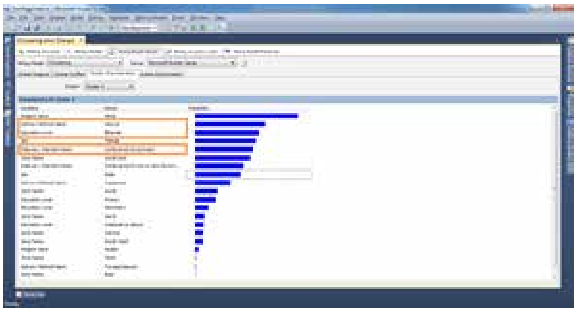

Fig -7: "Cluster 5" Characteristic for Birth data

The other clusters were explored by selecting different shading variables and its states. For example, variable "Religion" was selected with different states like "Hindu", "Muslim" and "Parsi".

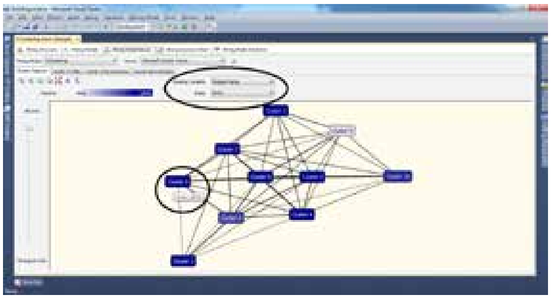

Fig -8: Cluster Diagram for Birth data (Religion Name = "Hindu")

The cluster diagrams for "Hindu" and "Muslim" religions are shown in the Figures 8 and 9 respectively. The results show that most clusters are populated with "Hindu" religion data and only "Cluster 9" is highly populated with "Muslim" religion.

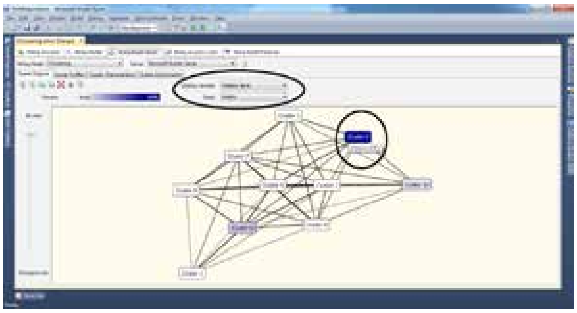

Fig -9: Cluster Diagram for Birth data (Religion Name = "Muslim")

The "Cluster 9" characteristics was analyzed and found that "Deliver Method" variable value is "Natural" with 90% probability as shown in the Figure 10. The variable "Education Level" is also present with value "Illiterate" and has probability of 50%.

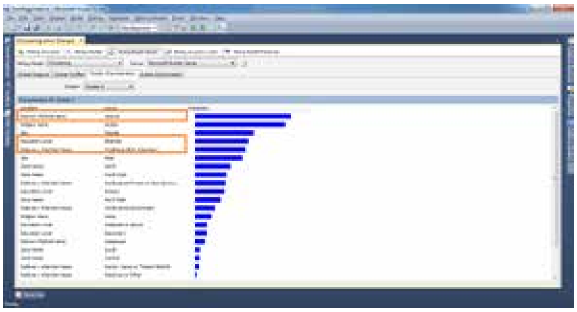

Fig -10: "Cluster 9" Characteristic for Birth data

From the Birth Cluster data, attempt was made to find association rules among Religion Names, Zone Names, Education Level and Delivery Method Name attributes. Three attributes Religion Names, Zone Names and Education Level are antecedents and the Delivery Method Name attribute is consequent. For example, if parents Education Level is "Graduate or above" and Zone is "South West", "South", "Central" or "North" than Delivery Method is likely to be "Caesarean".

Table -1: Association Rules for Zone Names, Education Level = "Graduate or Above" and Delivery Method Name attributes

Rule Conf.	Rule Importance	Association Rules
0.686	0.307338855	Zone Name = South West, Education Level = Graduate or Above -> Delivery Method Name = Caesarean
0.627	0.284129024	Zone Name = Central, Education Level = Graduate or Above -> Delivery Method Name = Caesarean
0.626	0.266437527	Zone Name = North, Education Level = Graduate or Above -> Delivery Method Name = Caesarean
0.479	0.235774425	Education Level = Graduate or Above -> Delivery Method Name = Caesarean
0.522	0.18906587	Zone Name = South, Education Level = Graduate or Above -> Delivery Method Name = Caesarean

Where as if parents are uneducated in that case Delivery Method is likely to be "Natural" for "North East", "East", "Central", "South East", "West, South", "South West" Zones. These association rules are given in the Table 2.

Table -2 Association Rules for Zone Names, Education Level = "Illiterate" and Delivery Method Name attributes

Rule Conf	Rule Importance	Association Rules
0.952	0.184434812	Zone Name = North East, Education Level = Illiterate -> Delivery Method Name = Natural
0.891	0.150231193	Zone Name = East, Education Level = Illiterate -> Delivery Method Name = Natural
0.843	0.121473832	Zone Name = Central, Education Level = Illiterate -> Delivery Method Name = Natural
0.83	0.119765218	Zone Name = South East, Education Level = Illiterate -> Delivery Method Name = Natural
0.828	0.114668323	Zone Name = West, Education Level = Illiterate -> Delivery Method Name = Natural
0.815	0.108869452	Zone Name = South, Education Level = Illiterate -> Delivery Method Name = Natural
0.809	0.102167849	Zone Name = South West, Education Level = Illiterate -> Delivery Method Name = Natural

The novel relationships were found considering different Religions. In this model, Religion Name and Zone as input attributes and Delivery Method as predict only attributes were used. The results suggest that in "North", "Central" and "South West" zones Delivery method is likely to be "Caesarean", where as in case of "North East" zone, Delivery method is likely to be "Natural" for Hindu Religion.

Table -3 Association Rules for Zone Names, Religion Name = "Hindu" and Delivery Method Name attributes

Rule Conf	Rule Importance	Association Rules
0.572	0.232351136	Zone Name = North, Religion Name = Hindu -> Delivery Method Name = Caesarean
0.523	0.206277973	Zone Name = Central, Religion Name = Hindu -> Delivery Method Name = Caesarean
0.503	0.171095265	Zone Name = South West, Religion Name = Hindu -> Delivery Method Name = Caesarean
0.845	0.134017658	Zone Name = North East, Religion Name = Hindu -> Delivery Method Name = Natural

The results of Muslim religion are contrasting with Hindu religion. The Delivery method is likely to be "Natural" for "North East" and "South East" zones where as for "South" zone Delivery method is likely to be "Caesarean" for Muslim religion.

Table -4 Association Rules for Zone Names, Religion Name = "Muslim" and Delivery Method Name attributes

Rule Conf	Rule Importance	Association Rules
0.912	0.157691416	Religion Name = Muslim, Zone Name = North East -> Delivery Method Name = Natural
0.474	0.134200943	Religion Name = Muslim, Zone Name = South -> Delivery Method Name = Caesarean
0.803	0.10031654	Religion Name = Muslim, Zone Name = South East -> Delivery Method Name = Natural

ACKNOWLEDGEMENTS AND LIMITATIONS

All results are based on data provided by the municipal corporation for the research purpose only. Hence results may change, if data mining concepts are applied on actual data sets.

CONCLUSIONS

The results obtained from Clustering and Association Rules Mining algorithm can be utilized by the Health Department of the Municipal Corporation. The Cluster Profile can be utilized to understand important attributes like Delivery Method and Delivery Attention of Birth registration data. The Association Rules Mining results provide novel relationship among important attributes like Zone, Religion Name, Delivery Attention and Delivery Method. These results can be utilized by the Corporation for better planning and service to the citizens.

REFERENCE

[1] Brian Larson (2008), Delivering Business Intelligence with Microsoft SQL Server 2008, McGrawHill. | [2] Hen and Kamber (2011), Data Mining Concepts and Techniques, Morgan Kaufmann Publishers. | [3] Jamie MacLennan et al. (2008), Data Mining with SQL Server® 2008, Wiley. | [4] Pushpal Desai and Dr. Apurva Desai (2011), The study on Data Warehouse Modelling and OLAP for Birth Registration System of the Surat city, In the proceedings of ICTSM 2011, CCIS 145, pp. 160–167, 2011. Springer-Verlag Berlin Heidelberg 2011 | [5] Pushpal Desai and Dr. Apurva Desai (2011), The Study on Data Warehouse and Data Mining for Birth Registration System of the Surat City, International Journal of Computer Applications, Number 4 - Article 2, 2011, pp. 1-5, ISBN: 978-93-80746-63-0. | [6] Pushpal Desai and Dr. Apurva Desai (2012), An empirical analysis based on association rules mining on E-Governance system, In the proceedings of International Conference & Workshop on Recent Trends in Technology 2012, TCET, Mumbai, India | [7] W. H. Inmon (2005), Building the Data Warehouse, Wiley | [8] W. H. Inmon et al. (2001),Corporate Information Factory, Wiley |