Visual Data Mining in Indian Election System

Prof. T. M. Kodinariya
Asst. Professor, Department of Computer Engineering,
Atmiya Institute of Technology & Science, Rajkot
Gujarat, India
trupti.kodinariya@gmail.com

Mr. Ravi Seta
Department of Information Technology,
Atmiya Institute of Technology & Science, Rajkot
Gujarat, India
Mr.raviseta@Gmail.com

Abstract - A good leaders or Government is the basic need to develop country. In India, who is largest democratic country in the world people are not fully involved in the selection process of Leaders. On an average there are 60-65% voting is done. For this purpose we have build the Data Warehouse that containing all the information related to election to increase awareness of voting. Using this we can find interesting patterns that are extracted and represented using Visual Data mining to arrange awareness program. The approach is divided into 5 phases: I) Data Preprocessing: II) Data Warehouse Creation: III) Task-Relevant Data Extraction; IV) Data Mining and V) Visualization. Generally, data collected from various cities of a state may be noisy, inconsistent and requires cleaning hence the database is preprocessed for missing values, normalization etc. in first phase. In Data Warehouse Creation phase, warehouse with vote as measure in fact and voter Gender, voter age, voter education, candidate, Religion, time, session, word as dimension. Word dimension has 4-level concept hierarchy of country, state, city, area/word id. Time dimension has concept hierarchy with year, Quarter, month, and date. Similarly, candidate has 2 level and Voter age, voter education, Religion, and session have level 1 concept hierarchy. In Last phase, results of data mining phase are represented using Visual Data mining techniques.

Keywords- Association, Concept Hierarchy, Election, Data Mining, Dimension, Data Warehouse, Fact, Visual Data Mining.

I. INTRODUCTION

The growth of a country depends upon good leader. In democracy, selection of leader is carried out by election procedure hence election system becomes a pillar of Indian democracy. The survey of Indian election data reveals that average 60-65% citizen contributes in voting. Rarely any party gets clear "Bahumat" hence it requires support of other parties which lead to mixture of party leader dominating Indian government rather than single party leader. That affects the approval of any bill from parliament.

To improve voting percentage, we have to aware a citizen about importance of his vote in growth of country. We have to analyze the education level, age level, population of different region of India to carry out different voting awareness street shows.

In this paper, we are focusing on analysis of such types of data and creating data warehouse for election system data and applying OLAP and visual data mining on the data.

II. PROPOSED ARCHITECTURE

In Proposed approach, data warehouse of candidate of election and voters are maintained, interesting patterns are extracted and represented using Visual Data Mining. The approach is divided into 5 phases: 1) Data Preprocessing; 2) Data Warehouse Creation; 3) Task- Relevant Data Extraction; 4) Data Mining and 5) Visual Data Mining as shown in Fig.1

A. Data Pre-processing

Generally, preprocessing of data ultimately reduces the elapsed time needed to retrieve the information from a data warehouse. There are number of data preprocessing techniques like data cleaning, data integration, data transformation and Normalization etc.

In the proposed system, we focus on only Data Cleaning and data Transformation. Some of information related to voter and candidate are missing, which was represented by constant "UNKNOWN" in data cleaning process. During data transformation, voter age is transform from numerical valve into categorical form like youth, middle and senior. Similarly, voter education level is transform into three different categorical valve ill-literal, school level and graduate level.

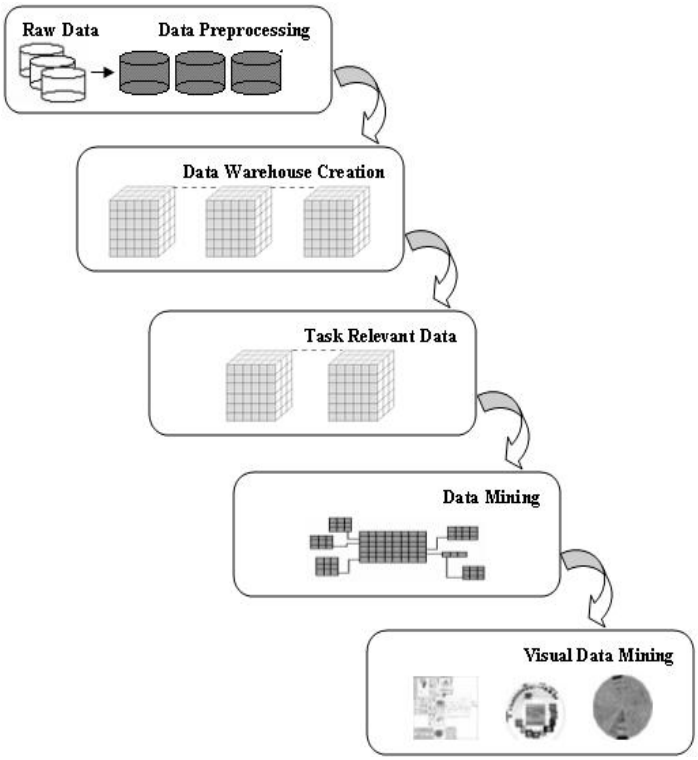


Figure 1: Proposed Architecture

B. Data Warehouse Creation

The entity-relationship model is used in the design of relational databases while multidimensional model is used for warehouse. Preprocessed data will be in the form of Relational Data Base. Using this Relational Data base a data cube is generated and populated with the data to create the warehouse. A data cube allows data to be viewed and modeled in multiple dimensions. Data cube is defined by dimensions and facts. Dimensions are the perspectives or entities with respect to which an organization wants to keep records. The fact table contains the names of the facts, or measures, as well as keys to each of the related dimension tables. Normally, data cube is n-dimensional [1].

There is a need to define concept hierarchy on each dimension to provide a sequence of mappings from a set of low-level concepts to higher-level, more general concepts. For example location dimension, city valves for location include Rajkot, Jamnagar, Surat, Delhi, Mumbai, Chennai etc. Each city, however, can be mapped to the province or state to which it belongs. Similarly, each state further mapped to country. These mappings form a concept hierarchy for the dimension location, mapping a set of low-level concepts (i.e., cities) to higher-level, more general concepts (i.e., countries). Concept hierarchies may be provided manually by system users, domain experts, or knowledge engineers, or may be automatically generated based on statistical analysis of the data distribution.

Word dimension has 4-level concept hierarchy of country, state, city and area/word id as shown in Fig. 2. Similarly, Time dimension has concept hierarchy with 4-level of year, quarter, month and Date. Concept hierarchy for remaining dimension is summarized in table 1.

Table 1: Concept Hierarchy for Dimension

	Dimension							
	Word	Time	Candidate	Religion	Education	Age	Voter	Session
0	Country	Year	C_Name	Name	Education Group	Age Group	Gender	S_name
1	State	Qtr	Party Name					
2	City	Month						
3	Word Id	Date						

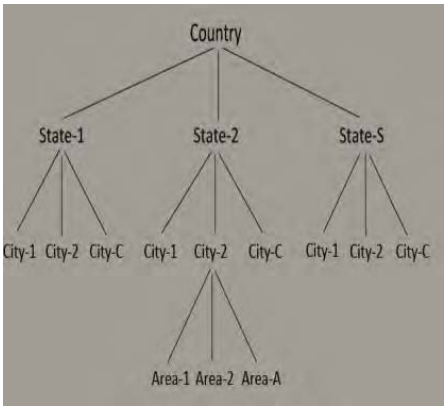


Figure 2: Concept Hierarchy for Location Dimension

For the election voting solution, an 8-Dimensional cube is created. A "Vote" is considered as a measure in the fact table while Word, Candidate, Time, Voter Education, Voter (Gender), Voter Age, and Voter Religion, Word and Session as Dimensions of the cube. Dimension and fact are not sufficient for warehouse but selection of appropriate multidimensional model is also important. The "Snow-Flake Schema" is selected for the Election system data warehouse. The snowflake schema is shown in Fig. 3.

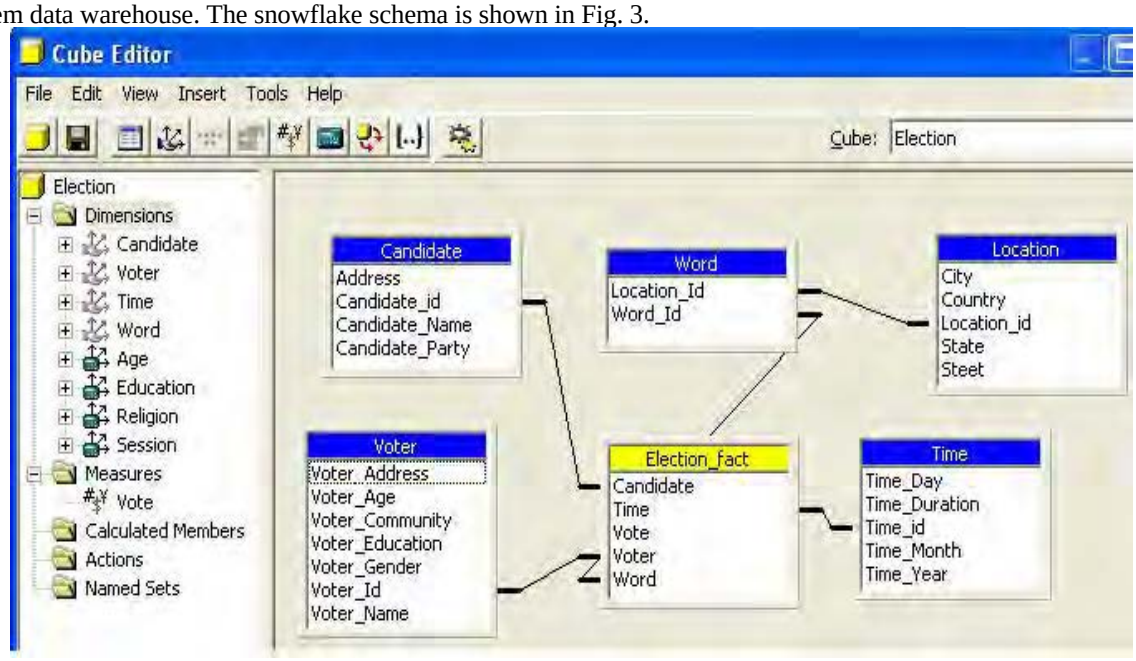


Figure 3: The Snowflake Schema for election system

In relational database, we have only 6 tables namely voter, candidate, time, location, word and Election_fact table. Using member properties like Voter_age, Voter_education and Voter_religion of Voter Dimension, we create 3-virtual dimension such as voter age, voter education and voter religion. Similarly, session member property of time dimension will create Session Dimension. The advantage of virtual dimensions is that they do not increase the size of the cube and increase the speed of processing of the cube. The data aggregations of virtual dimensions is not stored but calculated in memory. As a result if a virtual dimension is added to a cube the storage usage and processing time do not increase.

C. Task Relevant Data Extraction

Once cube is populated with the data, based on the knowledge to be mined, there is a need to extract the task relevant data. i.e. portion of the database or data warehouse is investigated. So, for the Election System Data warehouse, if a person is only interested in studying the association between Candidate and Word (Part of Word Dimension), the task relevant data can be specified with the following information:

- The name of the data warehouse (e.g., Vote Bank)
- The names of the tables or data cubes (in this case, the data cube is the data warehouse, but it could very well be part of a cube in a warehouse)
- *Conditions* for selecting relevant data (e.g., retrieve data pertaining to Candidate located in city)
- *The relevant attributes or dimensions* (e.g., *state* and *city* from *word* dimension)

D. Data Mining

Microsoft® SQL Server™ 2000 extends and renames the former OLAP Services component, now called Analysis Services. Analysis Services introduces data mining, which can be used to discover information in OLAP cubes and relational databases. Proposed Approach is implemented using Microsoft SQL Server 2000 with Analysis Services installed [2]. The Cube Editor is shown in Fig. 4.

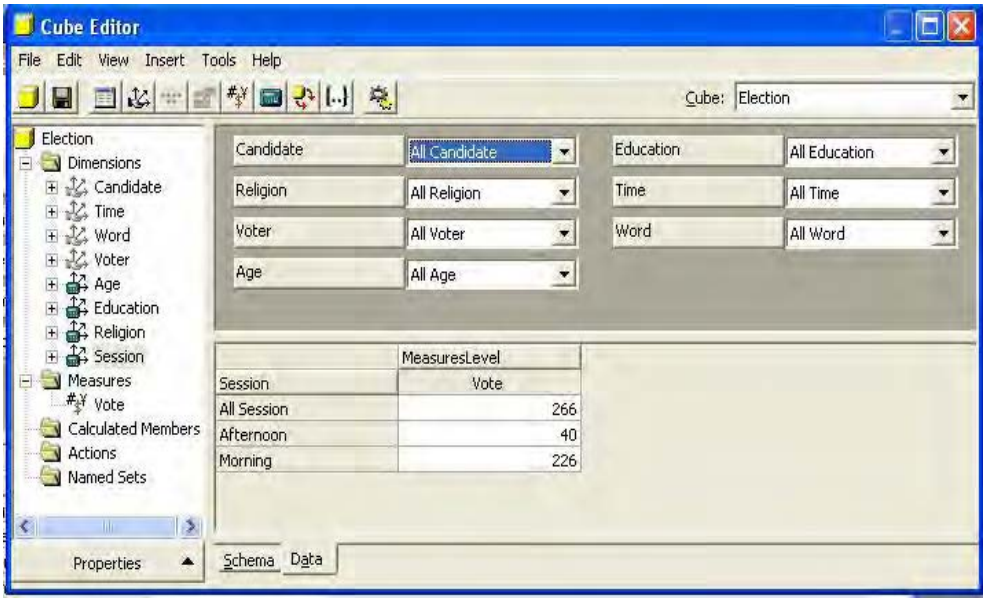


Figure 4: Detail of voting during different session in India

Output shown is based on the "Sample Election Database". Left Pan of Cube editor contains details of Dimensions, Measures, Calculated Member, and Actions etc. Upper Right Pan shows the details of the available dimensions based on which data can be extracted while Lower Right Pan contains the output of the selected dimension.

The total voting in country during morning and afternoon session is shown in Fig. 4. The dice operation defines a sub cube by performing a selection on two or more dimensions. The result of dice operation on dimensions age, religion and word with respect to youth, muslim and Gujarat selection is shown in Fig. 5.

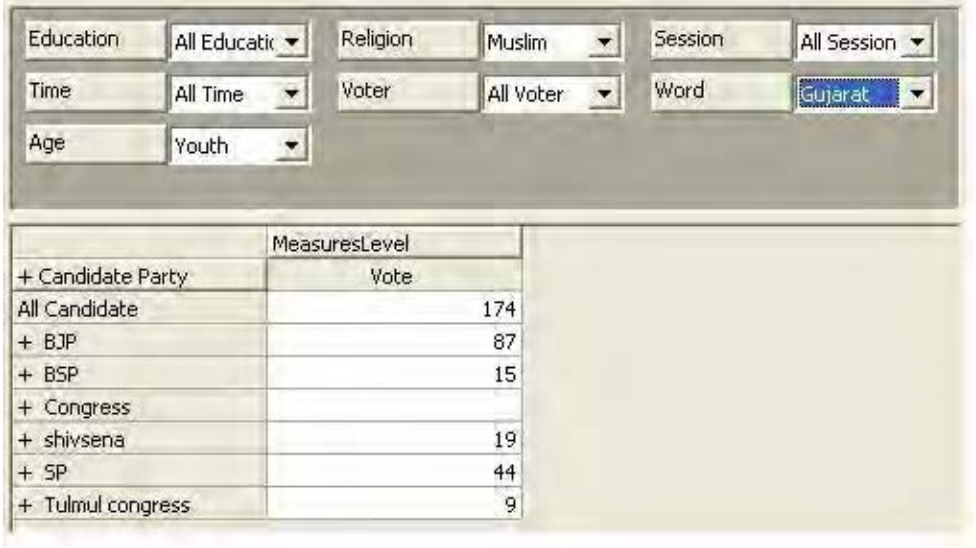
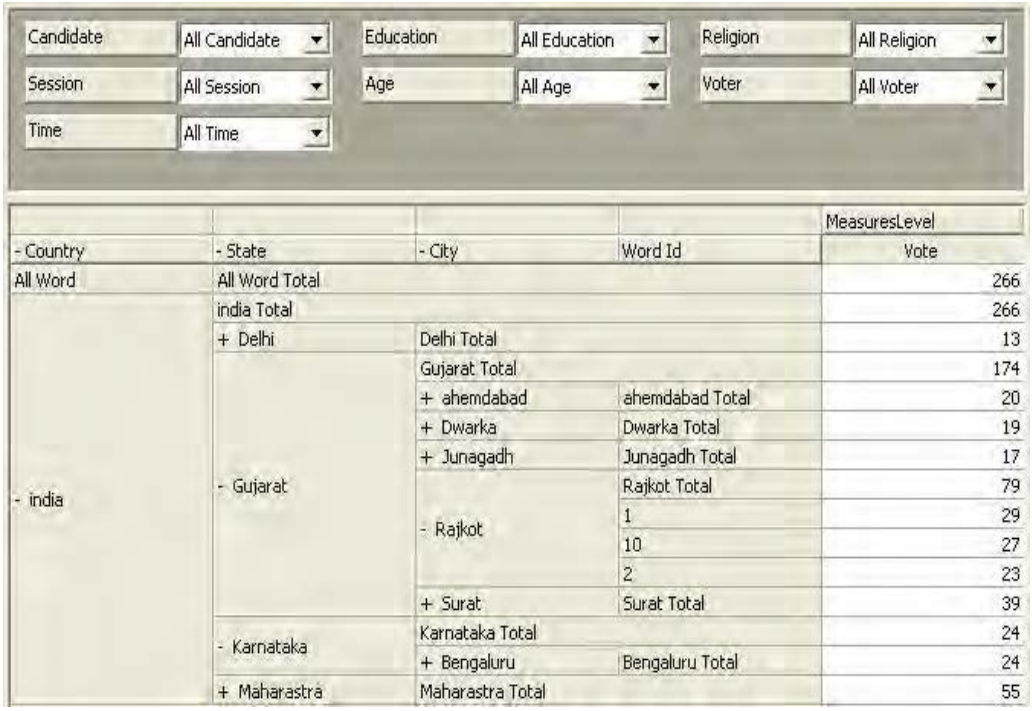


Figure 5: Detail of voting by "youth" age, "muslim" community in "Gujarat" state to different parties.

OLAP operations also include rollup (increasing the level of aggregation) and drill-down (decreasing the level of aggregation or increasing detail) along one or more dimension hierarchies. The result of drill down operation on dimension word is shown in Fig. 6.



Candidate	All Candidate	Education	All Education	Religion	All Religion
Session	All Session	Age	All Age	Voter	All Voter
Time	All Time				

				MeasuresLevel
- Country	- State	- City	Word Id	Vote
All Word	All Word Total			266
- india	india Total			266
	+ Delhi	Delhi Total		13
	- Gujarat	Gujarat Total		174
		+ ahemdabad	ahemdabad Total	20
		+ Dwarka	Dwarka Total	19
		+ Junagadh	Junagadh Total	17
		- Rajkot	Rajkot Total	79
			1	29
			10	27
			2	23
		+ Surat	Surat Total	39
	- Karnataka	Karnataka Total		24
		+ Bengaluru	Bengaluru Total	24
	+ Maharastra	Maharastra Total		55

Figure 6: Drill down operation on dimension word

E. Visualization

Consider the Association Rule Mining between Word and Education once again. Assume that there are 'A' areas in all cities, 'C' cities in all state, 'S' different states in country.

Different Possible Location keys are: A*C*S
Different Possible Education Levels are: 3
Thus, Different possible association rule for these two dimensions are: 3*A*C*S

If we assume value of 'a' 60, value of 'c' 40 and value of's' is 30 then number of association rules result into 2,70,000. We can obtain result of all these rules through cube editor but it is not possible to represent all these rules at once, to do so we need to use the visualization techniques like Pixel-Oriented, Icon-Based, Hierarchical, Graph-Based and Geometric Technique. Visualization techniques are used mainly for explorative analysis, confirmative analysis and presentation of facts [3].

Technique of visualization which is appropriate is pixel oriented technique as ours is an 8-dimensional data. Each dimension of the cube can be represented as a pixel. The Group of eight pixels represents one data cube tuple. Normal image size is 1024x1024 which is divided by 2x4 i.e. 2,70,000 tuples can be easily accommodated in this image. It is clear from this example that we can easily create a city wise 'Vote Map' if we slightly increase the size of the image. Fig.7 shows 'Election map' with 8 dimension created using Pixel-Oriented Technique.

Another well-known technique of visualization is Parallel Co-ordinate. Mondrian is a general purpose statistical data visualization system written in JAVA [4].

It features outstanding visualization techniques for data of almost any kind. Currently implemented plots comprise Mosaic Plot, Scatter Plots and SPLOM, Maps, Barcharts, Histograms, Missing Value Plot, Parallel Coordinates/Boxplots and Boxplots y by x. Mondrian works with data in standard tab-delimited ASCII files. There is basic Support for working directly on data in Databases (please contact me for further info).The current version of Mondrian has been tested on MacOS X, Windows XP and (various) Linux. Fig.8 shows the result of parallel co-ordinate on the sample data of Election system. It is clear from Education dimension in the figure that Education attribute have three levels. Also Age attribute have the same three levels, and Session dimension have two attribute that is morning and afternoon.

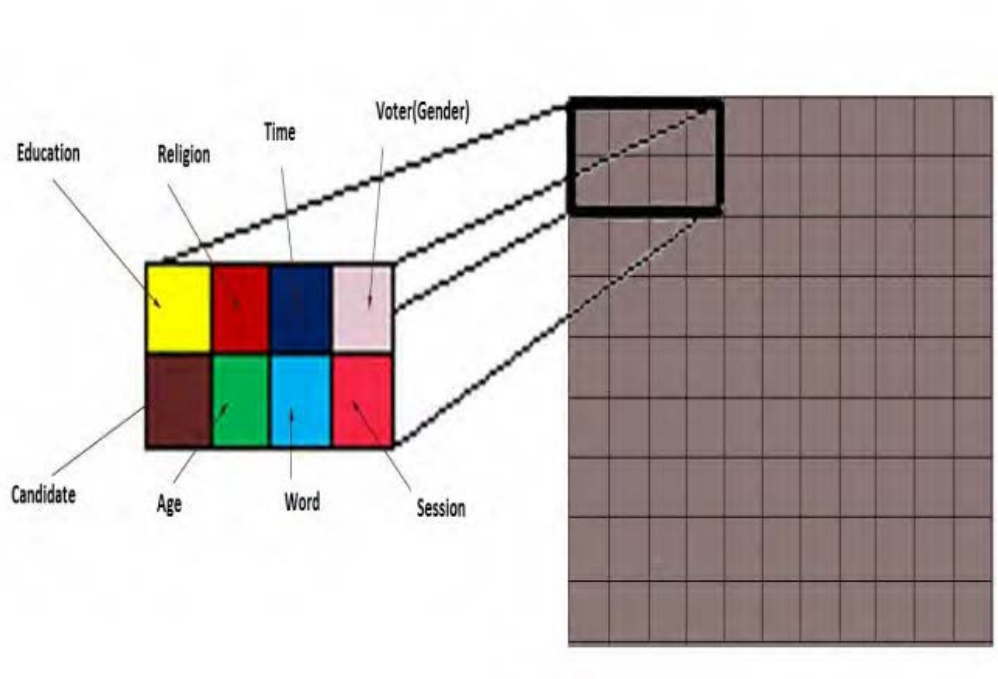


Figure 7: 'Election Map' generated using Pixel-Oriented Technique

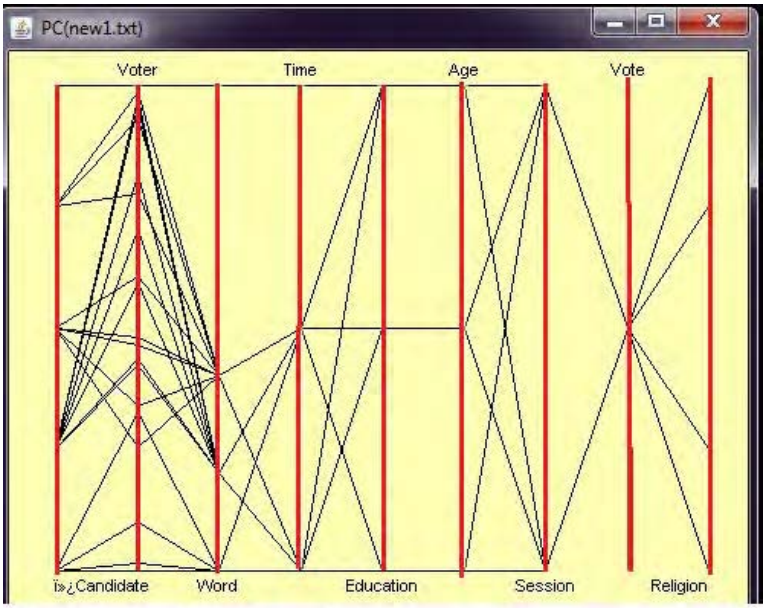


Figure 8: Parallel Co-ordinate output of Sample Data Using Mondrian

III. CONCLUSION

Visualization of the output generated using visual data mining techniques facilitates decision makers to make decisions like fluctuation in voting; voting ratio of male to female; voting rating of different party in a city/state. This type of analysis helps to arrange appropriate awareness program in different regions.

REFERENCES

[1] Data Mining Concepts & Techniques, Jiawei Han & Micheline Kamber, Morgan Kaufmann.
[2] Microsoft SQL Server 2000 Analysis Services help at *www.microsoft.com*
[3] Visual Data Mining : Techniques and Tools for Data Visualization & Mining, Tom soukup & Ian Davidson, John Wiley Publication
[4] Mondrian Software's Help Document at www.rosuda.org/Mondrian/

AUTHORS PROFILE

Trupti Kodinariya received M. E. Degree with specialization in image processing from Dharmasinh desai Univestity, Nadidad in may 2005. Her research interests are data warehousing and data mining, web mining, compiler design, theory of computation. She published many papers in national and international conference and Journals.
Presently she is working as Assistant professor at Atmiya Institute of Technology and Science-Rajkot. She is life time member of Indian Society of technical Education

Ravi Seta Born in Patnagadh, Orissa, Date of birth is 07-may-1990. He received Bachelor Degree of Engineering in Information Technology from Atmiya Institute of Technology & Science , Rajkot (Gujarat).
Her research interests are data warehouse & mining, computer networks and data structure.