Implementation of Data Mining and Data Warehousing In E-Governance

Sonali Agarwal
IIIT Allahabad,
India

Neera Singh
IIIT Allahabad,
India

Dr.G.N.Pandey
IIIT Allahabad,
India

ABSTRACT
India is a big democratic nation having multilevel administrative authorities from national level to state, district and block levels. Large amount of data is generated and disseminated by different government departments at various levels of administration. It is important to integrate the different department in terms of data sharing so that all departments can work under a single controlling authority without repetition of work. It is important to develop a framework for creating a centralized nationwide data warehouse which has horizontal as well as vertical interconnections having limited accessibility at lower level authorities and can be fully accessed at the higher levels. Use of efficient Data Warehousing and Data Mining techniques may surely enhance government decision making capabilities. A nationwide Data warehouse model especially for Indian context has been proposed. An agriculture data has been taken and mined to analyze the climatic and weather changes which affect the yield of the crop. This can help in predicting the climatic conditions so that better precautions can be taken to improve the agricultural output. .

Keywords Data Mining, data warehousing, Data Mart, E Governance

1. INTRODUCTION
India is fast moving towards modernization. It has or is in the process of upgrading and has computerized most of its departments. Now most of the information is stored digitally, although certain sections are following both methods of data storage that is computerized and manually. This had led to the creation of a huge Data warehouse at all levels i.e. national, state, district and block. Depending on the data, it can be categorized into Universal (social security number, PAN number etc.), Departmental (Health, Income Tax etc.) and Operational data (for day to day operations e.g. loan EMIs etc.)

The idea is to develop a data warehousing architecture where information is stored at all administrative levels, is stored only once to avoid repetitions and is accessible at all levels. To facilitate this there needs to be data isolation and integration. At lower levels (district and block) data accessibility has to be more discreet and only available to limited personnel and departments and at higher levels (national) it should be easily linked and available at all quarters. Thus, there is a need for more isolation at lower levels and more integration at higher levels.

Also there is a need to make this data available at functional levels, e.g. PAN Number is a unique ID for a lot of citizens but the individual has to produce it on demand because the personnel in governing authority still doesn't have the access to this information.

2. BASIC BUILDING BLOCKS OF PROPOSED E GOVERNANCE MODEL
The proposed E Governance model covers all important aspect of E Governance has been shown as Figure 1. There are four Basic Building Blocks of proposed E Governance Model. The lowest block is the Administration Block, which regulates the overall function of any country through efficient government.

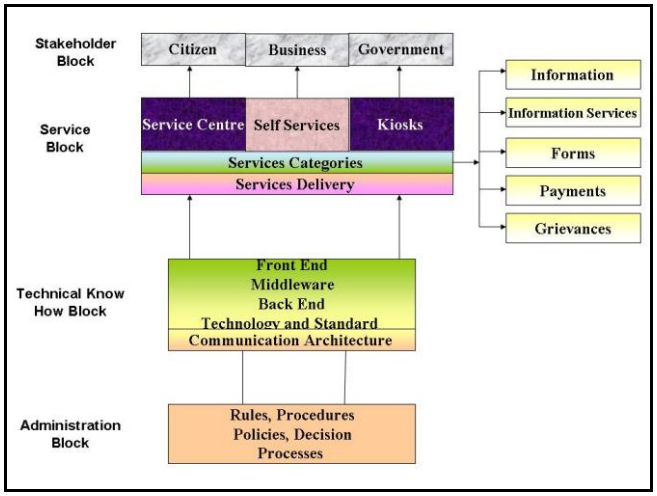

Fig. 1: Basic Building Blocks of Proposed E Governance Model

The Technical knowhow block includes computerization of manual processes, commonly agreed technological standard, Database related applications and easy access of information. The third block is Service Block, which includes all available operations of the E Governance. The upper block is Stakeholder Block, which has various categories of users that interact with the system. The user may be a Citizen, Business organization or any Government organization [4] .

3. PROPOSED DATA WAREHOUSE MODEL
Data Warehouses comprise the "building blocks" of the Data Mining based approach. A National Data Warehouse is an essential component of the proposed E Governance model. The objective of maintaining a National Data Warehouse is to integrate all distributed database at one place to facilitate an easy and quicker view of all its historical data. Different view of data

could be utilized in different report formats and reflect performance of the working organization. Figure 2 indicates the National level integration of various Data Warehouses. The proposed Data warehouse model is a 4 level system. Different levels of Data Warehouses interconnected together by using client-server architecture. In the proposed model, the District Data Warehouses are considered to be clients with respect to the State centers, and State centers are clients with respect to the central office.

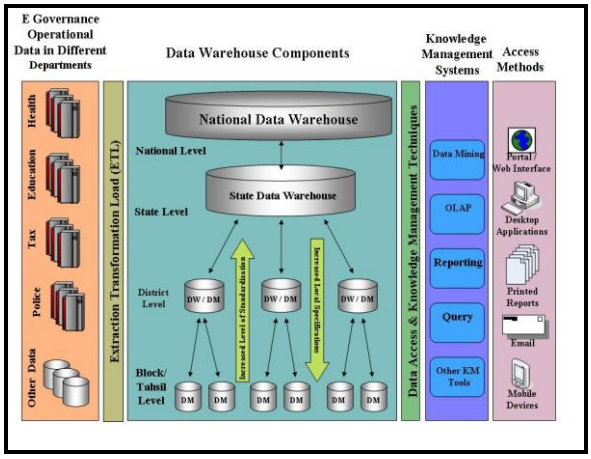

Fig.2.: Data Warehouse Model

The integration of Data Warehouses at District level and State level is shown in Figure 3. A State Data Centre has a state level database which provides summarized view of the data items. State Data Centre interconnects different separate department level database at District level and State level. Figure 4 explains different levels along with their interconnection of State Data Center at State and District Level databases of different Departments.

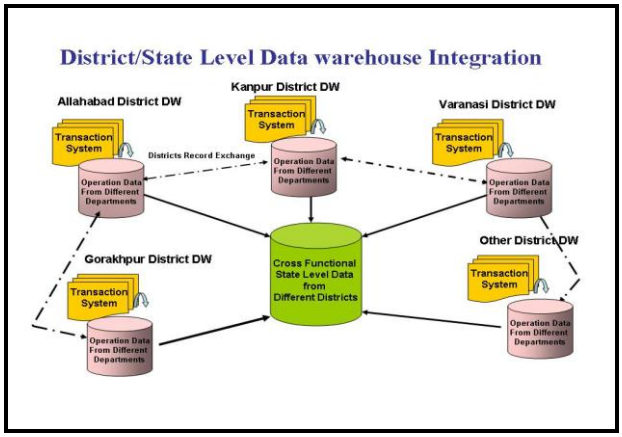

Fig.3. Integration of Data Warehouses in E Governance

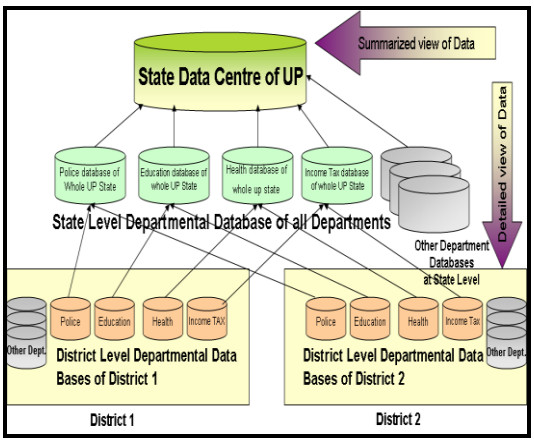

Fig.4. Interconnection of State Data Center at State and District Level

District level Database contains detailed view of data items. Only selected data items may transmit from District level to State level Database. An example has been taken from Transport department in which flow of data and information from a local Transport office to District level and State level is explained by using Figure 5.

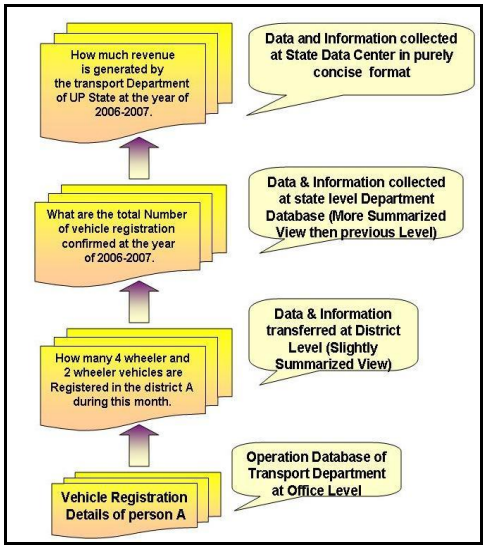

Fig.5. Flow of information from lower level to upper level

All department level databases should be interconnected and integrated in a State Level Data Warehouse in such a way so that any data updating in one department may easily reflect to the other Department. Any change in data fields could be automatically updated in all related department by using in single step.

Fig.6. Working of State Data Warehouse

Figure 6 shows steps for data updating in all concern departments with the help of state Data Warehouse. In Figure 6 a citizen changed his address by using a citizen portal. This change in address may automatically update his address in various concerned departments like election commission, Bank, Income Tax, Passport, Land Record and RTO Departments.

4 WORKING OF STATE DATA CENTRE

The flowchart given in figure 7, explains the working of State Data Center. Any government department may become source of new record related with any citizen. The new data or record is first needed to be checked in existing state database. If that record is not exists in the state database then that record may be created after proper authentication of appropriate department. If the record is exist in the state database then the old record has been updated only after the proper authentication of the appropriate department. This updating process held at state data centre as well as all related departments which are concerned with the data field.

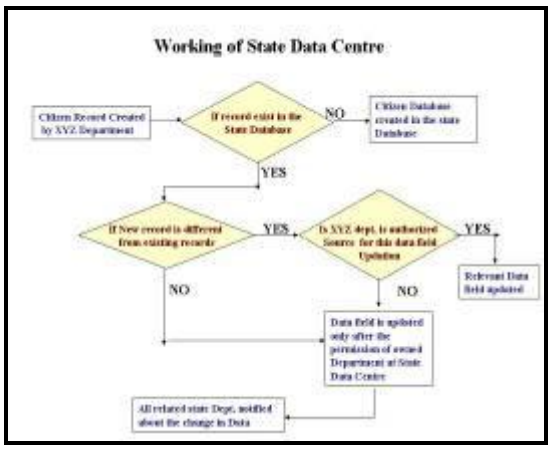

Fig.7. Flow Chart of State Data Center

There are following rules and responsibilities needed to be followed by the State Data Center and different department.

- Data capturing should be performed at the source system. That may be various State and District level departments.
- Any change regarding existing data field or new data must reflect in the State Data Centre in accordance with data ownership and data security policies.
- All type of Data Communication in between the participating departments may implemented as a process of automatic notifications, data file transfer or through reports [6].

5 DATA MODEL CONSIDERATION

A State level data center must be open and flexible to accommodate the information needs of many departments. The State level data center also has responsibility to ensure data quality, data usability and validity. The process of data creation, data updation and data maintenance is thoroughly checked by using different policies. The data categories are [7]:

- **Universal Data:** Universal data elements are very common data field which is useful for all departments. Universal data elements are shared by all the departments. The examples are Name, Address and Source System ID etc.
- **Departmental Data:** These data elements are created by one department but they may be useful for another department. The examples the data elements are bank account number, income or income indicators like asset holdings, land holdings and payments made.
- **Operational Data:** These data elements are useful for the concern department only. These data elements are detailed and may be very large in number [6].

Table 1. Policies for data creation, updating, maintenance and security

Data	**Data Creation Policy**	**Maintenance Policy**	**Security Policy**
Universal Data	Data must be created by a central authority. Data updating rights are also reserve to the central authority.	Data updating is only after proper authentication . For example address from Passport, Date of Birth from High School Mark sheet.	Data may be viewed by any department in read only format. Data Updation is only after approval.

Departmental Data	Departmental data may be created by any department according to their functional need.	Permission is required for the updating of any data field from its owning Department.	Data may be viewed by any department in read only format. Data Updation is only after approval.
Operational Data	Typically excluded from the Centralized Data Center.	Data viewed through queries	Data viewed through queries but not available at the Centralized Data Center

A National Data Warehouse must possess all of the above mentioned Data Elements. All these data elements are created, maintained and updated by using following policies

6 DATA MINING IN AGRICULTURE

Clustering, Classification, Association rule and Data visualization are the well known Data Mining Techniques. WEKA is a comprehensive open source Data Mining tool provides large range of Data Mining algorithms [1][2]. WEKA is suitable for Data Mining because it can handle easily available and maintainable Comma Separated Value (CSV) file formats [8].

Here the research work has objective to mine agriculture data containing climatic and crop related variables to explore interesting hidden and useful patterns. Rice, Temperature and Sugarcane data in CSV format has been taken for the experimental work [11].

Rice dataset with four attributes area, production, yield and % irrigation has been examined to model the production of rice and its dependency on other variables. The Expectation Maximization Algorithm suggests that the whole dataset can be divided into 4 different clusters having similar characteristics. Each cluster with its members may be semantically understood with the help of generated cluster description by EM algorithm as shown in Figure 8 [9] [10].

```
EM
==

Number of clusters selected by cross validation: 4


                Cluster
Attribute            0          1          2          3
                  (0.24)     (0.24)     (0.28)     (0.24)
=============================================================
area
  mean               43    37.0875    32.6119    40.1915
  std. dev.      1.0279     1.0555     1.9572     0.9519

production
  mean           78.7586    40.0135    28.0202     54.266
  std. dev.        5.601     2.3987     4.2228     5.9024

yeild
  mean         1829.5048 1078.4077   855.7712 1348.4871
  std. dev.      91.3204    40.0575    91.8349   126.6359

arri
  mean           49.0834    38.1189    35.0919    42.0413
  std. dev.        2.557     0.6423     1.9932     1.5284

Clustered Instances

0       12 ( 24%)
1       12 ( 24%)
2       14 ( 28%)
3       12 ( 24%)
```

Fig.8. Cluster description as output of EM Algorithm

With Data visualization approach it is apparent that the addition of fertilizers to the sugarcane crop increases the yield of sugarcane. However not all fertilizers are leading to an increase in total sugar recovery (1/2 V.C and ½ NPK fertilizers and combination of $1/4^{th}$ V.C and $3/4^{th}$ NPK fertilizers are leading to a decline in total sugar recovery).Hence it is important to focus on fertilizers which increase the sugarcane yield and sugar recovery as well.

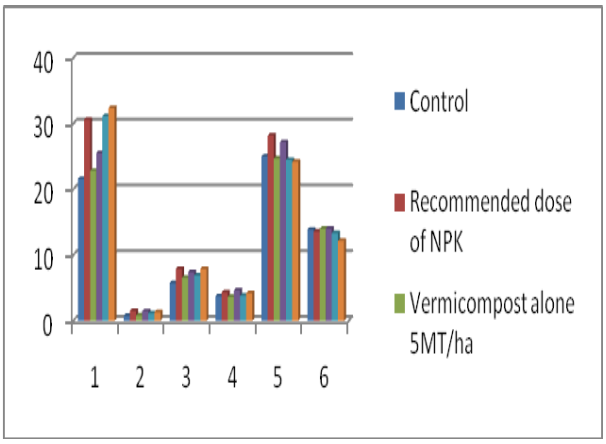

Fig.9. Showing the changes in various parameters (cane yield, cane weight, cane height, girth of internodes, number of internodes and sugar recovery) of sugarcane crop on addition of fertilizers. Fig. 10. and Fig. 11. representing to correlation between sugar recovery and sugarcane yield with usage of fertilizers.

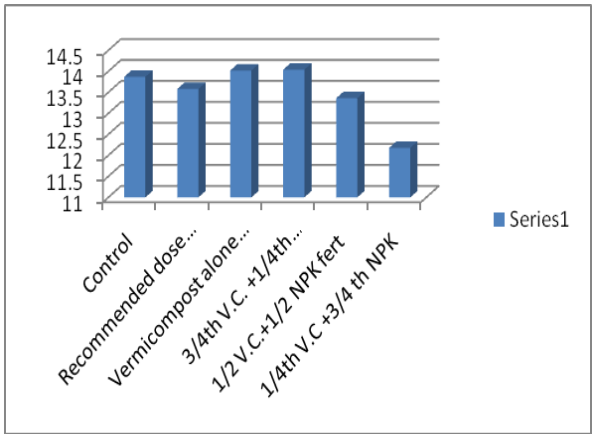

Fig.10. Showing effect on sugar recovery with usage of fertilizers

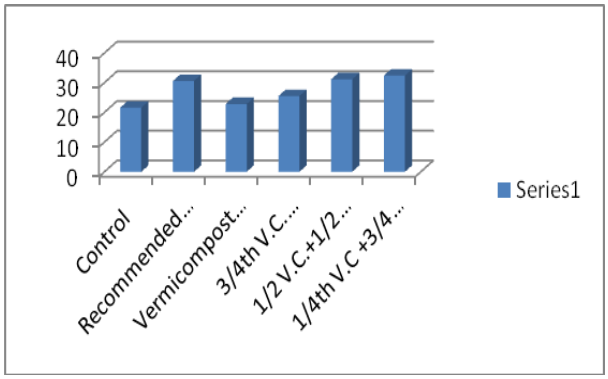

Fig.11 Showing the increase in sugarcane yield on treatment with fertilizer

7 CONCLUSION

A nationwide data warehouse is a feasible idea however it come with certain bottlenecks. There are issues related to privacy and security of the data. Also the amount of resources needed to maintain such a voluminous data source has to be justifiable. Nevertheless the concept carries huge potential and can resolve lot of India's problem concerning management and governance.

REFERENCES

1. S. Džeroski, A. Kobler, V. Gjorgijoski, P. Panov Using Decision Trees to Predict Forest Stand Heght and Canopy Cover from LANDSAT and LIDAR data. 20th Int. Conf. on Informatics for Environmental Protection – Managing Environmental Knowledge – ENVIROINFO 2006.
2. "WEKA 3: Data Mining Software in Java" (n.d.) Retrieved March 2007 from http://www.cs.waikato.ac.nz/ml/weka/`
3. Graham Williams, (2006) Data Mining Desktop Survival Guide http://www.togaware. com /datamining/survivor/Usage2.html.
4. "About Kiosk", (n.d.) E Governance of Government of West Bengal, Retrieved December 2006 www.wbgov.com/E-Gov/ENGLISH/Kiosk/AboutKiosk.asp
5. U.S. General Account Office (GAO)(2004) "Data Mining Federal Efforts Cover a Wide Range of Uses" GAO-04-548, http://www.gao.gov/new.items/d04548.pdf
6. Thomas Zwahr and Matthias Finger, (2004) "Enhancing the e-Governance model: Enterprise Architecture as a potential methodology to build a holistic framework" Proceedings of the International Conference on Politics and Information System: Technologies and Applications. Orlando, Florida, USA
7. Marcos M. Campos (2005) "Data-Centric Automated Data Mining" Proceedings of the Fourth International Conference on Machine Learning and Applications (ICMLA'05)0-7695-2495-8/05 $20.00 © 2005 IEEE
8. Jain A.K, Murty M.N., Flynn P.J., (1999) "Data Clustering: A Review" ACM Computing Surveys, 31, 3:264-323.
9. Apte C. & Weiss S.M. (1997) "Data Mining with Decision Trees and Decision Rules" T.J. Watson Research Center http://www.research.ibm.com/dar/papers/pdf/fgcsapteweiss_with_cover.pdf
10. S. Džeroski, A. Kobler, V. Gjorgijoski, P. Panov Using Decision Trees to Predict Forest Stand Heght and Canopy Cover from LANDSAT and LIDAR data. 20th Int. Conf. on Informatics for Environmental Protection – Managing Environmental Knowledge – ENVIROINFO 2006.
11. Official website of Government of India, Ministry of agriculture. http://agricoop.nic.in.