Proceedings of the Symposium on "Application of Database and Data Mining"

Chairman : Prof. Prem Narain

Convenor : Dr. R.C. Goyal

The following four papers were presented in the symposium.

1. Design and Development of Data Mart for Animal Resources in India – presented by Dr. Anil Rai, IASRI, New Delhi.
2. Need of Small Area Database for E-Governance in Uttar Pradesh – presented by Dr. L.R. Yadav, NIC, Lucknow.
3. An Appraisal for Database Resources in Dairying – presented by Dr. Ravinder Malhotra, NDRI, Karnal.
4. Statistical Data Mining – presented by Dr. V.K. Bhatia, IASRI, New Delhi.

After discussions at the end, following recommendations were emerged

1. For designing and development of databases for agriculture at country level, there is a strong need to evolve a standard coding system.
2. Small area databases may be developed at different levels, which finally may become source for data warehouse at country level.
3. There is a need to undertake studies to identify and evaluate suitable data mining techniques in the field of agriculture.

Design and Development of Data Mart for Animal Resources in India

Anil Rai, V.N.B.S. Madhavanand and P.K. Malhotra
Indian Agricultural Statistics Research Institute, New Delhi-110012

A data mart on animal resources in India is a logical subset of proposed data warehouse on agricultural resources in India, under NATP Mission Mode project "Integrated National Agricultural Resources Information System". Under the above project an attempt has been made to integrate the databases of

different subject areas in the form of data marts of a data warehouse for agricultural resources. This warehouse will help planners, researchers and development agencies for efficient use in this important area of agriculture. This paper presents the steps that are needed in developing a datamart and provides an overview about the reporting and cube building through which user is enabled to do Online Analytical Processing (OLAP) over the proposed datamart especially in the context of animal resources of the country.

The animal wealth of India is represented by a broad spectrum of native breeds of cattle, buffalo, goat, sheep, equines and camels. In addition several other forms like poultry, duck, geese, quails, yak, mithun, pigs are also important components of animal wealth of India and contribute a big share in animal production. India is bestowed with a large population of domestic animals with 204.5 million cattle, 83.5 million buffalo, 50.8 million sheep, 115.2 million goat, 1.99 million equines, 1.03 million camel, 12.8 million pigs, 0.06 million yaks and 0.13 million mithun. This national resource is a major contributor to national economy. Livestock sector accounts for 21 % of the value of output of agricultural sector, which is 29 % of the gross domestic product of the economy.

The livestock census is conducted every five years in India. This provides age and sex wise data at the district level on different categories of animals. State Animal Husbandry Departments, and National Sample Survey Organization are conducting surveys on production, prices and utilization of animal products all over the country. Surveys are also being conducted by different research organizations on animal diseases, animal genetic resources, migration of livestock etc. Still huge data gaps persist in the data on animal resources. These gaps are there because the data collected is not being compiled in its entirety. Most of data has not been computerized till now. Moreover a centralized database on animal resources does not exist in the country. Datawarehousing on animal resources can be of immense help for sustainable development of these resources in the country by collecting, compiling and analyzing these information centrally. The relationship of animal resources with other sectors of the country can also be studied through this warehouse.

Planners, researchers, development agencies and farmers require Information on animal resources for further studies and evolving realistic strategies for improvement and rearing of livestock and poultry. The data is also required for keeping a watch on prices and movement of animal products, animal feed, and establishment of services such as veterinary hospitals, artificial insemination (AI) centers, meat and dairy industries etc. Further, there is a need to study animal resources in relation to other aspects of agriculture, such as soils, vegetation, and climate, socio-economic, land use, water resources for overall development of agricultural production system. In a region the spatial data combined with attribute data on all these aspects of agricultural sectors and animal resources, which is also a component of this project, will lead to

development of geographical information system useful for planners and makes decision makers to visualize the spatial distribution of these agricultural resources and take judicious decision.

In this paper, the core concepts that are general to datawarehouses and data marts are applied on data pertaining to animal resources in the country for building data marts in general and data pertaining to the census of the livestock across agricultural years in particular.

Need of Small Area Databases for E-Governance in Uttar Pradesh

L.R. Yadav
National Informatics Centre, MCIT, Government of India, Lucknow-226001

The recent democratic decentralisation process initiated by the 73rd and 74th Amendments of the constitution of India giving greater responsibilities and powers to the Panchayats and Nagar Palikas as the third tier of governance offers a new window of opportunity for local planning, effective implementation and monitoring of various social and economic development programmes in the country at the local levels. Inspite of Central, State and District plans, Central and State governments have initiated a number of schemes like MPLAD, Border Area Development, Vidhayak Nidhi, Purvanchal, Bundelkhand, Paschimanchal, Madhyanchal Nidhies etc. in recent years for the local area development. There exists a decentralised statistical system in the country from the early seventies, by which information are collected/compiled from gross root level to national level. Over the period attempts have been made to strengthen the data collection system from local levels (Village/Village Panchayat/Block/District levels).

After the implementation of 73rd and 74th Amendments of the constitution, administrative readjustment of local level workers has also been made in Uttar Pradesh. Presently there are mainly 2 workers at grass-root levels. One is Patwari looking after the revenue related works of allotted villages and another is Multipurpose Officer in every Village Panchayat looking after the developmental activities. Patwari is responsible for the maintenance of Khatauni, Khasra, Sajara, Agriculture/Livestock Census etc. and Multipurpose Officer is responsible for Kutumb/Pariwar Register, Arthic Register, Village Level Basic Amenities (VLBA) information etc. Other local level information are also collected through Population Census, National Sample Survey Organisation etc. from time to time.

VLBA information on about 40 parameters whether it is available in the village itself or the distance from where it is available have been computerised as a joint project of NIC and Planning Department, Government of Uttar Pradesh (GOUP) along with Sankhyakiya Patrika (SP) from the year 1994. These databases are available from the year 1995-2002 and updation for the year 2003 is in progress. NIC has also developed a Web-based Query System by which different types of queries can be generated through the bilingual website of Planning Department. SP databases can also be retrieved through the website. Any individual can also obtain his Khatauni from the computers installed in the tehsils.

As per recommendation of National Statistical Commission, India proforma for data callection at different levels have been redesigned and are being incorporated in the IT model of Planning Department. In this paper experiences gained in implementation of IT solutions and strengthening it further has been discussed in detail.

An Appraisal of Database Resources in Dairying

D. K. Jain, Adesh K. Sharma and Ravinder Malhotra
National Dairy Research Institute, Karnal-132001 (Haryana)

The database development efforts were originally started in the late 80's at the Computer Centre on the basis of the demands of different user groups. These subdatabases were developed separately on different aspects using tools like Unibase/Unix, HIcalc/Unix, Lotus 1-2-3/MS-DOS, dBase III Plus/MS-DOS, MS-FoxPro/ MS-Windows, etc. Only a few years back, structured efforts were initiated to convert all such scattered databases into a single unified database system with a graphical user interface (GUI) for easy storage and retrieval of information through the advent of object oriented application programming tools like MS-Visual Basic which was commercially available in India only in late 90's.

Database management system (DBMS) is a collection of programs that facilitates the user to store, modify, and extract information from a database. From a technical standpoint, DBMSs can differ widely. The terms relational, network, and hierarchical all refer to the way a DBMS organizes information internally. The internal organization can affect how quickly and flexibly information can be extracted. An RDBMS (abbreviation for relational database management system) is a type of DBMS that stores data in the form of related tables. Relational databases are powerful because they require few assumptions

about how data is related or how it will be extracted from the database. As a result, the same database can be viewed in many different ways. Another important feature of relational systems is that single database can be spread across several tables. Requests for information from a database are made in the form of a query. The set of rules for constructing queries is known as a query language. Different DBMSs support different query languages, although there is a semi-standardized query language called SQL (structured query language). Sophisticated languages for managing database systems are called fourth-generation languages (4GLs).

Graphical user interface (GUI) is a computer working environment that represents files and operations visually, using icons, buttons, windows, and other imagery that one can manipulate with a mouse. Hence, GUI is a program interface that takes advantage of the computer's graphics capabilities to make the program easier to use. Well-designed graphical user interfaces can make the user free from learning complex command languages by allowing use of graphic objects rather than just words to represent the input and output of a program.

Microsoft Visual Basic 6.0 is one such programming language that provides a visual programming environment for developing user interfaces by using sophisticated controls such as buttons and dialog boxes, and then defining appearance and behavior. Although not a true object-oriented programming language in the strictest sense, yet Visual Basic nevertheless offers an object-oriented philosophy. It is sometimes called an event-driven language because each object can react to different events such as a mouse click. For this reason Visual Basic was selected as the front-end tool to develop the graphic user interface for the present database system.

Several relational-database oriented information systems e.g. National Dairy Scene, Global Dairy Scene, Web Enabled Information System for Dairy Cultures, MSI-NDRI, Management Information Service etc. have been developed at this Institute on production, processing and management aspects of dairying.

Moreover, some more set of reports, viz., Breeding Efficiency Report, Heifers having not come in heat report, Animals having taken five or more services report, Pregnancy diagnosis report, etc. are being prepared from the database which is being regularly maintained on all the related parameters and are being sent to the concerned Division/Section every month for their perusal and corrective action, wherever needed. For this purpose, data on different parameters periodically include "Expected Producing Ability (EPA)" and "Expected Transmitting Ability (ETA)" of animals for assisting the concerned division in taking culling decision.

Statistical Data Mining

V. K. Bhatia
Indian Agricultural Statistics Research Institute, New Delhi-110012

Data Mining

We are drowning in information but starved for knowledge: John Naisbitt

Data mining sits at the interface between statistics, computer science, artificial intelligence, machine learning, database management and data visualization etc.

Data mining is the process of identifying valid, novel, potentially useful, and ultimately comprehensible knowledge from data (bases) that is used to make crucial decisions. Data mining is not a product that can be bought. Data mining is a discipline and process that must be mastered a whole problem solving cycle. The main part of data mining is concerned with the analysis of data and the use of software techniques for finding patterns and regularities in sets of data. The idea is that it is possible to strike gold in unexpected places, as the data mining software extracts patterns not previously discernible or so obvious that no one has noticed them before. The analysis process starts with a set of data, uses a methodology to develop an optimal representation of the structure of the data during which time knowledge is acquired. Once knowledge has been acquired this can be extended to larger sets of data working on the assumption that the larger data set has a structure similar to the sample data. This is analogous to a mining operation where large amounts of low-grade materials are sifted through in order to find something of value. The choice of a particular combination of data mining techniques, to apply in a particular situation, depends on both the nature of the data-mining task to be accomplished and the nature of the available data. From a statistical perspective, many data mining tools could be described as flexible models and methods for exploratory data analysis. In other words many data mining tools are nothing else than multivariate (statistical) data analysis methods.

Data mining techniques have been used for knowledge discovery in many sectors including insurance, retail, meteorology, agriculture, animal science and medicine to name just a few. These techniques exploit methodologies from several fields including statistics, artificial intelligence, information theory, uncertainty theory, fuzzy logic and heuristics. They are used for applications such as clustering data objects, classifying objects into groups or hierarchical structures, regression and association. Until recently, data mining techniques had not been used widely for biological research. Recent applications include clustering expression profiles in a microarray analysis. Statistical techniques are widely used to study linkage disequilibrium between markers and the linkage

disequilibrium mapping of traits is an important area of research. A new European/US project, GENE-MINE is developing bioinformatics tools for the analysis of germplasm data. Under this, by using statistical data mining techniques, one can study genome wide linkage disequilibrium and possibly map complex traits from haplotype data.

Before a data-mining study is formulated, it is essential that we must avoid a cycle of unrealistic expectations, which may result unnecessary disappointment in the end. It is advisable that we must understand the facts instead so that our data mining efforts are successful. Finally it is seen that data mining should only be used to discover patterns and relationships in the data in order to arrive at better decisions. Thus data mining cannot be ingored - the data is there, the methods are numerous, and the advantages that knowledge discovery brings to a study are tremendous. *'The secret of success is to know something that nobody else knows'* is the real essence of data mining.

The important aspects of statistical data mining, include the foundations of probability, the foundations of statistical data analysis, and most of the classic machine learning and data mining algorithms. These include classification algorithms such as decision trees, neural networks, Bayesian classifiers, support vector machines and case-based (non-parametric) learning. They include regression algorithms such as multivariate polynomial regression, MARS, Locally Weighted Regression, GMDH and neural nets. And they also include other data mining operations such as clustering (mixture models, k-means and hierarchical), Bayesian networks and Reinforcement Learning.

In view of above developments taken place in the theory and application of statistical data mining techniques, an attempt is made in the present paper to highlight some of the issues.