"Data Mining: Knowledge Discovery in Databases"

Authors

[1.] **Dr. Maulik N. Pandya**
HOD, Assistant Professor
M.Sc. IT Department
Shri A. N. Patel P. G. Institute
Anand

[2.] **Mr. Bhavin B. Patel**
Assistant Professor
M.Sc. IT Department
Shri A. N. Patel P. G. Institute
Anand

Abstract *- Database is a technology for data loading, storing, manipulating, querying, sharing and controlling. Leading-edge technology areas, such as data warehousing, web-based applications, object oriented databases, distributed databases, and front end tools are being increasingly utilized by organizations to manage large volume of generated data and information like a Data Mining. This paper explores various statistics methods related with data exploration. Data Exploration is mainly dependent upon the factors like (A) Summary statistics (B) Visualization (C) Online Analytical Processing (OLAP). Thus Data exploration is quite useful to understand the trend, behavior of data and information provided by data mining techniques.*

Keywords: *Data Mining, Data Exploration, OLAP, Data Visualization,*

1. Volumes of Data – The Biggest Challenge

- ❖ The largest challenge a data miner may face is the sheer volume of data in the warehouse.
- ❖ It is quite important, then, that summary data also be available to get the analysis started.
- ❖ A major problem is that this sheer volume may mask the important relationships the analyst is interested in.
- ❖ The ability to overcome the volume and visualize the data becomes quite important.

2. What is Data Mining

Data Mining can be define in simple words as Extracting useful information from large data sets. Data mining is the process of using raw data to infer important business relationships. Despite a consensus on the value of data mining, a great deal of confusion exists about what it is. It is a collection of powerful techniques intended for analyzing large datasets. There is no single data mining approach, but rather a set of techniques that can be used in combination with each other.

One can say that Data mining is the process of discovering meaningful new correlations, patterns and trends by sifting through large amounts of data stored in repositories, using pattern recognition technologies as well as statistical and mathematical techniques. Data mining is the process of exploration and analysis, by automatic or semi-automatic means, of large quantities of data in order to discover meaningful patterns and rules here automatic and semi-automatic means "feeling it shortchanged the role of data exploration and analysis".

3. Introduction of Data Exploration

Data Exploration is a preliminary investigation & exploration of the data to better understand its characteristics & dimensions. Data exploration includes the activities like selecting the perfectly required tools preprocessing or analysis also to recognize the patterns in narrowly focused data.

The availability of very large volumes of data in the electronic form has created a problem of deriving from them useful, task-oriented knowledge. Traditional data analysis techniques, which include statistical and numerical methods, are oriented primarily toward the extraction of quantitative data characteristics, and as such have inherent limitations. For example, statistical techniques cannot produce conceptual descriptions of dependencies among data items or explain reasons why these dependencies exist. Nor can they justify found relationships in the form of higher-level logic-style descriptions, or draw an analogy between the discovered regularity and regularity in another domain.

To address such tasks as above, a data exploration system has to be equipped with a substantial amount of background knowledge, and be able to perform symbolic reasoning involving that knowledge and input data. Unless our data project is very narrowly focused on answering a specific question determined in advance (in which case it has drifted more into the realm of statistical analysis than of data mining), an essential part of the job is to review and examine the data to see what messages it holds.

4. Process-Framework in Data Mining

Although all data mining endeavors are unique, they possess a common set of process steps:

4.1 Infrastructure preparation: choice of hardware platform, the database system and one or more mining tools

4.2 Exploration: looking at summary data, sampling and applying intuition

4.3 Analysis: each discovered pattern is analyzed for significance and trends

4.4 **Interpretation:** Once patterns have been discovered and analyzed, the next step is to interpret them. Considerations include business cycles, seasonality and the population the pattern applies to. 4.5 **Exploitation:** this is both a business and a technical activity. One way to exploit a pattern is to use it for prediction. Others are to package, price or advertise the product in a different way.

5. Data Mining & Data Exploration - Where to apply

Data mining is used in a variety of fields and applications. The military might use data mining to learn what roles various factors play in the accuracy of bombs. Intelligence agencies might use it to determine which of a huge quantity of intercepted communications are of interest. Security specialists might use these methods to determine whether a packet of network data constitutes a threat. Medical researchers might use them to predict the likelihood of a cancer relapse.

6. Application Areas and Opportunities in Data Mining

- Marketing: segmentation, customer targeting
- Finance: investment support, portfolio management
- Banking & Insurance: credit and policy approval
- Security: fraud detection
- Science and medicine: hypothesis discovery, prediction, classification, diagnosis
- Manufacturing: process modeling, quality control, resource allocation
- Engineering: simulation and analysis, pattern recognition, signal processing
- Internet: smart search engines, web marketing

7. Classes of applications of Data Mining

- Market analysis, target marketing, customer relation management, market basket analysis, cross selling, market segmentation.
- Risk analysis
- Forecasting, customer retention, improved underwriting, quality control, competitive analysis.
- Fraud detection
- Text (news group, email, documents) and Web analysis.

Data Exploration, also encompassing a broader area of Data Visualization, involves the application of various technologies to examining large collections of data for structure, patterns, faults, and other characteristics. Data Exploration, while related; to the large universe of data mining (since it is concerned with exposing patterns and relationships in the data) take, in general, a more human centered approach to this pattern discovery. Exploration involves tools and techniques that are used by analysts and researchers to retrieve and examine data through what is often an interactive and intuitive-based process of trial and error. As an approach to understanding data, Data Exploration generally involves four broad technologies:

- **Statistical descriptions of the data,**
- **Structured queries against databases,**
- **Multi-dimensional visualization**
- **Automatic clustering and organization of data around common features**

Data Exploration typically centers on content analysis in a particular context. Exploration is usually a less formal activity than data mining but is often a more critically important process for the analyst, engineer, architect, and systems developer.

As with data mining, there are no specifications as to how these methodologies are to be implemented. But the analogy to actual exploration is very enlightening. An exploration is an activity in which any of a great number of paths and techniques might be utilized and it may take place over a very long period of time. The key to managing such an exploration is to be organized. Keeping records about the exploration, recording your thoughts and ideas along the way, and organizing your findings are all important. This is a complex undertaking, though possibly very rewarding.

8. Data Exploration is mainly dependent upon the following factors

(A) Summary statistics

(B) Visualization

(C)Online Analytical Processing (OLAP)

(A) Summary statistics

Summary statistics are numbers that summarize properties of the data .Summarized properties include frequency, location and spread. Most summary statistics can be calculated in a single pass through the data

Examples: location – mean

Spread - standard deviation

Frequency and Mode

The frequency of an attribute value is the percentage of time the value occurs in the data set. For example, given the attribute 'gender' and a representative population of people, the gender 'female' occurs about 50% of the time. The mode of an attribute is the most frequent attribute value. The notions of frequency and mode are typically used with categorical data

Percentile

For continuous data, the notion of a percentile is more useful. Given an ordinal or continuous attribute x and a number p between 0 and 100, the pth percentile is a value of x such that p% of the observed values of x are less than. For instance, the 50th percentile is the value X50% such that 0% of all values of x are less than X50%.

Measures of Location: Mean and Median

The mean is the most common measure of the location of a set of points. However, the mean is very sensitive to outliers. Thus, the median or a trimmed mean is also commonly used.

$$\mathrm{mean}(x) = \bar{x} = \frac{1}{m}\sum_{i=1}^{m} x_i$$

(B) Visualization

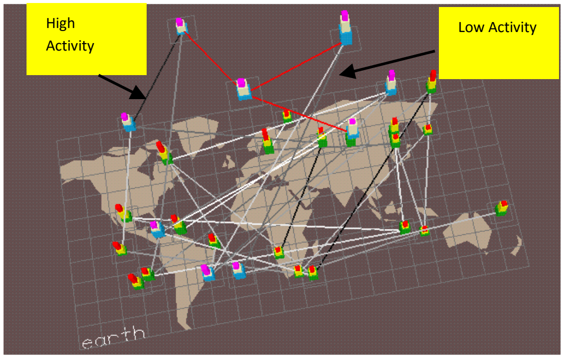
Visualization is the conversion of data into a visual or tabular format so that the characteristics of the data and the relationships among data items or attributes can be analyzed or reported. Visualization of data is one of the most powerful and appealing techniques for data exploration. – Humans have a well developed ability to analyze large amounts of information that is presented visually

– Can detect general patterns and trends

– Can detect outliers and unusual patterns

Exploration by Projections

Example: Global Private Network Activity



Example: An "Enlivened" Risk Analysis Report

Federal Funding Risk Mgt
Unit Risk ($000)
Sensitivity & Unit Risk Analysis
Yield Curve Level & Shape
Two Day Sensitivity Comparis
USD
YTM

Representation

Is the mapping of information to a visual format Data objects, their attributes, and the relationships among data objects are translated into graphical elements such as points, lines, shapes, and colors.

Example:

– Objects are often represented as points

– Their attribute values can be represented as the position of the points or the characteristics of the points, e.g., color, size, and shape

– If position is used, then the relationships of points, i.e. whether they form groups or a point is an outlier, are easily perceived.

Selection

It is the elimination or the de-emphasis of certain objects and attributes. Selection may involve the choosing a subset of attributes

– Dimensionality reduction is often used to reduce the number of dimensions to two or three

– Alternatively, pairs of attributes can be considered

Selection may also involve choosing a subset of objects

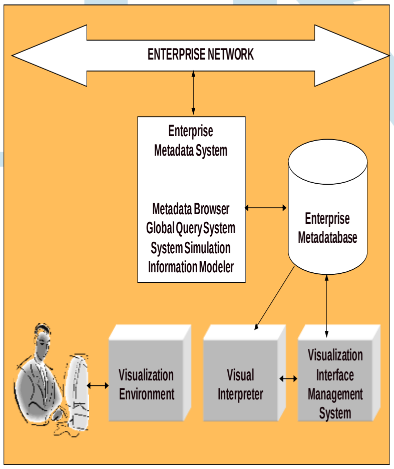
– A region of the screen can only show so many points

– Can sample, but want to preserve points in sparse areas

9. Components of Future Visualization Applications

- ❖ The data visualization environment links the critical components and enables the smooth flow of information among the components.
- ❖ In the future, the bounds between computers, graphics and human knowledge will become more blurred.
- ❖ Many advances in technology will be needed to handle the visualization environment of the future. Intelligent file systems and data management software will contend with thousands of coupled storage devices.

Figure1. Conceptual Mapping of Information Architecture



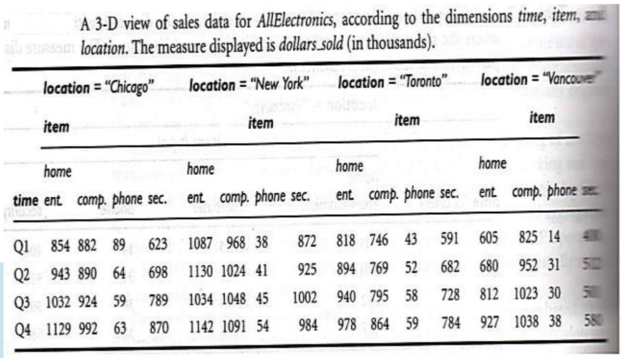
(C) OLAP

On-Line Analytical Processing (OLAP) was proposed by E. F. Codd, the father of the relational database. Relational databases put data into tables, while OLAP uses a multidimensional array representation. Such representations of data previously existed in statistics and other fields. There are a number of data analysis and data exploration operations that are easier with such a data representation.

10. OLAP Operations

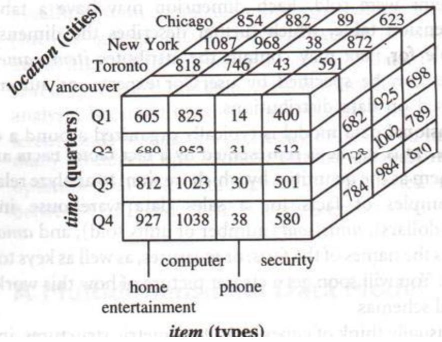
The key operation of an OLAP is the formation of a data cube a data cube is a multidimensional representation of data, together with all possible aggregates. By all possible aggregates, we mean the aggregates that result by selecting a proper subset of the dimensions and summing over all remaining dimensions. For example, if we choose the species type dimension of the Iris data and sum over all other dimensions, the result will be a one-dimensional entry with three entries, each of which gives the number of flowers of each type.

10.1 Data Cube Example



A 3-D view of sales data for *AllElectronics*, according to the dimensions *time*, *item*, and *location*. The measure displayed is *dollars_sold* (in thousands).

	location = "Chicago"				location = "New York"				location = "Toronto"				location = "Vancouver"			
	item				item				item				item			
time	home ent.	comp.	phone	sec.	home ent.	comp.	phone	sec.	home ent.	comp.	phone	sec.	home ent.	comp.	phone	sec.
Q1	854	882	89	623	1087	968	38	872	818	746	43	591	605	825	14	[illegible]
Q2	943	890	64	698	1130	1024	41	925	894	769	52	682	680	952	31	[illegible]
Q3	1032	924	59	789	1034	1048	45	1002	940	795	58	728	812	1023	30	[illegible]
Q4	1129	992	63	870	1142	1091	54	984	978	864	59	784	927	1038	38	580



CA 3-D data cube representation according to the dimensions time, item and location. The measure displayed is dollars_sold (in thousands) Consider a data set that records the sales of products at a number of company stores at various dates. This data can be represented as a 3 dimensional array There are 3 two dimensional aggregates (3 choose 2), 3 one-dimensional aggregates, and 1 zero-dimensional aggregate (the overall total)

10.2 OLAP Operations: Slicing and Dicing

Slicing is selecting a group of cells from the entire multidimensional array by specifying a specific value for one or more dimensions. Dicing involves selecting a subset of cells by specifying a range of attribute values. This is equivalent to defining a sub-array from the complete array. In practice, both operations can also be accompanied by aggregation over some dimensions.

10.3 OLAP Operations: Roll-up and Drill-down

Attribute values often have a hierarchical structure.

– Each date is associated with a year, month, and week.

– A location is associated with a continent, country, state (province, etc.), and city.

– Products can be divided into various categories, such as clothing, electronics, and furniture. Note that these categories often nest and form a tree or lattice

– A year contains months which contains day

– A country contains a state which contains a city this hierarchical structure gives rise to the roll-up and drilldown operations.

– For sales data, we can aggregate (roll up) the sales across all the dates in a month.

– Conversely, given a view of the data where the time dimension is broken into months, we could split the monthly sales totals (drill down) into daily sales totals.

– Likewise, we can drill down or roll up on the location or product ID attributes.

Conclusion:

It’s a human mentality that one can understand any concepts or techniques or information in better way if content is displayed with some graphics and pictorial representation. Data exploration does the same thing. It uses various graphics techniques to explain the information and trends in better way.

Reference:

1. Data Mining Concepts & Techniques by Jiawei Han and Micheline Kambler
2. Principal of Data Mining by David Hand,Heikki Mannila and Padhraic Smyth
3. An introduction to Building the Data Warehouse – IBM
4. Data Mining Practical Machine Learning Tools & Techniques By Ian H. Witten & Eibe Frank
5. Piatetsky-Shapiro, Gregory; Parker, Gary (2011). "Lesson: Data Mining, and Knowledge Discovery: An Introduction". Introduction to Data Mining. KD Nuggets. Retrieved 30 August 2012
6. Zhu, Xingquan; Davidson, Ian (2007). Knowledge Discovery and Data Mining: Challenges and Realities. New York, NY: Hershey. pp. 163–189.ISBN 978-1-59904-252-7.
7. Seltzer, William. The Promise and Pitfalls of Data Mining: Ethical Issues
8. O'Brien, J. A., & Marakas, G. M. (2011). Management Information Systems. New York, NY: McGraw-Hill/Irwin.
9. IEEE Transactions on Evolutionary Computation14 (15): 671–687. doi:10.1109/TEVC.2010.2058118
10. MicroStrategy, Incorporated (1995). "The Case for Relational OLAP" (PDF). Retrieved 2008-03-20
11. Nigel Pendse (2006-06-27). "OLAP architectures". OLAP Report. Retrieved 2008-03-17

Autors

1. Dr. Maulik N. Pandya

He is Head of Department and Assistant Professor in M.Sc. IT Department at Shri. A. N. Patel P. G. Institute, Anand. He completed his Ph.D. in mobile communication technology. He has registered a patent for innovative technology of money transaction over small computing devices. His research interest includes JAVA Technologies, Database Management.

2. Bhavin B. Patel

He is an Assistant Professor in M.Sc. IT Department at Shri. A. N. Patel P. G. Institute, Anand and received his M.C.A. (2005) degree from Jalgaon University, Maharashtra. His research interest includes RDBMS.