Intelligent Multidimensional Modelling

Swati Hira
Computer Science and Engineering
VNIT, Nagpur, India
swati.heera@gmail.com

Dr. Parag S. Deshpande
Computer Science and Engineering
VNIT, Nagpur, India
psdeshpande@cse.vnit.ac.in

Abstract- On-Line Analytical Processing (OLAP) systems considerably ease the process of analyzing business data and have become widely used in industry. Such systems primarily employ multidimensional data models to structure their data. However, current multidimensional data models fall short in time and skills to model the complex data found in some real-world application domains. Multidimensional data Analysis is based on Measure, Dimensions and Hierarchies. Process to find them manually is very crucial and time consuming because large and complex data is involved across multiple regions, products, and employees. This paper presents an Intelligent Multidimensional modelling system which helps the modeller in building multidimensional model and provides working at logical level by hiding heterogeneousity of physical database. The paper proposes the process to identify Measures, Dimensions, and Hierarchies to generate multidimensional model.

Keywords: Multidimensional Data Modelling, Measures, Dimensions, Hierarchies.

I. INTRODUCTION

Multidimensional data is a type of data that records facts related to various entities, called dimensions. It analyzes and categorizes data based on multiple dimensions and measures. For example, when a customer purchases an item from a retailer, this information may be recorded as transaction in following format:

Customer X has purchased item Y from retailer Z on date W of amount A.

A. Usage

Multidimensional data analysis helps in the mathematical modelling of business process; especially where large and complex data is involved across multiple regions, products, and employees. The analysis generated from these models can help in decision-making and planning activities across the gamut of business operations.

B.Process

Multidimensional modelling of data contains various tasks, which include:

(a) The identification of relevant information at the source side in the form of files (html, excel, txt, csv) or RDBMS; (b) The extraction of this information; (c) the customization and integration of the information coming from multiple sources into a common format; (d) the cleaning of the resulting data set, on the basis of database and business rules, and (e) the propagation of the data to the data warehouse and/or data marts.

II. NEED OF MULTIDIMENSIONAL MODELING

Today's business scenario demands collating data from various sources to arrive at a precise and accurate conclusion for various activities through data analysis. Analysis to provide insights across the gamut of stakeholders, channels, and geographies within an organization; therefore requires multidimensional data analysis.

The varied nature of datasets makes it extremely difficult to integrate data, primarily due to the inherent complexities, as well as its inter-relationship with data from stakeholders within the organization. For example, data for a specific purpose may be collated from different sources, such as branding teams, market research houses, and channel partners and within the organization itself. Therefore it is challenge to ensure the easy integration and analysis of transaction data from varied sources, such as touch point data, customer data, and data from relatively newer technologies. This is where we need to employ sophisticated multidimensional data analysis process and technologies.

It is difficult for clients to do whole process. They require an expertise to do multidimensional modelling and suggest various possible models. Process to identify models is:

Identify measures: A multidimensional dataset contains a set of specified values associated with a transaction; the quantities are called measures or variables of a given dataset. For example, the sale within Mumbai, for a particular item X and on a 21st of June 2009 may be Rs. 6000/-. In this case sale value indicates the measure.

Identify dimensions: Selecting proper dimensions in data analysis is indeed very crucial for multidimensional analysis and gaining greater insights from the data. Dimensional modelling is very important in data analysis because business queries can be satisfied only if dimensions are defined properly and adequately. For example, when a medical representative visits a doctor in some city, the dimensions on which data may be analyzed are Region (Doctor city), Specialty, Medicine Prescribed, and Time. Dimensional modelling; therefore, involves determining the dimensions, such as Product, Region, and Time.

Identify dimensions hierarchies: Data hierarchy is defined as data that comprises both sets and subsets; and

DOI: 10.5176/2010-3043_2.3.191

where the latter ranks lower than the set itself. Hierarchies are present in the dimensions that represents grouping of data .It can be formed after dimensions are decided; the same data can be represented as separate hierarchies or dimensions depending on the analysis required. For example regional data has hierarchical characteristic, such as Country-State-District-Town.

Identify dimensions properties: A specific dataset (containing data from various sources) can encompasses numerous dimensions. However, it must be ensured that relevant dimensions are selected while analyzing data for specific purposes.

Properties: Data Analysis is based on hierarchy, sequential property and dependency relationships.

Hierarchy property: Hierarchies are present in the dimensions that represent grouping of data. The same data can be represented as separate hierarchies or separate dimensions depending upon the analysis required. The analysis changes according to varied data dimensions. Suppose hierarchy is

Country-State-Region-City

Then the information like top three states having largest cultivation of rice can be retrieved.

Sequential property: It refers to the presence of a definite predecessor and successor value in a dataset. A dataset may be characterized by the sequential property when it has a definite predecessor and successor value. For example, sale in the months of January, February, and March may be recorded in following format and growth of sale is calculated for each month.

Table 1

Sales in the Month of January, February, and March 2009			
Month	***January 2009***	***February 2009***	***March 2009***
Sale value	32	47	54

Dependency Relationship in Data: In a dataset, some variables may be dependent on others in such a way that their relationship or dependency is fixed. For example, Sales depends on advertisement, may be specified with the following expressions:

Sales= k*advertisement amount

III. MODEL

Various models can be designed for any given data. Each model can address specific business analysis task. Modelling is very crucial in multidimensional analysis. It addresses the following issues:

1) Dimensions should be built to address a given address problem; 2) Hierarchies for which the analysis is required; 3) Measures to work with the given dimensions and data hierarchies; 4) Behaviour of variables at higher levels of data.

Example of Multidimensional Data Analysis

Sales Force Management

Axis Pharmaceutical is supplying 69 drugs in various categories. It has distributers located in various cities. The grouping of cities is done on the basis of the area parameter. For promotion of the products, sales representatives are appointed, who visits doctors on regular Basis. The visits are recorded in the following format:

Table2

Visits						
Date	***Representative Name***	***City***	***Doctor Name***	***Doctor Age***	***Doctor Specialization***	***Product Name***

Table3

Sales Information from Invoices			
Month	***Product***	***City***	***Sale Amount***

Using these data two models can be built.

Model I: This model is used for analyzing product calls and sales. The hierarchies and dimensions for this model are as follows:

Dimensions and Hierarchies

- Region-All-State-Area-City
- Product-All-Items
- Time-All-Year-Quarter-Month

Variables (Measures)

- Amount
- Product calls(number of visits)

Model II: In this model, greater importance has been given to analyze doctor's data. The analysis helps in specific targeting of doctors. The hierarchies and dimensions for this model are as follows:

Dimensions and Hierarchies

- Region-All-Area-City
- Doctor-All-Specialization-Age Group
- Product-All-Items
- Time-All-Year-Quarter-Month

Variables (Measures)

- Amount
- Product calls(number of visits)

IV. PROPOSED MODEL

Intelligent Multidimensional modelling will automatically map physical data into logical model. As in above Sales Force Management data example user has to select dimensions, hierarchies, measures and facts. This is

very difficult, time consuming and an expert is required to do all those process.

Aim of intelligent multidimensional modelling is to provide a view that can be used by BI applications.

Automatic Multidimensional modelling process:

1) Collect data tables which in different formats and transform those formats into a common format.
2) Find dimensions for all tables in given data field.
3) Group the dimensions to reduce number of tables or number of columns.
4) Find hierarchy after analyzing the dimensions for tables to provide solutions for all BI queries at every level.
5) Find measures for all tables in given data field.
6) Models will be generated on the basis of dimensions, hierarchy, measures and represent different view to users.

Therefore users can analyze and query data models to solve business problems.

In this modelling process users have to just enter the source of data and they will get all possible models for given data set.

Rules to generate model

(A) Identifying Measures

(a) Identify data type of column, which should be numeric.(b) Values should not have patterns, example (1.1, 2.1, 3.1, 4.1...10.1); (c) Values should not be discrete.

(B) Identifying Time:

(a)First few rows or columns of given table will be scanned; (b) If year is find in any one of the row or column example (1950-1951 to 2009-2010), a new column would be added to transform column as year; (c) Date components like month, day will be identified.

(C) Identifying Dimensions and Hierarchies: A relational database used for OLTP system is a set of tables. Each table is having table name, table comments, column name, column data type, constraints and rows as information. This information is scanned to identify dimensions and hierarchies.

Dimensions and hierarchies are identified by using the following heuristics:

(a) Remove all columns having names which describe general information like remarks, comments.

(b) Remove all columns having large data types.

(c) Remove all words in table heading names which represents general words and construct new columns for each remaining words. For each table construct extended table definition with new columns.

(d) For each extended column, calculate row count. Calculate global row count for each extended column. Identified extended columns as dimensions column if its global row count is above threshold.

(e) Determine whether two dimension columns are having one to many relationships. If such relationship exists construct hierarchical relationship between two dimension columns.

(f) Prepare view in the following form

D_{11} D_{12} D_{13} D_{21} D_{22} .....Measure value

Where D_{12} is dimension column in the lower hierarchy than D_{11}.

(g) Construct queries to map physical data into logical data.

Example: The above model was developed for "Crop data of Indian economy" and logical view is constructed as follows:

*(A)Data Source (*Step1)

Table 4.1

State-wise Area for Kharif Pulses (Other than Tur) in India		
(In ' 000 Hectare)		
States/UTs	***Pulses***	***2010-11****
Arunachal Pradesh	Urad	350

Table 4.2

Area under Crops in India - Part I							
(1950-1951 to 2009-2010)							
(' 000 Hectare)							
Year	***Rice***	***Jowar***	***Bajra***	***Maize***	***Ragi***	***Wheat***	***Barley***
1950-51	3500	1500	9000	3000	2000	1000	300

Table 4.3

State-wise Area Under Crops (Food Crops) in India - Part XII							
(1992-93 to 1994-95)							
(Thousand Hectare)							
Food Crops							
	Fresh Fruits					***Dry Fruits***	
States/UTs/Year	***Mango***	***Citrus fruit***	***Banana***	***Grape***	***Papaya***	***Others***	***Almond***
Tripura							
1992-93	5	13	-	4	1	-	9

Table 4.4

State-wise Area under Crops in India - Part III							
(1950-1951 to 2009-2010)							
(' 000 Hectare)							
Condiments / Spices							
States/UTs	***Blac***	***Chilli***	***Ging***	***Turme***	***Cardam***	***Bet***	***Othe***

/Year	*k Pepper*	*es*	*er*	*ric*	*om*	*el nuts*	*rs*
Andhra Pradesh							
1950-51	50	60	240	250	100	140	160

(B)*Intermediate Data* (Step2)

Table 4.1.1

Country	**State**	**Year**	**Crop**	**Crop_ category-I**	**Area(Hectare)**
India	Arunachal Pradesh	2011	Kharif pulses	Urad	350

Table 4.2.1

Country	**Year**	**Crop**	**Area(Hectare)**
India	1951	Rice	3500
India	1951	Jowar	1500

(C)Generated Logical View

Table 4.3.1

Country	**State**	**Year**	**Crop**	**Crop_ category-I**	**Crop_ category-II**	**Area(Hectare)**
India	Tripura	1993	Food crops	Fresh fruits	Mango	5000
India	Tripura	1993	Food crops	Dry fruits	Almond	9000
India	Andhra Pradesh	1953	Spices	Black pepper		500

Model:

The following hierarchies and dimensions are used for this model:

Dimensions and Hierarchies:

- Region-All-Country-state
- Product-All-Crop- Crop_ Category-I- Crop- Crop Category-II- Crop Name
- Scheme-All-Scheme
- Time-All-Year-Month

Variables (Measures):

- Area of Crop in Hectares

Advantage

The following business queries may be addressed:

1. In which regions are certain crop patterns more evident?
2. Which States have more acreage of wheat?
3. What has been the effect of rain on crops across cultivable areas?
4. What are the reasons for change in crop patterns in certain areas?
5. What income can be generated from crop cultivation?
6. What is the harvesting position of crops under last year?
7. In which locations should training camps be conducted for pest control?
8. If a new variety is introduced in a particular geographical area, what has been its effect on the crop pattern?

Once robust model is built then information can be easily retrieved.

V. CONCLUSION

Dimensions modelling for business Intelligence is very time consuming and requires specialized skills. The proposed model helps the modeller in building multidimensional model and provides working at logical level by hiding heterogeneousity of physical databases.

REFFERENCES

[1] Dr. P. S. Deshpande, Arijay Chaudhry,Multidimensional Data Analysis and Data Mining.
[2] www.indiastat.com
[3] Jaiwe Han, Micheline Kamber, Jian Pei,"Data Mining Concepts and Techniques".
[4] Torben Bach Pedersen, Christian S. Jensen,"Multidimensional Data Modeling for ComplexData", Denmark.
[5] Anirban Sarkar, Chaki," Implementation of Graph Semantic Based Multidimensional Data Model",Journal of Computer Information Systems, 2011.
[6] Chaogui Zhang, Zhiyong Zheng," Multidimensional Traffic GPS Data Quality Analysis Using Data Cube Model".
[7] Bhaskar Reddy Moole," Forecasting Demand Using Probabilistic Multidimensional Data Model", Walden University

ABOUT THE AUTHORS:

1) Dr.P.S.Deshpande is currently Associate Professor at National Institute of Technology, Nagpur. He did his M.Tech. in Computer Science from IIT, Bombay and Ph.D. from National Institute of Technology, Nagpur. He has published several research papers in Data mining,Pattern recognition and Machine Intelligence in International Journals and Conferences. He has written 9 books on Data warehousing, Data Mining and Data analysis.

Dr.Parag Deshpande
psdeshpande@cse.vnit.ac.in
psdeshpandevnit@gmail.com

2) Swati Hira is currently Research Scholar at National Institute of Technology, Nagpur.She is pursuing Ph.D. under Dr.P.S.Deshpande. She did his M.Tech. in Computer Science from DAVV, Indore.

Swati Hira
swatihira@cse.vnit.ac.in
swati.heera@gmail.com