Discovering the Association Rules in OLAP Data Cube with Daily Downloads of Folklore Materials[*]

Galina Bogdanova, Tsvetanka Georgieva

Abstract: *Association rules mining is one kind of data mining techniques which finds interesting relationships among attributes in analyzing data. This paper presents an application that discovers the association rules by using data cube structure and applying OLAP operations. It allows to be performed the association analysis of the data in the cube created by using a WEB based client/server system that contains an archival fund with folklore materials.*

Keywords: *Data Warehouse, Data Cube, OLAP (Online Analytical Processing), Data Mining, Association Rules, MDX (Multidimensional Expressions).*

INTRODUCTION

The data warehouse has enormous value to the organization by arranging operational data into meaningful information. Often, data warehouses are designed for online analytical processing (OLAP), where the queries aggregate large volumes of data. OLAP operates efficiently with data organized in accordance with the common dimensional model used in data warehouses. OLAP organizes data warehouse data into multidimensional cubes based on this dimensional model, and then preprocesses these cubes to provide maximum performance for queries that summarize data in various ways. However, much of the information required for proactive activities of an organization cannot be accommodated simply through organized views of historical data. Data mining allows empirically navigate the organization to profitability, while simultaneously setting the focus of the organization and adding insight into its processes, objects and attributes [6]. Data mining process is efficient only with presence of summarized data that is stored in data warehouses.

The components, characteristics and architecture of the data warehouse and the means of OLAP are exhibited in [13]. A method of efficiently maintaining aggregate views is proposed in [12]. Some tasks of data mining are discussed in [14].

Data mining aims at the discovery of useful summaries of data. Association rule mining is a form of data mining to discover interesting correlation relationships among data. The discovered rules may help decision making in different area. The problem of association rules mining originates with the problem of market analysis on sales basket data and it was first introduced in [1]. In [16] is proposed and developed a method which integrates OLAP technology with association rules mining methods. An algorithm for discovering distribution intervals of association rules in time by computing the fractal dimension and exploring a data cube structure is represented in [7]. This algorithm is applied to the data cube created by using a WEB based client/server system that contains an archival fund with folklore materials of the Folklore Institute at BAS. In [4] are reported the advantages of the applying the OLAP operations to analyze the data in mentioned data cube.

In present paper is represented an application that discovers the association rules in data cube with daily downloads of folklore materials, whose creation is described in detail in [5]. The rest of the paper is organized as follows. Section 2 reviews the concepts of the association rules and OLAP-based association rule mining. In section 3 is described an application for discovering the association rules in data cube by using the OLAP operations. Section 4 presents the realized ways for visualization of association rules and

[*] Supported partially by the Bulgarian National Science Fund under Grant MM-1405/2004

advantages of different presentations of association rules. Section 5 gives the conclusion of this paper.

ASSOCIATION RULES DISCOVERY AND OLAP TECHNOLOGY

Association rules mining is to find interesting associations or correlation relationships among a large set of data, i.e. to identify sets of attribute-values (predicate or item) that frequently occur together, and then formulate rules that characterize these relationships [1].

An association rule is an implication of the form $X \rightarrow Y$, where $X = \{x_1, \ldots, x_m\}$ and $Y = \{y_1, \ldots, y_n\}$ are sets of items with $X \cap Y = \varnothing$. The rule $X \rightarrow Y$ has support s if s% of all itemsets contain $X \cup Y$. The rule $X \rightarrow Y$ has confidence c if c% of itemsets that contain X also contain Y. The problem of mining association rules is to generate all association rules that have support and confidence greater than the user-specified minimum support *min_supp* and minimum confidence *min_conf* respectively [2]. The problem of discovering association rules can be decomposed into two subproblems:

1) Find the set F of all itemsets with support above minimum support *min_supp*. These itemsets are called *frequent itemsets*.
2) Use the frequent itemsets to generate the desired rules. For every $X \in F$ check the confidence of all rules $X \setminus Y \rightarrow Y$, $Y \subset X$, $Y \neq \varnothing$ and eliminate those that do not achieve *min_conf*.

It is sufficient to calculate all support values of the subsets of X to determine the confidence of each rule.

In [2] is proposed algorithm *apriori*: first, the set of frequent 1-itemsets L_1 is found; then the set of candidate k-itemsets C_k with $k \geq 2$ is generated by joining frequent (k – 1)-itemsets L_{k-1} and eliminating those having a (k – 1)-subset that is not frequent; on the next step of algorithm the set of frequent k-itemsets L_k is generated from C_k by comparing the support of each candidate in C_k with *min_supp*. During each iteration k increases ($k = k + 1$) and this process continues while $L_k = \varnothing$ for some k.

OLAP-based association rules mining integrates OLAP technology and association rules mining that facilitates flexible mining of interesting knowledge in data cubes because data mining can be performed at multidimensional and multilevel abstraction space in a data cube [8, 10]. In [10] are proposed efficient algorithms by either using an existing data cube or constructing of a data cube. The common idea is to initially find the frequent 1-itemsets in each dimension, and then use frequent (k – 1)-itemsets to obtain frequent k-itemsets by applying appropriate OLAP operations for selection of multidimensional parts from data cube. The performance analysis shows that this method outperforms apriori algorithm which uses a relational table-based structure and requires multiple scans of the data.

The multidimensional data structure facilitates efficient mining of multilevel association rules. The aggregate values needed to discovering the association rules are computed and stored in data cube which facilitates the association testing and filtering. A count cell of a cube stores the number of occurrences of the corresponding multidimensional data values. A dimension count cell stores the sum of counts of the whole dimension. With this structure, it is straightforward to calculate the values of the support and confidence of association rules based on the values in these summary cells.

DISCOVERING THE ASSOCIATION RULES IN DATA CUBE *FOLKLORECUBE*

The data cube *FolkloreCube* is created by using a WEB based client/server system that contains an archival fund with folklore materials of the Folklore Institute at BAS. This system uses OLTP (online transaction processing) database that is created in accordance to the classification schema described in [11]. It keeps detailed information of the documents and materials, which can be downloaded by the users and contain audio, video and text information. The database in data warehouse is designed and the data is extracted from the OLTP database, transformed to match the data warehouse schema, and loaded into data warehouse database periodically by execution a batch job.

The data cube *FolkloreCube* consists of four dimensions – *Document*, *Link*, *User* and *Time*. The lattices for the dimensions hierarchies of examined data cube are shown in figure 1. The measure of the examined data cube is count of downloads of the folklore materials from the documents by the users. A data cell in the cube *FolkloreCube* *c*[*Document* = *d*, *Link* = *l*, *User* = *u*, *Time* = *t*] stores the *count* of the corresponding rows of the initial relation; a summarizing cell in the cube *c*[*Document* = “*all*”, *Link* = *l*, *User* = *u*, *Time* = *t*] stores the *sum* of the *counts* of the whole dimension *Document*, i.e. without grouping by that dimension and so on; the value of the cell *c*[*Document* = “*all*”, *Link* = “*all*”, *User* = “*all*”, *Time* = “*all*”] is summation of the counts obtained without grouping by any dimension.

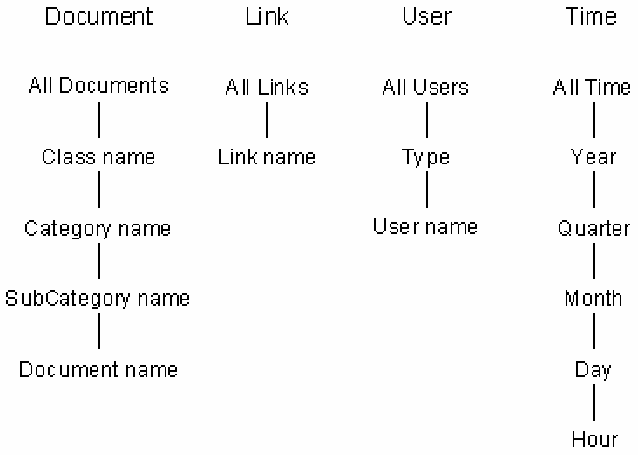


Fig. 1 Dimension hierarchy lattices defined in *FolkloreCube*

For example, based on the data shown in Table 1 we obtain the following values stored in the cube: the value stored in the count cell *c*[*Document* = “*Кръщене*”, *Link* = “*songlink062*”, *User* = “*User0001*”, *Time* = “*19/02/2005*”] is 1; *c*[*Document* = “*Имен ден*”, *Link* = “*songlink031*”, *User* = “*all*”, *Time* = “*11/2004*”] is 2; *c*[*Document* = “*all*”, *Link* = “*all*”, *User* = “*User0002*”, *Time* = “*2004*”] is 4 and so on; *c*[*Document* = “*all*”, *Link* = “*all*”, *User* = “*all*”, *Time* = “*all*”] is 10.

Table 1 Example data of the materials downloaded from the users

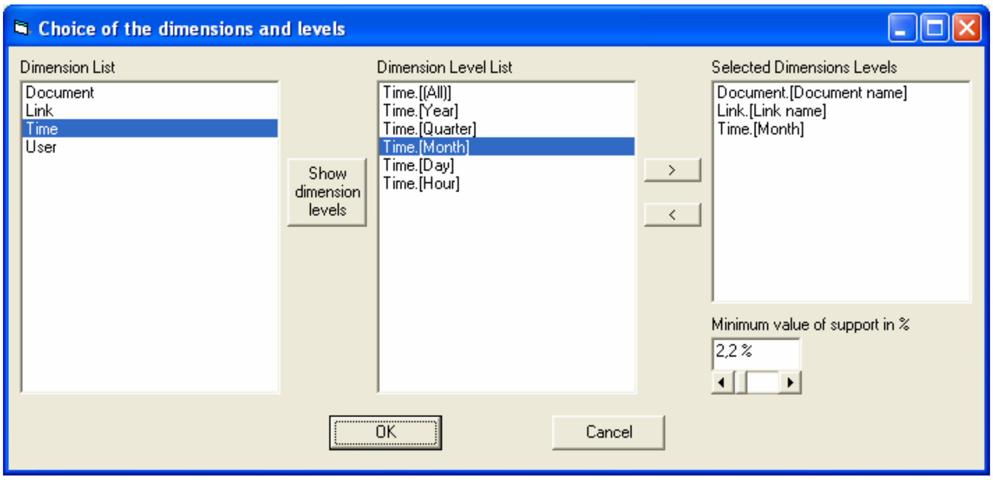
Document	**Link**	**User**	**Time**
Кръщене	songlink062	User0001	19/02/2005
Годеж	songlink013	User0001	09/01/2005
Имен ден	songlink031	User0001	27/11/2004
Имен ден	songlink032	User0001	27/11/2004
Имен ден	songlink031	User0002	20/11/2004
Кръщене	songlink061	User0002	20/11/2004
Семейство	songlink071	User0002	20/11/2004
Годеж	songlink013	User0002	29/12/2004
Годеж	songlink012	User0002	01/03/2005
Годеж	textlink011	User0002	28/05/2005

The represented application discovers the association rules in data cube *FolkloreCube* by using the OLAP operations. It is realized by using the languages MDX (*Multidimensional Expressions*) [17, 18, 19] and Visual Basic [3, 15].

The user that starts the application has possibility to (fig. 2):

- Choose the dimensions and the levels of the dimensions to be analyzed.

Usually the user is interested in specified subset of attributes and wants to extract interesting relationships among chosen attributes. Therefore a facility with a friendly interface should be provided to specify the set of attributes to be mined and exclude the set of irrelevant attribute from examination. User-controlled generalization of attributes increases or decreases the levels of abstraction for attributes.



Choice of the dimensions and levels

Dimension List
Document
Link
Time
User

Dimension Level List
Time.[(All)]
Time.[Year]
Time.[Quarter]
Time.[Month]
Time.[Day]
Time.[Hour]

Selected Dimensions Levels
Document.[Document name]
Link.[Link name]
Time.[Month]

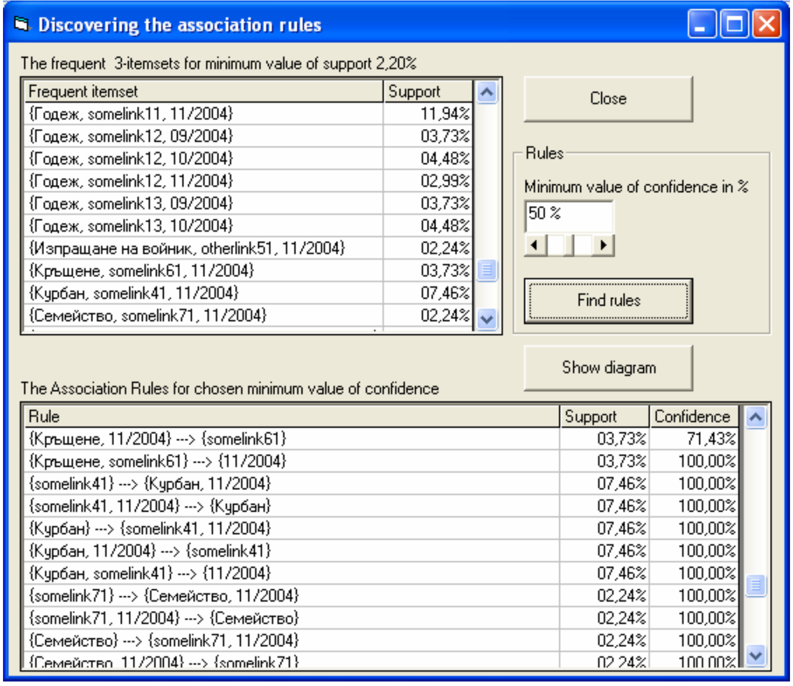
Minimum value of support in %: 2,2 %

Fig. 2 Selection of dimensions, levels and minimum support

- Set the minimum value of the support *min_supp*.

The support reflects the usefulness of a rule. The minimum support *min_supp* that an associations rule must satisfy means that each value to be examined must to be occurred a significant number of times in the corresponding attribute of the initial relation. The user can set different values of minimum support when mining items at the different levels of abstraction to generate sufficient meaningful association rules at low levels and sufficient interesting association rules at high levels.

It is easy to group data according one or a set of dimensions using the cube structure. The application finds the set of frequent k-itemsets L_k with selected dimensions and levels in data cube *FolkloreCube* and chosen minimum support by using MDX query. Then it computes the confidences of relevant rules by composing and executing MDX queries. These values are compared with minimum confidence that is user-defined (fig. 3).



Discovering the association rules

The frequent 3-itemsets for minimum value of support 2,20%

Frequent itemset	Support
{Годеж, somelink11, 11/2004}	11,94%
{Годеж, somelink12, 09/2004}	03,73%
{Годеж, somelink12, 10/2004}	04,48%
{Годеж, somelink12, 11/2004}	02,99%
{Годеж, somelink13, 09/2004}	03,73%
{Годеж, somelink13, 10/2004}	04,48%
{Изпращане на войник, otherlink51, 11/2004}	02,24%
{Кръщене, somelink61, 11/2004}	03,73%
{Курбан, somelink41, 11/2004}	07,46%
{Семейство, somelink71, 11/2004}	02,24%

Rules

Minimum value of confidence in %: 50 %

The Association Rules for chosen minimum value of confidence

Rule	Support	Confidence
{Кръщене, 11/2004} ---> {somelink61}	03,73%	71,43%
{Кръщене, somelink61} ---> {11/2004}	03,73%	100,00%
{somelink41} ---> {Курбан, 11/2004}	07,46%	100,00%
{somelink41, 11/2004} ---> {Курбан}	07,46%	100,00%
{Курбан} ---> {somelink41, 11/2004}	07,46%	100,00%
{Курбан, 11/2004} ---> {somelink41}	07,46%	100,00%
{Курбан, somelink41} ---> {11/2004}	07,46%	100,00%
{somelink71} ---> {Семейство, 11/2004}	02,24%	100,00%
{somelink71, 11/2004} ---> {Семейство}	02,24%	100,00%
{Семейство} ---> {somelink71, 11/2004}	02,24%	100,00%
{Семейство, 11/2004} ---> {somelink71}	02,24%	100,00%

Fig. 3 Results from discovering the association rules in *FolkloreCube*

The confidence reflects the certainty of a discovered rule. For example the following rule is generated from the data cube with daily downloads of folklore materials:

{*Document*(“*Кръщене*”), *Time*(“*11/2004*”)} → {*Link*(“*somelink61*”)}

with support = 3,73% and confidence = 71,43%. This rule means that one of the most downloaded materials from document “*Кръщене*” during November 2004 is the material “*somelink61*” (with 71,43% confidence) and such downloads represent 3,73% from all downloads under study.

VISUALIZATION OF ASSOCIATION RULES

Different ways for visualization of association rules allow users to work in an interactive environment with ease in analyzing the rules. Implemented possibilities for visualizing the mining results in represented application are tabular view and graphical view.

In tabular view of association rules, all discovered rules are represented in a tabular table with each row corresponding to a rule and represents information about rule support and confidence. All the rules can be displayed in different order – by the alphabet; by the support or confidence in ascending or descending order. By this way the user has a clearer and complete view of the rules and can locate a specific rule more easily. The tabular view facilitates understanding the large number of rules. This presentation of resulting association rules is shown in figure 3.

The tabular view is not very convenient when the user needs to obtain a comprehensive view of the relationships between rules and items. The graphical visualization provides a more clear and vivid view of the rules and items. Figure 4 shows a graph for visualization of some association rules. The existence of a column (bar) on the grid represents an association between the left-hand side items and the right-hand side items. The height of the column depicts the support of the rule it represents.

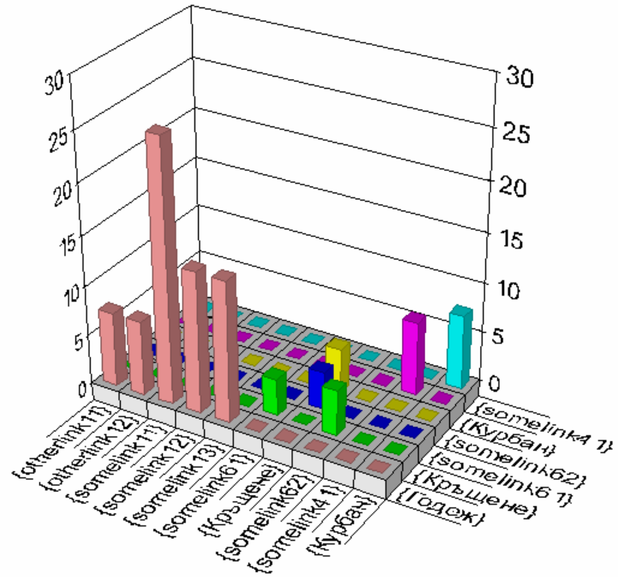


Fig. 4 Graph for visualizing association rules

CONCLUSION

Discovering the association rules is an important data mining function. Data cube structure increases flexibility and efficiency of association rules mining. Represented application provides a possibility to association analysis of daily downloads of folklore materials according to dimensions of interest. Discovered association rules can be displayed in different ways when the user of this application needs to view the rules from different aspects.

REFERENCES

[1] Agrawal, R., T. Imielinski, A. Swami. Mining Association Rules between Sets of Items in Large Databases, In Proc. of the ACM SIGMOD International Conference on Management of Data, Washington, 1993, pages 207-216.

[2] Agrawal, R., R. Srikant. Fast Algorithms for Mining Association Rules, In Proc. of International Conference on Very Large Databases, 1994, pages 487-499.

[3] Bekuit, B., VB.NET: A Beginner's Guide, AlexSoft, 2002, pages 330 (in Bulgarian).

[4] Bogdanova, G., Tsv. Georgieva. Applying the OLAP Operations to Analyzing the Data in a WEB based Client/Server System Containing Archival Fund with Folklore Materials, In Proceedings of the National Workshop on Coding Theory and Applications, Bankya, 9-12.12.2004, page 4.

[5] Bogdanova, G., Tsv. Georgieva. Analyzing the Data in OLAP Data Cubes, International Journal on Information Theory and Applications, 2005 (submitted).

[6] Charran E., Introduction to Data Mining with SQL Server (Part 2), http://www.sql-server-performance.com/ec_data_mining2.asp, 2002.

[7] Georgieva, Tsv., Using the Fractal Dimension of Sets to Discover the Distribution Intervals of Association Rules in OLAP Data Cubes, In Proceedings of the First International Conference on Information Systems and DataGrids, Sofia, 17-18.02.2005, pages 88-98.

[8] Han, J., Towards On-Line Analytical Mining in Large Databases, SIGMOD Record (ACM Special Interest Group on Management of Data), 1998, pages 97-108.

[9] Hipp, J., U. Güntzer, G. Nakhaeizadeh. Algorithms for Association Rule Mining – A General Survey and Comparison, ACM SIGKDD, 2000, pages 58-64.

[10] Kamber, M., J. Han, J. Chiang. Using Data Cubes for Metarule-Guided Mining of Multi-Dimensional Association Rules, Technical Report, CMPT–TR–97–10, School of Computing Sciences, Simon Fraser University, 1997, pages 6.

[11] Mateeva, V., I. Stanoeva. Classification Scheme of the Typological Catalogue in the Folklore Institute, Bulgarian Folklore, v. 2-3, 2001, pages 96-109 (in Bulgarian).

[12] Mumick, I., D. Quass, B. Mumick. Maintenance of Data Cubes and Summary Tables in a Warehouse, In Proc. ACM SIGMOD Conf. on Management of Data, Tuscon, Arizona, 1997, pages 100-111.

[13] Peneva, J., G. Tuparov. Databases, Regalia 6, 2004, pages 230 (in Bulgarian).

[14] Ullman, J. (http://www.db.stanford.edu/~ullman/mining/mining.html).

[15] Wang, W., Visual Basic 6: A Beginner's Guide, AlexSoft, 2002, pages 562 (in Bulgarian).

[16] Zhu, H., Online Analytical Mining of Association Rules, Master Thesis, Simon Fraser University, 1998, pages 117.

[17] http://www.georgehernandez.com/xDatabases/MD/MDX.htm

[18] http://www.microsoft.com/data/oledb/olap

[19] http://www.microsoft.com/sql

ABOUT THE AUTHORS

Assoc. Prof. Galina Bogdanova, PhD, Institute of Mathematics and Informatics, Bulgarian Academy of Sciences, Veliko Tarnovo, P.O.Box: 323, E-mail: galina@moi.math.bas.bg

Tsvetanka Georgieva, University of Veliko Tarnovo "St. St. Cyrill and Methodius", Department of Information Technologies, Phone: 0889823216; E-mail: cv.georgieva@uni-vt.bg