Mathematica
Balkanica

New Series Vol. 20, 2006, Fasc. 3-4

Algorithm for Discovering the Distribution Intervals of the Association Rules in OLAP Data Cubes[1]

Tsvetanka Georgieva

Presented by Bl. Sendov

The application of the OLAP operations and their integration in the algorithms for discovering association rules, realize an effective method of data mining. The present paper represents an algorithm for finding the intervals of distribution of the association rules in time by computing the fractal dimension. The significant change of its value indicates the beginning of a new interval. It contains a realization of this method by using the languages MDX and Visual Basic.

Key Words: Data Cube, OLAP (*Online Analytical Processing*), Association Rules, Fractal Dimension, MDX (*Multidimensional Expressions*)

1. Introduction and Motivation

Online analytical processing (OLAP) technology uses a dimensional view of summarized data and provides a quick answer of queries, performed to analyze the data. It is designed to provide superior performance for the queries processing large amounts of data, used by decision support systems (DSS). The data liable to OLAP is organized in multidimensional cubes, created according to the dimensional model used in data warehouse. They store preprocessed summaries of the data. The data cubes creation and usage eliminates the need of joining the tables and preprocessing the values returned from the most frequently performed queries. OLAP includes a set of operations for manipulation of the dimensional data organized in multiple levels of abstraction. The basic OLAP operations are roll-up (increasing the level of aggregation), drill-down (decreasing the level of aggregation), slice-and-dice (selection and projection), and pivot (re-orienting the multidimensional view of data).

[1] Supported partially by the Bulgarian National Science Fund under Grant IO-3/2006

With an enormous amount of data stored in databases and data warehouses, it is increasingly important to develop powerful data warehousing and data mining tools for analysis of this collected data and mining interesting knowledge from it. OLAP mining integrates online analytical processing with data mining so that mining algorithms can be performed in different portions of data in data warehouses and at different levels of abstraction [11, 12]. One of the basic advantages of the OLAP mining is the usage of data extracted from data warehouses. The data is loaded into data warehouse after it is previously integrated, consolidated, cleaned, and transformed. This motivates a study on mining of distribution intervals of association rules in time using the data cube structure.

Some common tasks of data mining are considered in [8, 15]. The problem of association rules mining originates with the problem of market analysis on sales basket data and it was first introduced in [1]. A method which integrates OLAP technology with association rules mining methods is proposed in [17]. In [7] an application that discovers the association rules in data cube with daily downloads of folklore materials is represented. The present paper represents an algorithm, which discovers the frequent k-itemsets and by computing the fractal dimension finds the interval of distribution of relevant association rules in time. The proposed approach for finding separate intervals explores a data cube structure. The previous study on fractal mining of distribution intervals of association rules [3] uses a relation table-based structure and requires multiple scans of the data in order to find all frequent itemsets. The fractal dimension is useful tool for the clustering of large datasets [2] and quickly selection of the most important attributes for a given dataset [14].

The algorithm proposed in the present paper 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 of BAS. The data cube creation is described in detail in [6]. An earlier and shorter version of this paper appeared as [5]. The conference paper [9] contains the results in a preliminary form.

The remainder of the paper is divided as follows. Section 2 reviews the concepts of the association rules and their distribution in time. Section 3 presents the main fractal properties. Section 4 gives an approach for finding intervals of distribution of association rules in OLAP data cubes. Section 5 presents the realization of the algorithm and some results of applying this approach. Section 6 gives the conclusion of this paper.

2. Association Rules and their Distribution in Time

Association rules mining is one kind of data mining techniques, which discovers interesting relationships among attributes of analyzing data. The discovered association rules can be used for the decision support in different area. An association rule shows the frequently occurring patterns of a set of data items in a database. Association rules are rules of the form $X \rightarrow Y$ where X and Y are sets of items with $X \cap Y = \emptyset$. For example the rule $\{Store("StoreA")\} \rightarrow \{Product("ProductA")\}$ means that the customers in "*StoreA*" will most possibly buy "*ProductA*".

There are two parameters associated with a rule - support and confidence. Let with T is denoted a set of itemsets. The rule $X \rightarrow Y$ has support s in the set T if s% of elements in T contain $X \cup Y$. The rule $X \rightarrow Y$ has confidence c if c% of elements in T that contain X also contain Y [1]. The most difficult part of the association rules discovery algorithm is to find the itemsets $X \cup Y$, that have support no less than a pre-defined minimum value *min_supp* (*threshold*). These sets are called *frequent itemsets.*

In addition to the association rules the important practical applicability has their distribution in time. For example, if the previous rule $\{Store\ ("StoreA")\} \rightarrow \{Product\ ("ProductA")\}$ has a support of 0.75, all it is known is that the out of all the purchases analyzed in dataset, 75% of them are realized in "*StoreA*" and contain "*ProductA*". The rule says nothing about the distribution of purchases that originated that support. It is not known whether all the customers bought "*ProductA*" in "*StoreA*" over mostly during a short period of time (for example as a response to a promotion), or buying of this product in the store responds to a pattern that is the same, regardless of the scale chosen (self-similar).

3. Fractals and the Fractal Dimension

A fractal dataset is known by characteristic to be self-similar. This means that the dataset has roughly the same properties for a wide variation in scale or size, i.e. parts of any size of the fractal are similar (exactly or statistically) to the whole fractal. This idea is illustrated in Fig. 1, which shows the first three steps in building the Koch curve.

The Koch curve is constructed from a line segment of unit length by removing the middle third; replacing it by the other two sides of the equilateral triangle based on the removed segment and recursively repeating this procedure for each of the resulting four smaller segments.

The fractal dimension is a characteristic of the fractal sets [10, 13] and is defined with:

$$D_q = \frac{1}{q-1} \frac{\partial \log \sum_i C_{r,i}^q}{\partial \log r},$$

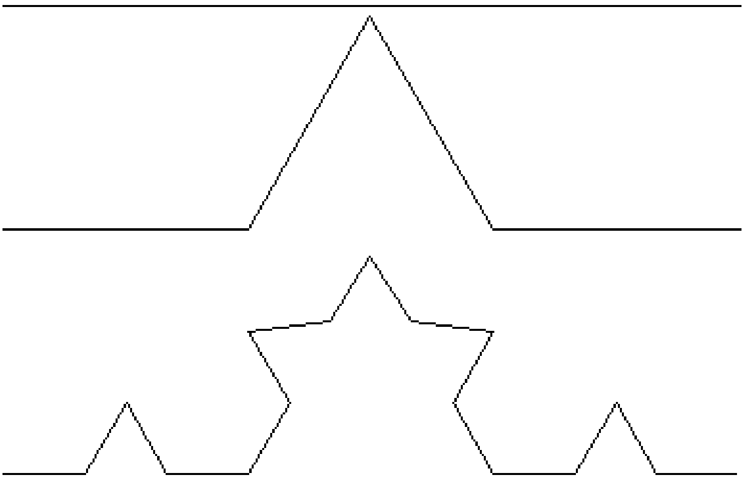


Fig. 1: The Koch curve

where the dataset is embedded in an n-dimensional grid which cells have sides of size r; $C^q_{r,i}$ is the frequency with which data points fall into the i-th cell.

With $q = 0$ comes the Hausdorff fractal dimension, which can be computed in the following way: for a set of p points, each of n dimensions, one divides the space in grid cells of size r; if $N(r)$ is the number of cells occupied by points in the dataset, it finds the polynomial from first degree $y = ax + b$, which is the best approximation of $y_i = \log(N(r_i))$, $x_i = \log(r_i)$; the Hausdorff fractal dimension D_0 corresponds to the absolute value of the coefficient a.

For example, the Hausdorff dimension of the Koch curve can be computed in the following way: the set, obtained after k iterations, consists of $N = 4^k$ pieces, each of length $r = (\frac{1}{3})^k$; by using the grid cells of size $r = (\frac{1}{3})^k$, we find 4^k of the cells populated with points; therefore, we obtain a line slope $-\log 4/\log(1/\frac{1}{3}) = -\log 4/\log 3 = -1.262$. The value 1.262 is the fractal dimension of the Koch curve.

4. Algorithm for Finding the Intervals of Distribution of the Association Rules in OLAP Data Cubes

The present approach explores association rules mining and finding the intervals of distribution in time using a data cube structure. An n-dimensional data cube $C[A_1, \ldots, A_n]$ is an n-D database where $A_1, \ldots, A_n$ are n dimensions (fig. 2). Each dimension of the cube A_i contains $|A_i| + 1$ rows where $|A_i|$ is the number of distinct values in the dimension A_i. The first $|A_i|$ rows are data

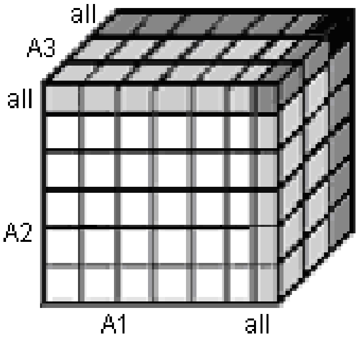


Fig. 2: 3-D data cube with dimensions A_1, A_2, A_3

rows and represent the distinct values of A_i. The last row is used to store the summation of the counts of the corresponding columns of the above rows.

The data cube can be mapped to an $(n+1)$-attribute table with each attribute representing a dimension and the $(n+1)$-th representing the *count*. A data cell in the cube $C[a_{1,i_1}, \dots, a_{n,i_n}]$ stores the *count* of the corresponding rows of the initial relation $r(A_1 = a_{1,i_1}, \dots, A_n = a_{n,i_n}, count)$. A summarizing cell in the cube $C[all, a_{2,i_2}, \dots, a_{n,i_n}]$ stores the *sum* of the *counts* of the generalized rows, which share the same values for all the second to the n-th columns, i.e. $r(all, A_2 = a_{2,i_2}, \dots, A_n = a_{n,i_n}, sum)$. A data cube can be viewed as a lattice of cuboids. The n-D space consists of all data cells, i.e. without "*all*" values; $(n-1)$-D space consists of all the cells with a single "*all*" value, i.e. $r(all, a_{2,i_2}, \dots, a_{n,i_n}, sum_1)$, ... , $r(a_{1,i_1}, a_{2,i_2}, \dots, all, sum_n)$, and so on; finally 0-$D$ space consists of one cell with "*all*" values of dimensions, i.e. $r(all, all, \dots, all, TotalCount)$.

The algorithm for finding the intervals of distribution of the association rules in OLAP data cubes can be described by the following way:

```
1) A direct mining of the association rules from $n$-$D$ cube.
Denote the frequent $k$-itemsets by $L_k$. Examine the cell count of each $k$-$D$ cell.
If a cell count satisfies $min\_supp$, then add to $L_k$.
2) For chosen element $I \in L_k$
        3) make $begin_I = interval(t_0)$
        4) make $end_I = interval(t_1)$
        5) make $M_I = \{begin_I, end_I\}$
        6) compute the fractal dimension $D_0(M_I)$
        7) for every $t_m$; $m = 2$; $m++$; $m < q - 1$ do
```

```
8) make $end_I = interval(t_m)$
9) add $end_I$ to $M_I$
10) compute the fractal dimension $D_0(M_I)$
11) if a change in $D_0(M_I)$ of more than $\varepsilon$ is obtained
        12) output found interval $begin_I$: $end_I$
        13) make $begin_I = interval(t_m)$
        14) make $end_I = interval(t_{m+1})$
        15) make $M_I = \{begin_I, end_I\}$
        16) compute the fractal dimension $D_0(M_I)$
```

In step 1 the algorithm finds L_k directly by examination of the k-D cells since the summary layers of the cube are computed. For chosen frequent k-itemset I (Line 2) it makes the lower bound of the interval equal to the first value t_0 of the time dimension in the data cube (Line 3) and the upper bound - the second value t_1 (Line 4); initializes the multidimensional set M_I (Line 5); computes the fractal dimension $D_0(M_I)$ (Line 6). With q is denoted the number of distinct values in the time dimension. Then the algorithm iterates until no more values t_m exist, generating the next set M_I (Line 9), computing the fractal dimension of the set M_I, and if a significant change is found (Line 11), the interval is output (Line 12) and a new interval is declared.

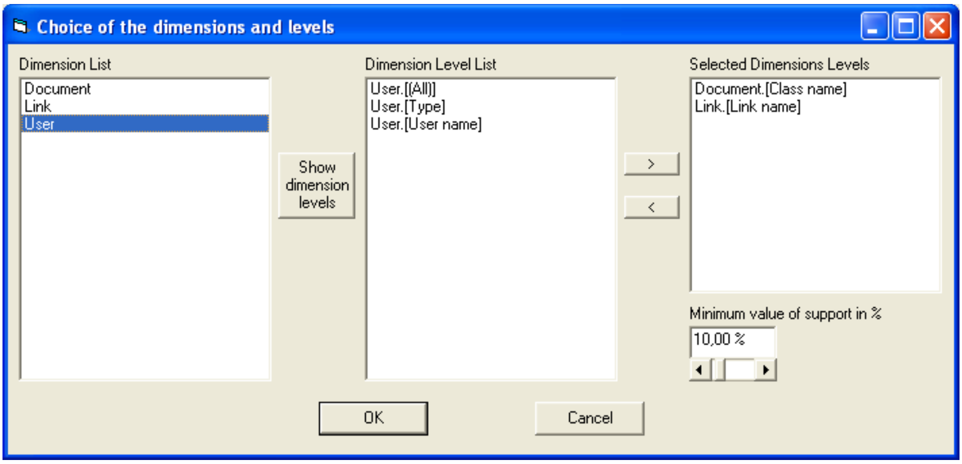
5. Realization of the Approach to Finding the Intervals of Distribution of the Association Rules in OLAP Data Cubes

The represented application is realized by using the language for OLAP queries - MDX (*Multidimensional Expressions*) [18, 19, 20] and Visual Basic [4, 16]. 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 and chosen minimum support by using MDX query. Let with $v_1, \ldots, v_k$ are denoted the numbers of distinct values in the dimensions $A_{\alpha_1}, \ldots, A_{\alpha_k}$ at chosen levels. Then in step 1 the algorithm must examine at most $v_1 \ldots v_k$ to the number of cells in order to find L_k.

Let with M_I is denoted the following subset with cells of the data cube

$$M_I = \Big\{ C[a_{\alpha_1}, \ldots, a_{\alpha_k}, t_m, all, \ldots, all],\ \textit{for each } t_m \in [t_{begin}, t_{end}]\ \textit{ and such that stores the sum of the counts } sum_m\ \textit{ satisfying the inequality } \frac{sum_m}{TotalCount} > min\_supp \Big\},$$

where $TotalCount$ is the value of the cell $C[all, \ldots, all]$, i.e. the summation of the counts obtained without grouping by any dimension; t_{begin} and t_{end} are between the first t_f and last t_l value of chosen level of the dimension of dates and hours Time; $I = \{a_{\alpha_1}, \ldots, a_{\alpha_k}\} \in L_k$ is the selected frequent k-itemset.



Choice of the dimensions and levels

Dimension List
Document
Link
User

Dimension Level List
User.[(All)]
User.[Type]
User.[User name]

Selected Dimensions Levels
Document.[Class name]
Link.[Link name]

Minimum value of support in %
10,00 %

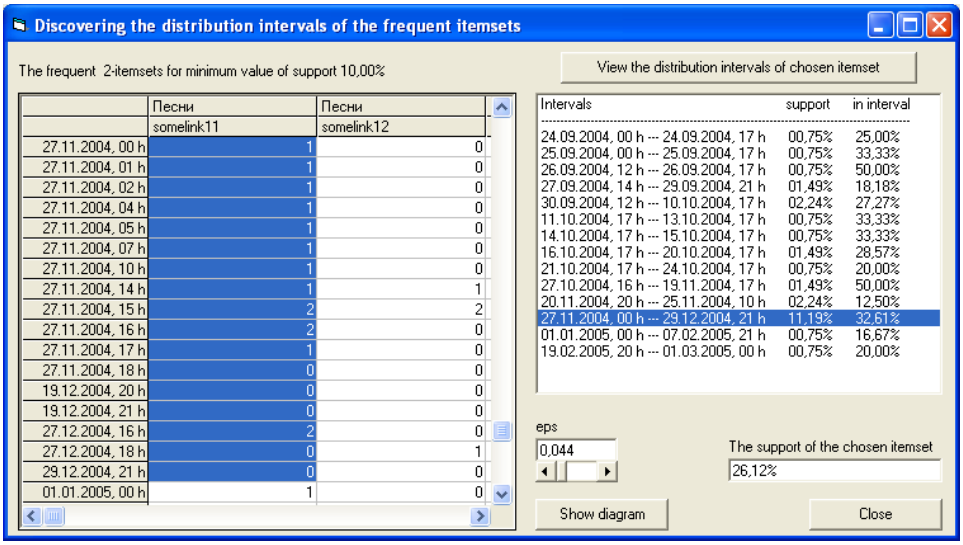
Fig. 3: Selection of dimensions, levels and minimum support

For given element $I \in L_k$ the application incrementally computes the fractal dimension of the set M_I that sequentially includes each of the values t_{h_i} of the lowest level - Hour of the dimension Time of examined data cube by composing and performing the MDX query. Consequently the algorithm is linear on the number of distinct values in the time dimension q.

It is created an user-defined function Logar that is external for MDX and is used to find the polynomial from first degree which is the best approximation by the method of the least squares of the values $x_1, \ldots, x_i$, and the values $y_1, \ldots, y_i$. It is called from the MDX query computing the fractal dimension of relevant set. The tracing of the change of its values with incremental addition of the points of the set is realized by the procedure created by using the language Visual Basic. If a significant change of the fractal dimension of the set is found, the beginning of a new interval is declared.

The present approach is applied to the data cube FolkloreCube created by using a WEB based client/server system that contains an archival fund with folklore materials of the Folklore Institute of BAS. The investigated archive keeps detailed information of the documents and materials, which can be downloaded by the users and contain audio, video and text information. The data cube FolkloreCube consists of four dimensions - Document, Link, User and Time. The measure of the examined data cube is count of downloads of the folklore materials from the documents by the users.

In the conference paper [9] the case of discovering 2-frequent itemsets is implemented. The application is developed more precise and in the present realization the user has possibility to determine how many and to choose which


Discovering the distribution intervals of the frequent itemsets
View the distribution intervals of chosen itemset
Show diagram
Close

The frequent 2-itemsets for minimum value of support 10,00%

	Песни	Песни
	somelink11	somelink12
27.11.2004, 00 h	1	0
27.11.2004, 01 h	1	0
27.11.2004, 02 h	1	0
27.11.2004, 04 h	1	0
27.11.2004, 05 h	1	0
27.11.2004, 07 h	1	0
27.11.2004, 10 h	1	0
27.11.2004, 14 h	1	1
27.11.2004, 15 h	2	2
27.11.2004, 16 h	2	0
27.11.2004, 17 h	1	0
27.11.2004, 18 h	0	0
19.12.2004, 20 h	0	0
19.12.2004, 21 h	0	0
27.12.2004, 16 h	2	0
27.12.2004, 18 h	0	1
29.12.2004, 21 h	0	0
01.01.2005, 00 h	1	0

```
Intervals                                  support   in interval
-----------------------------------------------------------------
24.09.2004, 00 h --- 24.09.2004, 17 h     00,75%    25,00%
25.09.2004, 00 h --- 25.09.2004, 17 h     00,75%    33,33%
26.09.2004, 12 h --- 26.09.2004, 17 h     00,75%    50,00%
27.09.2004, 14 h --- 29.09.2004, 21 h     01,49%    18,18%
30.09.2004, 12 h --- 10.10.2004, 17 h     02,24%    27,27%
11.10.2004, 17 h --- 13.10.2004, 17 h     00,75%    33,33%
14.10.2004, 17 h --- 15.10.2004, 17 h     00,75%    33,33%
16.10.2004, 17 h --- 20.10.2004, 17 h     01,49%    28,57%
21.10.2004, 17 h --- 24.10.2004, 17 h     00,75%    20,00%
27.10.2004, 16 h --- 19.11.2004, 17 h     01,49%    50,00%
20.11.2004, 20 h --- 25.11.2004, 10 h     02,24%    12,50%
27.11.2004, 00 h --- 29.12.2004, 21 h     11,19%    32,61%
01.01.2005, 00 h --- 07.02.2005, 21 h     00,75%    16,67%
19.02.2005, 20 h --- 01.03.2005, 00 h     00,75%    20,00%
```

eps: 0,044

The support of the chosen itemset: 26,12%

Fig. 4: Results of discovering the distribution intervals of selected frequent k-itemset with chosen $min\_supp = 0.1$ for the data from FolkloreCube

dimensions and levels of the dimensions to be analyzed as well as to set the minimum value of the support min_supp (fig. 3).

The represent application discovers the intervals of distribution of found frequent k-itemsets in data cube FolkloreCube for selected dimensions, levels and the minimum value of the support. The value of ε used by comparison with the change in the fractal dimension of the set M_I also is user-defined. It depends on how precise the resulting intervals need to be determined. Some results are shown in Fig. 4. The visualization of the results includes information about the support of selected frequent k-itemset and its support within each found interval. The support of a given itemset $\{a_{\alpha_1}, \ldots, a_{\alpha_k}\}$ within an interval is measured as the number of itemsets containing $\{a_{\alpha_1}, \ldots, a_{\alpha_k}\}$ in the interval divided by the number of all itemsets in the interval.

For instance it is interesting to find out if the common occurrence of a frequent itemset such as $\{\mathsf{Document}(\text{''}Pesni\text{''}), \mathsf{Link}(\text{''}somelink11\text{''})\}$ is clustered in some region of time or if it exhibits a self-similar behavior, i.e. occurs with the same distribution regardless of the used scale. The application provides the visualization of the behavior of the association rules in the time dimension (fig. 5).

It is created an install program. To successful starting of the program of present application is necessary to save the file of user-defined library *MyLibLog.dll* containing the user-defined function Logar that is external for MDX

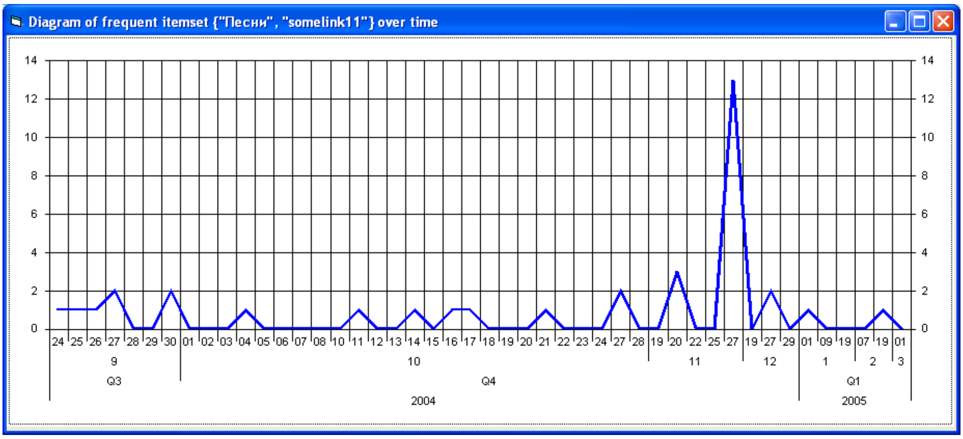


Fig. 5: *Graph for visualizing chosen frequent itemset over dates*

and the file of the local data cube *FolkloreCube.cub* in the directory in which the application file is saved.

6. Conclusion

In the present paper is described an algorithm that uses the fractal dimension to uncover the behavior of the association rules in time dimension. Previous methods dealt with this problem by using a relation table-based structure and require multiple scans of the data. The integration of online analytical processing of data with the association rules mining considerably increases the effectiveness of the data analyzing process and the discovering of interesting relationships. Therefore OLAP operations are applied in the represented algorithm for selection of the frequent k-itemsets as well as for calculation of the fractal dimension of the relevant sets. It is successfully implemented the code, which realizes the proposed algorithm.

References

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

[2] D. Barbara, P. Chen. Using the Fractal Dimension to Cluster Datasets, In: *Proceedings of the* 6^{th} *International Conference on Knowledge Discovery and Data Mining (KDD-200)*, 2000, 260-264.

[3] D. Barbara, Z. Nazeri. Fractal Mining of Association Rules over Interval Data, Technical Report, *George Mason University*, 2000, 9.

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

[5] G. Bogdanova, 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, 4.

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

[7] G. Bogdanova, Tsv. Georgieva. Discovering the Association Rules in OLAP Data Cube with Daily Downloads of Folklore Materials, In: *Proceedings of the International Conference on Computer Systems and Technologies*, Varna, 16-17.06.2005 (to appear).

[8] H. Garcia-Molina, J. D. Ullman, J. Widom. Database Systems: The Complete Book, *Williams*, 2002, 1083 (in Russian).

[9] Tsv. Georgieva. 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, 88-98.

[10] K. Falconer. *Fractal Geometry: Mathematical Foundations and Applications*, John Wiley & Sons, Chichester, 1990, 287.

[11] J. Han. OLAP Mining: An Integration of OLAP with Data Mining, In: *Proc. 1997 IFIP Conf. Data Semantics (DS-7)*, Leysin, Switzerland, 1997, 1-11.

[12] M. Kamber, 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, 6.

[13] B. Mandelbrot. *The Fractal Objects: Form, Fortuity and Dimension*, University of Sofia "St. Kliment Ohridski", Sofia, 1996, 275 (Bulgarian).

[14] C. Traina, A. Traina, L. Wu, C. Faloutsos. Fast Feature Selection Using the Fractal Dimension, In: *XV Brazilian Symposium on Databases (SBBD)*, 2000.

[15] J. Ullman. http://www.db.stanford.edu/∼ullman/mining/mining.html.

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

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

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

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

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

Department of Information Technologies
University of Veliko Tarnovo,

e-mail: cv.georgieva@uni-vt.bg

Received 09.06.2006