About one Approach to the Development of Fuzzy OLAP Cube in Decision Making Support System in the Sphere of Foreign Policy

Gulnara Nabibayova
Institute of Information Technology, Baku, Azerbaijan
gulnara@iit.ab.az

Abstract— **This paper presents the fuzzification process on the data warehouse attribute of decision making support system in the sphere of foreign policy, as the result of which the fuzzy OLAP-cube is obtained. For this purpose, comparative analysis of clustering algorithms was fulfilled, and then selected algorithm CLARANS was applied in clustering the warehouse attributes, using medoids obtained as the result of clustering, the values of membership functions were derived.**

Keywords— fuzzification; clustering; OLAP-cube; data warehouse; attribute; medoid

I. INTRODUCTION

Recently, a data warehouse (DW) attracts increasing attention as an important tool for the institution management. Various organizations have become more acutely aware of the need to build DW, which became one of the most important elements of an enterprise infrastructure [1].

As a key element of decision making support systems (DMSS), DW is a single and rather large repository of data used for decision making. According to Inmon [2] DW - is "an object-oriented, time bound and unchanging collection of data to support management decision making". DW is built by consolidating the data from set of heterogeneous data sources, which are converted into a single format and contain information in a detailed and aggregated form. Data aggregation of multi-dimensional DW is performed by the attributes (dimensions) to facilitate decision-making. Detailed and aggregated are data stored in multidimensional OLAP-cube, which provides a rapid response to users realizing an quiry to DW.

Noted that the main advantage of DW over the other source types is the presence of the semantic layer, which enables the user to operate the subject field terms to form analytic query to the warehouse [3].

OLAP tools, which are the key elements of DW, provide a variety of operations, such as folding, unfolding, cross section, cut, ranging, cross-tables. As a result of these operations, users receive a response in numerical expressions; therefore they can be effective and useful only in those cases, when the users need to answer it in numerical expressions. However, the relationship between data and queries to the data are often of fuzzy nature. For example, it is very important for decision makers to response to the query - whether this or that result is good, medium or bad, that is, qualitative aspect of the result is very important for them. This issue is close for the sphere of foreign policy, as well. Solving the problem of human resource management in the directions, such as analysis of human resources, their placement, training, planning staff promotion, not the specific age of an employee is very important for policy-makers, but the fact that he/she is *young*, *middle aged* or *elderly*. Or, for example, requesting a date of any event not the exact date is important, but the fact that the event happened *recently* or *long ago*. To solve such requests DW should be able to provide responses to queries in linguistic terms, which are term-sets corresponding to linguistic variables. In this regard, currently the use of the equipments of the theory of fuzzy sets is widespread in information search issues [4].

The paper aims to develop a fuzzy OLAP-cube based on the DW information in the information and analytical systems for decision making support in the field of foreign policy.

To achieve this goal the following problems were solved:

- clustering algorithms PAM, CLARA and CLARANS were studied, their comparative analysis was implemented, on the basis of which algorithm CLARANS was chosen to be used;
- clustering algorithm CLARANS was used for the attributes "DATE" and "AGE", noting required number of clusters.
- the value of membership function were found with the help of medoids (medoid - the center of the cluster, which is one of its points) of obtained clusters to fuzzificate these attributes and to obtain the fuzzy OLAP-cube.

II. SELECTION OF CLUSTERING ALGORITHM FOR THE TASK

The main aim of clustering is to separate the studied set of objects into the groups of "similar" objects, called clusters. Clustering algorithms are classified as follows [7]. Two main categories of algorithms are hierarchical and non-hierarchical algorithms.

Unlike the hierarchical algorithms, non-hierarchical algorithms are based on the search of the optimal variant of partition of the n objects into k clusters. Note that the clusters obtained by partitioning method tend to have higher quality (i.e., the elements of a cluster are closer) than the clusters obtained by hierarchical method. Due to this, partitioning methods have become one of the main research directions using cluster analysis. Several partitioning methods were developed: basing on the k-means, k-medoids, etc. As the basic algorithm among them, we chose the method of k-medoids, for several reasons, the main of which is that the methods of *k*-medoids are tolerant to the presence of outliers, as well as the fact that the methods of *k*-medoids, described below, can work with large data sets quite effectively.

Algorithms based on the partitioning method pass two main stages [9]:

- an initial stage, in which an initial set of k objects is selected as medoids;
- an evaluation stage, in which an attempt is made to minimize the target function, usually based on the sum of the overall distance between unselected objects and their medoids, i.e.:

$$\sum_{j=1}^{n} d(r_i,s_j), \text{ where } s_j \in S \text{ and } d(r_i, s_j)<d(r_c,s_j),\ r_i,r_c \in R,\ r_i \neq r_c$$

The smaller is the sum of the distances between the medoids and all other objects in their clusters, the better is the clustering.

The three most well-known algorithms based on the method of k-medoids, are PAM (Partitioning Around Medoids), CLARA (Clustering LARge Applications) and CLARANS (Clustering Large Applications based upon RANdomized Search).

PAM is one of the primary algorithms based on the method of k-medoids [8].

The basic principle of PAM clustering process is to search all the objects, which currently are not medoids, in order to calculate the distances from the selected medoids. PAM algorithm leads to a high quality of the clusters, but as it tries each possible combination, and it works effectively for small data sets. And due to its computational complexity its use is not practical for clustering large data sets.

The computational complexity of PAM algorithm was the motivation to develop CLARA algorithm - clustering algorithm based on a sample [8]. CLARA selects a few data samples out of a data set and applies PAM on each sample and finds the sample medoids. As the sample is produced randomly, the sample medoids can be considered as the medoids of entire data set. To choose the best approximations, CLARA creates multiple samples and produces the best clustering at the output. Here, for more accuracy, clustering quality is determined basing on the average distance of all objects in the entire data set, not only the objects in the samples. The study [8] shows experimentally, that five samples of (40 +2) size give satisfactory results.

CLARANS was developed within the framework of spatial data analysis. Searching for a better medoid at the evaluation stage, CLARANS selects objects out of the remaining (n-k) objects randomly. The number of objects, selected at this stage, is limited to the parameter specified by the user (*maxNeighbor*). If a better solution is not found after *maxNeighbor* efforts, then a local optimum is considered to be achieved. The procedure continues as long as *numLocal* local optimum is found [10]. The CLARANS algorithm uses random search strategy to improve the algorithms PAM and CLARA in terms of efficiency (computational or time complexity) and the effectiveness (i.e., quality of a clustering - the average distance values), respectively.

III. Application of Medoids to Find Membership Function

The notions "mathematical theory of fuzzy sets" and "fuzzy logic" were first proposed by the American scientist of Azeri origin Lotfi Zadeh in 1965. The main reason of the emergence of new theory was the presence of fuzzy and approximate reasoning in describing the processes, systems and objects by a human.

The papers [5], [6] provide information about the fuzzy sets theory, as well as the description of fuzzy operators, various combinations which are used in operations on fuzzy OLAP-cube.

Let's remind the basic definitions. The characteristic of a fuzzy set is the membership function. Let us denote the degree of membership of x through μ (x) to the fuzzy set, which is a generalization of the characteristic function of an ordinary set. Then the fuzzy set C will be a set of ordered pairs in the form of C={μ(x)/x}, herewith μ(x) can take any value in the interval [0, 1], $x \in X$. Meanwhile, the value μ(x)=0 means the absence of the membership to the set of x, and μ(x)=1 is full membership to the set of x.

The environment of information search means a pair: query - database. In terms of non-fuzziness – fuzziness, the medium of information search can be in four options (Table 1) [4].

TABLE I. POSSIBLE FUZZINESS OPTIONS IN THE MEDIUM OF INFORMATION SEARCH

Query	Database
non-fuzzy	non-fuzzy
non-fuzzy	fuzzy
fuzzy	non-fuzzy
fuzzy	fuzzy

As shown in TABLE 1 the most common is the situation 4.

Let us study DW DMSS of foreign policy sphere, in which we should fuzzificate, for example, 2 attributes out of the entire set of attributes, in order to develop fuzzy OLAP-cube. These are the attributes of DATE and AGE. Let us apply clustering algorithm CLARANS for them, specifying the number of clusters k=3 per each. As a result of the algorithm performance for each attribute 3 medoids will be obtained, on the basis of which these attributes then will be fuzzificated. Note that 3 clusters obtained after the fuzzification will turn into 3 linguistic terms (term-sets). This is {*recently, not-long ago-long ago*} for the DATE attribute and {*young, middle aged, old*} for the AGE attribute. Herewith, the attributes DATE and AGE become linguistic variables. Note that besides the basic terms - *recently* and *long ago*, we used a modifier – *not long ago* for the attribute DATE.

The paper [11] shows how to obtain membership values (fazzification) using k medoids. In our case, medoids of 3 clusters are (m_1, m_2, m_3) after applying CLARANS. In a database with n attributes, each medoid can be represented as the vector $m_i=\{a_{il}.\ .\ .\ .\ .\ a_{in}\}$, where $i=\overline{1,3}$.

That is, we have:

$m_1=\{a_{1l}.\ .\ .\ .\ .\ a_{1n}\}$
$m_2=\{a_{2l}.\ .\ .\ .\ .\ a_{2n}\}$
$m_3=\{a_{3l}.\ .\ .\ .\ .\ a_{3n}\}$

The set $\{a_{lj},a_{2j},a_{3j}\}$ represents the value of 3 medoids of the j-th attribute, which will be fazzificated.

The process of attribute fazzification is described below.

The input data are:

J = set of values of the j-th attribute;

Fij(x) = membership function to obtain membership values of all the values of the j-th attribute in the i-th fuzzy set;

i - number of fuzzy sets.

The output data are:

The membership values of all the values of the j-th attribute in each fuzzy set

For all x є J we perform:

```
If i=1 /*for the first fuzzy set*/
$F_{ij}(x)=1.0$                          if  $x\leq a_{ij}$.
$F_{ij}(x)=(x-a_{2j})/ (a_{1j} -a_{2j})$       if  $a_{1j} <x<a_{2j}$.
$F_{ij}(x)=0$                            if  $x\geq a_{2j}$.

If i=2 /*for the second (middle) fuzzy set*/
$F_{ij}(x)=0$                            if  $x\leq a_{(i-1)j}$.
$F_{ij}(x)=(x-a_{(i-1)j})/a_{ij} -a_{(i-1)j}$      if  $a_{(i-1)j}<x<a_{ij}$.
$F_{ij}(x)=1.0$                          if  $x=a_{ij}$
$F_{ij}(x)=(x-a_{(i+1)j})/ (a_{ij} -a_{(i+1)j})$   if  $a_{ij} <x<a_{(i+1)j}$
$F_{ij}(x)=0$                            if  $x\geq a_{(i+1)j}$.

If i=3 /*for the third (last) fuzzy set*/

$F_{ij}(x)=0$                            if  $x\leq a_{(i-1)j}$.
$F_{ij}(x)=x-a_{(K-1)j}/a_{kj}-a_{(k-1)j}$       if  $a_{(i-1)j}<x<a_{ij}$.
$F_{ij}(x)=1$                            if  $x\geq a_{ij}$.
```

This process is applied to the input attributes DATE and AGE. Fig.1 and Fig.2 show a graphic description of the term-sets of corresponding attributes.

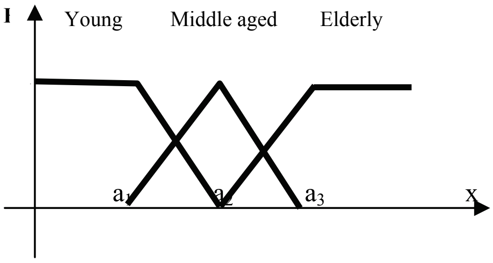

Figure 1. Graphical description of linguistic variables "AGE"

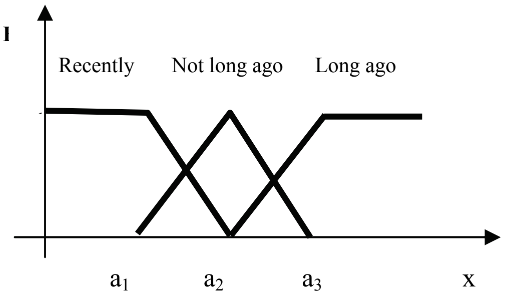

Figure 2. Graphical description of linguistic variables "DATE"

The TABLE II and TABLE III represent the approximate values of the attributes AGE and DATE before and after fazzification.

TABLE II. TABLE OF VALUES OF AGE AND DATE ATTRIBUTES BEFORE FAZZIFITCATION

AGE	LOCATION	DATE
30	Turkey	2010
42	Ukraine	2002
31	Russia	2008
35	Germany	2007
46	Turkey	2007
45	Georgia	2011
40	Ukraine	2005
42	Russia	2007
25	Germany	2001

TABLE III. TABLE OF VALUES OF AGE AND DATE ATTRIBUTES AFTER FAZZIFITCATION

AGE	LOCATION	DATE
Young	Turkey	Recently
Elderly	Ukraine	Long ago
Young	Russia	Not long ago
Young	Germany	Recently
Elderly	Turkey	not long ago
Elderly	Georgia	Recently
Middle aged	Ukraine	Not long ago
Middle aged	Russia	Recently
Young	Germany	Long ago

IV. CONCLUSION

Typical data warehouse that integrates data from multiple data sources, provides decision-makers with the quantitative analysis. And making foreign policy decisions in the management of human resources, financial resources in the sphere of foreign policy, which play an important role in strategic decision-making, decision-makers have to implement not only quantitative analysis, but also qualitative analysis. This is due to the fact, that the relationship between the data and query to the data are inherently fuzzy. The development of fuzzy OLAP cube provides decision makers to carry out a qualitative analysis.

REFERENCES

[1] A.Arsentev, Data warehouse becomes an infrastructure component. №1.CNews Analytics. In 2010. [Electronic resource].

[Http://retail.cnews.ru/reviews/free/BI2010/articles/articles6.shtml].

[2] W.Immon, Building the Data Warehouse, Second eds, John Wiley & sons Inc., New York, 1996

[3] N.B.Paklin, V.I.Oreshkov, "Business analytics: From Data to Knowledge", Publication PITER, 2010

[4] A.P.Ryzhov, Models of information search in a fuzzy environment. Publication house Mechanics and Mathematics Faculty of MSU, Moscow, 2004.

[5] A.P.Ryzhov. Elements of the theory of fuzzy sets and fuzzy measurement. Moscow, Dialog-MSU, 1998.

[6] Klir George J, Bo Yuan "Fuzzy sets and fuzzy logic", Prentice Hall, 1995

[7] I.A.Chubukova, Data Mining. Textbook. - M.: Internet-University of Information Technology; BINOM. Knowledge Laboratory, 2006. p. 382

[8] L.Kaufman, P.J. Rousseeuw, Finding Groups in Data: an Introduction to Cluster Analysis. John Wiley & Sons, 2005.

[9] Maria Camila N. Barioni, Humberto L. Razente, Agma J. M. Traina, Caetano Traina Jr. An efficient approach to scale up k-medoid based algorithms in large databases. XXI Brazilian Simposium on Databases. In 2006. http://www.lbd.dcc.ufmg.br:8080/colecoes/sbbd/2006/018.pdf

[10] Ng, R. T. and Han, J. Clarans: A method for clustering objects for spatial data mining. IEEE Transactions on Knowledge and Data Engineering (TKDE), 2002, 14 (5), Page (s) :1003-1016.

[11] KVNN Pavan Kumar1, P.Radha Krishna.Supriya Kumar De, Fuzzy OLAP Cube for Qualitative Analysis. Intelligent Sensing and Information Processing, 2005. Proceedings of 2005 International Conference on, 2005, Page (s): 290 – 295