A hybrid data mining model for effective citizen relationship management: a case study on Tehran municipality

Ali Mohammad Ahmadvand
Department of industrial engineering, Imam Hossein University
Tehran, Iran
Alimohamad.ahmadvand@gmail.com

Behrooz Minaei Bidgoli
Department of computer engineering, Iran university of Science and technology
Tehran, Iran
minaeibi@cse.msu.edu

Elham Akhondzadeh
Department of industrial engineering, Shaehd University
Tehran, Iran
e_akhondzadeh@yahoo.com

Abstract- Currently, many governments are actively promoting implementation of ICT to be more citizen-oriented. For effective citizen relationship management, it is important to identify the needs of different citizen groups and to provide respective services for each group accordingly. In this way, the application of data mining tools would be very useful to understand citizen's needs.

In this paper, focusing on the CiRM concept, we apply a data mining framework on the database of Tehran municipality. This framework consists of clustering and the association rule to improve citizen satisfaction. The main objective is to find the factors those affect the rate of satisfaction.

Firstly, we use the K-means algorithm to cluster the subjects that cause citizens complaint. Every data point is identified in terms of the following features: the frequency, the number of days that at least one complaint occurred and the interval time between the first and the latest time of each subject during a season. Secondly, the association rule is used to identify the factors that affect the rate of satisfaction in the cluster of subjects that occur regularly during the season and have a high number of complaints. The results of the research are very useful to build a strategy recommendation system in order to improve the rate of citizens' satisfaction. This study could be notable as one of the first studies on using data mining tools in CiRM.

Keywords- Citizen Relationship Management (CiRM); Urban service management system; Data mining; Clustering; Association rule.

I. INTRODUCTION

The governmental services are one of the most primitive and predominant service domains in any community with a wide array of services. As such, urban services and the concept of citizen relationship management (CiRM) are currently gaining importance in this domain [7, 8].

The main purpose of CiRM is to understand the needs of different citizen groups and to provide respective services for each group accordingly. In this way, many governments are actively promoting the use of ICT. Sasaki, T., A.Watanabe, Y. and Minamino, K. (2007) declared in [8] that the main issues of this approach are "how e-governments can manage effectively" and "be more citizen-oriented". In other words, ICT can be strategically significant in promoting e-government effectiveness through understanding the citizen requirements. In this domain, the application of data mining tools seems to be useful. Appropriate data mining tools, which are good at extracting and identifying useful information and knowledge from enormous customer databases, are one of the best supporting tools to obtain a deeper understanding of citizen’s characteristics and needs.

In this paper, focusing on CiRM, we apply data mining tools to improve citizen satisfaction. We use the database of Tehran municipality to find the most important citizens’ needs. The main objective of this paper is to find those subjects that have a high number of complaints, the root of them and the factors that affect the rate of satisfaction.

In order to accomplish this objective, we use a framework that consists of the clustering of subjects in terms of the frequency, the time interval between the first and the latest time and the number of days that at least one complaint of each subject occurred and extracting the association of the items which was identified as the prior citizens' needs via the cluster analysis. The result of this research is expected to be useful for development of citizen satisfaction.

The remaining of this paper is structured as follows. In Section 2, we review the studies related to CiRM, clustering and the association rule. In Section 3, we introduce the hybrid method that we use. In Section 4, we apply the proposed model on the data of Tehran municipality in Iran.

978-0-7695-3948-5/10 $26.00 © 2010 IEEE
DOI 10.1109/IC4E.2010.114

Finally, we conclude in Section 5 and summarize the study results which provide valuable information for the effective management of urban services.

II. Literature Review

In this section, we briefly review related studies of the Citizen Relationship Management, clustering, K-means algorithm, association rule and the Apriori algorithm.

A. CiRM

Citizen Relationship Management is the application of customer relationship management (CRM) in an e-governmental context and establishment guiding principles on how customer relations can be effectively strategized in the public sector. Specifically, it talks about the differences in CRM practices between public and private institutions. [7].

The essential purpose of CiRM is to change government-oriented management into citizen-oriented one in governments. In this way, it is necessary to understand what citizen needs and improve the citizen satisfaction. In fact, CiRM applies CRM (Customer Relationship Management) to improve citizen satisfaction [7, 8].

Pan,S., Tan, C., Lim, E. (2006) claimed in [7] that despite the irrefutable relevance of customer relations in influencing the adoption of e-government systems, there has not been a corresponding rise in the study of CRM practices within the context of public eservices.

1) Using ICT in CiRM

As discussed in the previous section, understanding citizen's needs and improving the citizen satisfaction are the most important points in CiRM. In this way, channels to understand citizen's needs are required. So, it is necessary to restructure the inner environment and communication tools for sharing opinion that citizens think. By the progress of ICT, it has been easy to achieve [8]. Many tools are used to connect between public office and citizen by the advantage of ICT, but according to Sasaki,T., A.Watanabe,Y. and Minamino,K.(2007) in [8], the public sector is not able to provide citizen-oriented services because of the bureaucratic sectionalism. The way to solve this problem is to introduce "the citizen's voice database" which all public officers are able to access the database. By using this kind of call center, it is possible to contact between the government and citizens and a review and a feedback process at the whole public office is constructed. However, it is important not only to construct the review and feedback process but also managing the framework to apply the opinion from citizens and the knowledge management to share their knowledge are effective and required[8]. In this way, the application of data mining tools is expected to be useful.

Many researchers have used data mining tools in CRM, but based on our researches, no one has used data mining techniques for CiRM. Ngai, .E.W.T., Xiu, L. and Chau, D.C.K. (2008) declared in [6] that Appropriate data mining tools, which are good at extracting and identifying useful information and knowledge from enormous customer databases, are one of the best supporting tools for making different CiRM decisions.

Data mining techniques could be applied to discover unseen patterns of citizens' complaints. The root of the problems may also be uncovered by investigating the association between complaints from different citizens.

B. Clustering

Clustering is the process of grouping a set of data objects into groups such that objects from the same cluster are more similar and the objects from different clusters are more dissimilar [4].

The difference between clustering and classification is that clusters are unknown at the time the algorithm starts. In fact, there are no predefined classes in clustering. Common tools for clustering include neural networks and discrimination analysis.

Clustering techniques have been used in CRM by Many researchers [1, 5, 6]. But based on our research no one has applied cluster analysis for citizen relationship management.

1) K-means algorithm

K-means is one of the well-known algorithms for clustering and it has been used extensively in various fields including data mining, statistical data analysis and other business applications[2]. The K-means algorithm for partitioning is based on the mean value of the objects in the cluster.

The K-means algorithm proceeds as follows:

First, it randomly selects K of the objects, each of which initially represents a cluster mean or center. For each of the remaining objects, an object is assigned to the cluster to which it is the most similar, based on the distance between the object and the cluster mean. It then computes the new mean for each cluster. This process iterates until the criterion function converges. Typically, the square-error criterion is used for cluster evaluation [4].

C. Association rule

According to Tan, P., Steinbach, M. and Kumar, V. (2006) in [9], Association aims to establish relationships between items which exist together in a given record. An association rule is an implication expression of the form $X \rightarrow Y$ where X and Y are disjoint itemsets, i.e., $X \cap Y = \phi$. The strength of an association rule can be measured according to the support and confidence metrics.

The support metric determines how often a rule is satisfied in the transaction database. It is obtained by dividing the support count for $X \cup Y$ by the total number of transactions. The confidence metric determines how often items in Y appear in transactions that contain X.

Originally, association rules emerged in the domain of shops and customers. It is one of the most commonly used data mining techniques in CRM. Market basket analysis is a typical example of this technique. Storekeepers use the results of market basket analysis for various marketing purposes, such as how to decide on what to put on sale, how to place merchandize on shelves to maximize a cross selling effect, and how to advertise. The association rule is not limited to marketing problems. It is widely applied to other decision making problems [3, 6].

Association techniques have been used in CRM by Many researchers [1, 5, 6]. But based on our research no one has used association rules for citizen relationship management.

Common tools for association modeling are statistics and apriori algorithms [6]. We use apriori algorithm in this paper to extract the association rules. The main characteristics of this algorithm are summarized as follows:

1. Apriori is a level-wise algorithm that generates frequent itemsets one level at-a-time, from itemsets of size-1 to the longest frequent itemsets

2. At each level, new candidate itemsets are created using frequent itemsets discovered at the previous level.

3. At each level, the transaction database is scanned once to determine the actual support count of every candidate itemset[9].

III. Methodology

In this study, we use a hybrid method incorporating clustering and association rule in to a data mining approach, in order to offer useful strategic information for different kinds of complaints. It enables the municipality to find the important points for effective management of urban services.

At first, we use clustering analysis to group the subjects that may cause citizens complaint. After that, we use association rules to discover the correlation among items that affect the rate of citizen satisfaction in the preferred cluster.

The mining process is introduced step by step as follows:

At step 1, we divide the subjects which cause complaint into several groups using K-Means clustering via following three measures:

1- Frequency of complaints: frequency refers to the number of complaints of each subject during a season, for example 100 times during the winter.

2- Time interval between complaints: time interval refers to the interval between the first and the latest time that a complaint of each subject happened during a season, for example 22 days.

3- Number of days: number of days refers to the number of days that at least one complaint of each subject occurred.

The advantage of clustering analysis is useful to identify groups of complaints with similar characteristics.

In step 2, we use association rules to determine which factors affect the satisfaction rate in those kinds of complaints which occur permanently during the season with a high frequency.

IV. Case Study

The Tehran municipality constructed an urban service management system by using the ICT technology to connect between the municipality and citizens to improve citizens' satisfaction. This system is focusing on managing the complaints against urban services and requirements. This call center has been established to provide appropriate and timely responses to citizens' demands and complaints. Currently, the urban service management system responses more than 3500 phone calls per day.

The database of this system includes proper information about citizens' needs and would be a key to new services and point of improvement. The data warehouse stores detailed information about citizens' calls; so we know when, why and at which point of geographical region, a complaint occurred.

Many researchers have used data mining tools in CRM, but based on our research, no one has used data mining for CIRM. Besides, the number of articles focusing on the application of data mining techniques in complaints management is low.

A. *The pre-processing phase*

Raw data of this study are one-year complaint data of the data warehouse from 2007/03/21 to 2008/03/19. This dataset is composed of 45 data fields and 1,116,249 records of complaints. We apply the 2-step method on the data which refers to the winter and 5 data fields among 45 data fields. We substitute the mode value for missing values and eliminate inaccurate values. The fields and an example of the data are shown in table 1 and table 2 respectively.

As the citizens' needs may be different from one geographical region to another, we use the hybrid method for each region.

As discussed in the previous section we decide to cluster the subjects via the following features: the frequency, the number of days that at least one complaint occurred and the interval time between the first and the latest time of each subject during a season.

At first, we should calculate the above attributes for each subject. In order to accomplish this objective, we create a new field as 'day of season' from the 'message time' field that shows the day of season on which the complaint occurred; for example the second day, the fifty forth day or the ninth day. By creating this field, it is possible to calculate the number of days on which each subject occurred and the time interval. Now a dataset is created that each data point shows a subject and is characterized by the proposed attributes.

TABLE I. THE DEFINITION OF ATTRIBUTES

ID	Attribute	definition
1	Subject ID	The kind of complaint
2	Region	The geographical point that a complaint occurs.
3	Message Time	The time of complaint
4	unit ID	The unit of the municipality that is responsible for the complaint
5	Last state ID	The state of being satisfied or dissatisfied

TABLE II. THE EXAMPLE OF DATA

Complaint ID	Subject ID	Region	Message Time	unit ID	Last state ID
12	413	7	07/02/2008 10:16	250	27
13	420	15	07/02/2008 10:16	154	28

B. Clustering

In this section, we use the K-means algorithm to cluster the subjects which cause complaint in region 1, via the frequency, the time interval and the number of days.

The K-means algorithm is required to define the number of initial clusters K. For that, we run the K-means algorithm using different values of K. according to the results, the K=3 is a suitable choice for K. The results are shown in table 3.

Based on the results, two items has been identified as outliers that refer to 'the removal of ice from slippery paths' and 'clearing snow from roads and paths' subjects and the remaining items are segmented in to three clusters, while 68 subjects belong to cluster 1, 34 subjects belong to cluster 2 and 32 subjects belong to cluster 3.

We also use the Tow-Step algorithm which has the advantage of automatically estimating the optimal number of clusters and handling large datasets. The result of the Two-Step algorithm shows that the number of clusters is three.

The characteristics of each group are as follows:

Cluster 1: most of the items in this cluster refer to the 'collection and installation', 'trees' and 'parks and green lands', 'employees behavior' and 'animals' problems. The average of the interval time is about 78, but the frequency and the number of days are on average 19 and 29 respectively. It means that these kinds of needs occur sometimes during the winter and not during a specified period of time; they occur almost during the whole winter with a low frequency.

Cluster 2: most of the subjects in this cluster refer to the 'construction' and 'traffic' problems. The average of frequency, interval time and the number of days are about 5, 13 and 4 respectively. It means that citizens grumble occasionally about these kinds of subjects and not during the whole winter. It seems that citizens are rather neutral about these subjects.

Cluster 3: most of the items in this cluster refer to the 'garbage and waste', 'cleaning', and 'asphalt'. The average of frequency, interval time and the number of days are about 358, 86 and 70 respectively. It means that these kinds of complaints occur permanently during the whole winter with a high frequency. It is notable that both subjects which we eliminated as outliers have the same characteristics as this cluster. Therefore, the complaints which refer to the 'snow', 'garbage and waste', 'cleaning' and 'asphalt' occur with the highest frequency and need more attention as the prior citizens' needs.

TABLE III. THE RESULTS OF K-MEANS ALGORITHM

K	2	3	4	5	6
SSE	52.32	22.88	46.28	34.22	30.21

C. Association rules

As discussed in the previous section, the 'snow', 'garbage and waste', 'cleaning' and 'asphalt' are specified as the important prior citizens' needs in region 1 and the municipality should give the top priority to them. These kinds of complaints include 15 different subjects. The results show that 80 percents of the complaints in this region refer to the above subjects.

The last state ID shows the state of being satisfied or not. We use this field and the subject to find the conditions in which citizens are dissatisfied. The results show that citizens are pleased about the service of the municipality in 11 subjects but are not always satisfied with the four remaining items which are identified in table 4. The feedback information shows that the performance of municipality in handling these kinds of complaints is sometimes acceptable, but not often.

Now, we use the association rules to find the factors that affect the state of satisfaction in handling these kinds of complaints. We analyze association rules using apriori algorithm by four input fields and confidence levels of 75%.

Table 5 shows 8 examples of the results. These rules show the state of satisfaction according to the month on which the complaint occurs and the responsible unit. For example the seventh rule means that when a complaint of 'asphalt layer' occurs in the first month of winter and the unit 250 is responsible for handling it, the state of satisfaction is dissatisfied with the probability of 100%.

The rules make it possible to understand the effect of 'time of complaint' and 'unit' for each subject on the rate of satisfaction.

Generally, a high rate of support and confidence has a strong mutual relation. But significant association rules are those which have a lift value higher than 1. All rules have shown the lift measure above 1. Therefore, we can say that the rules have a strong relationship.

Some results of this research are as follows:

The subjects of 'snow', 'garbage and waste', 'cleaning', and 'asphalt' occur permanently during the whole winter with a high frequency. So, the municipality performance is not proper in providing services which related with these subjects and the municipality should give top priority to the above items.

TABLE IV. THE DEFINITION OF SUBJECTS

Subject ID	Subject definition
379	asphalt settlement
380	asphalt layer
413	Garbage collection
524	Installation of garbage ban

TABLE V. EXAMPLES OF ASSOCIATION RULES

ID	Antecedent	Consequent	Support %	Confidence %	Lift
1	month = 12	satisfied	28.947	75.214	1.089
2	unit = 87	dissatisfied	0.987	95	3.234
3	subject = 379	satisfied	12.5	75.51	1.079
4	month = 10 and subject = 413	dissatisfied	0.987	100	3.234
5	month = 10 and unit = 83	dissatisfied	0.329	100	3.234
6	unit = 83 and subject = 524	satisfied	0.658	100	1.448
7	month = 10 and subject = 380 and unit = 250	dissatisfied	0.329	100	3.234
8	subject = 380 and month = 12 and unit = 250	satisfied	7.237	81.818	1.184

The municipality performance is not proper in handling the complaints which refer to 'asphalt settlement', 'asphalt layer', 'garbage collection' and 'installation of garbage ban' subjects.

The probability of being satisfied is 70% in the last month of winter.

The unit 82 which refers to the section 2 of region 1 has the best performance in the third month of winter in comparison with the other months in handling the 'garbage collection' subject.

The performance of municipality in handling the complaints of 'installation of garbage ban' subject in the first month of winter is very perfect and citizens feel satisfied with the probability of 100%.

The unit 83 which refers to the geographical section 3 of this region has the best performance and can be seen as a benchmark for the other sections.

Citizens feel dissatisfied with the unit 87 which refers to the section 5 of this region with the probability of 95%. This unit has a good performance just in handling the complaints which refer to the 'asphalt settlement' subject.

The subjects of 'asphalt layer' and 'garbage collection' cause a high number of complaints in the first month of winter while the municipality performance is worse.

V. CONCLUSTION

Governments have to consider convenient channels and services to connect between the governmental managers and citizens. Furthermore, data mining tools are needed to manage citizens' requirements and process which collect, analyze, reflect, and evaluate their needs.

In this research, we have used clustering and association rules on the data of the urban service management system in Iran to find the subjects that cause complaint and the factors that affect the rate of satisfaction. The results show that the municipality should give top priority to the 'snow', 'garbage and waste', 'cleaning' and 'asphalt' requirements and specially to the 'asphalt settlement', 'asphalt layer', 'Garbage collection' and 'Installation of garbage ban' subjects during the winter in geographical region1.

Analyzing the rules, make it possible to understand the impact of factors such as time and responsible units on the rate of satisfaction. Besides, units with a perfect or worse performance in providing services and handling complaints are identified.

The results of the research are very beneficial in providing improved urban services and the development of citizens' satisfaction. This study could be notable as one of the first studies on using data mining tools in CiRM.

REFERENCES

[1] Ahn. J. And Young, S., "Customer pattern search for after-sales service in manufacturing". Journal of Expert Systems with Applications, vol.36, pp.5371–5375, 2009.

[2] Cheng, Ch. and Chen, Y., "Classifying the segmentation of customer value via RFM model and RS theory", journal of Expert Systems with Applications, vol.36, pp.4176–4184, 2009.

[3] Cock, M. D., Cornelis, C. and Kerre, E. E., "Elicitation of fuzzy association rules from positive and negative examples. Journal of Fuzzy Sets and Systems", vol.149, pp.73–85, 2005.

[4] Han, J. and Kamber, M., Data Mining: Concepts and Techniques, Second ed., Morgan Kaufman Publisher, 2006, pp. 383-407.

[5] Jukic, N. And Nestorov, S., "Comprehensive data warehouse exploration with qualified association-rule mining", Journal of Decision Support Systems, vol.42, pp.859–87, 2006.

[6] Ngai, .E.W.T., Xiu, L. and Chau, D.C.K., "Application of data mining techniques in customer relationship management: A literature review and classification", journal of Expert Systems with Applications, vol. 36, pp. 2592–2602, 2008.

[7] Pan,S., Tan, C., Lim, E., "Customer relationship management (CRM) in e-government: a relational perspective", Decision Support Systems, vol. 42, pp. 237– 250, 2006.

[8] Sasaki,T., A.Watanabe,Y. and Minamino,K, "An Empirical Study on Citizen Relationship Management in Japan", Proc. PICMET conference of IEEE Xplore., Aug. 2007, pp. 2820-2823.

[9] Tan, P., Steinbach, M. And Kumar, V, Introduction to Data Mining, Addison Wesley, ISBN: 0-321-32136-7, 2006, pp. 171-180.