International Journal of Electrical and Computer Engineering (IJECE)
Vol. 4, No. 1, Feburary 2014, pp. 7~15
ISSN: 2088-8708

Improving Semantic Clustering Using Ontology and Rules

Elham Bahrami Foroutan*, Hasan Khotanlou **
* Departement of Electrical and Computer Engineering, Malayer Azad University
** Departement of Electrical Engineering, Bu-Ali Sina University

Article Info

Article history:

Received Aug 8, 2013
Revised Dec 18, 2013
Accepted Jan 3, 2014

Keyword:

Clustering
Ontology
Protégé
Rapidmine
Retrieve
Semantic

ABSTRACT

Access to huge data is needed for an appropriate structure and grouping of data such that the access to the data becomes easier, the status which clustering algorithms are doing this for us. However, special attentions are paid in recent years on semantic data clustering which in semantic interpretation of the input data is needed. In this paper, three modified clustering methods are used and the results of these techniques are evaluated. Based on this, first a technique is developed in which some rules are applied to prevent confusion within clusters. A rule-based clustering can be applied to the given data. Then, next technique performs these rules with applying ontology-based semantics. And last and basic technique changes presumed ontology and then rules applied on clusters. The result shows that the clusters derived from the information provided within them were very similar and very different from other clusters and a significant reduction in the k-distance of these clusters is also occurred and the correlation is increased.

Corresponding Author:

Elham Bahrami Foroutan
Departement of Electrical and Computer Engineering, Malayer
Malayer Azad University,
Email:elhforoutan@gmail.com

1. INTRODUCTION

Choosing an appropriate clustering for data mining is a very important part in the discussion of the data processing, because different clustering techniques have been considered in the literature. For the reason that wide variety of clustering techniques do not specify which clustering is beneficial to obtain good results of existing data in a dataset that contains not only numeric data values but also include semantic values, a method for ontology-based clustering is formed, recently. A large variety of clustering techniques have been developed to create clusters so that data within a cluster are very similar and have the greatest differences between different clusters. Clustering means to cluster data based on numerical aspect. It is clear that in this form, text data are not considered and the data within clusters are highly heterogeneous. Because of this reason, a technique has been applied that emphasizes on the discovery of knowledge obtained from data using clustering [1]. Some works have been practiced in the community objects which have created some generalizations [2] and as a result a structure of a complex dataset is provided. An idea was that the intelligent data analysis methods how to improve the semantic knowledge that was called Domain-Driven Data Mining (DDDM) [3]. Among the related tasks is [4] that are done in the new field DDDM [5]. Thus the internal data by clustering algorithm is divided into specified number of clusters. For this clustering, we use a data matrix where the rows are properties regarding the features available in [6] that is numeric or a comparison between categorical features based on equal or unequal values of them [7,8]. Many works in the numerical and categorical variable field has been done in [9]. A source of knowledge can assess semantic association and similarity based on existing evidence [10].

There are some methods to calculate similarity semantic between concepts and terminology that differ in the distribution information of their primary sources. There is a method in which an ontology and IC is used which provides a structured view on the basis of similarity calculation [13]. Another way creates an

ontology that is based on a graph model [14, 15 and, 16]. Finally, to interpret the semantic characteristics, similarity or distance between terms is calculated based on a hierarchical clustering algorithm [17]. We have also used this approach.

The non-numeric data management in traditional AI methods can be applied. Recently, ontology-based semantic clustering technique is introduced by the name that is presented by Batet. In addition, numerical features and the clustering result will be more efficient. Semantic features consider based on existing ontology and clustering is done based on all the features [18]. The purpose of this paper is to improve ontology-based clustering. The features used in this method were not considered by Batet and, ontology has been changed so that a more regular profile of dataset is available and the rules have been created so that they satisfy in their clusters. In this work, rules are applied for cluster ordering and ontology to encompass semantic features. Finally, we modify an existing ontology so that clusters within the data will result more consistently. This ontology was modified by adding a feature that will affect the result of clustering. In section 2, a review of proposed clustering algorithm and the used methods are presented fully theoretically without implementation of its simulation. In section 3, the final implementation and results are fully expressed.

2. SUMMARY OF CLUSTERING ALGORITHM

Although clustering is usually done simply, it seems that the use of matrix and ontology and restrictions of the clustering improve the results. Then we introduce a clustering algorithm based on ontology that in addition to the applying restrictions uses different ontology. The ontology-based clustering for travel and accommodation will be done.

2.1. Semantic Clustering of the Proposed Rules

1- Constructing a matrix for two-dimensional properties.
2- Do the following procedure as long as the cluster size is larger than one.
- I- If rules are true then do merge operations
- II- Do hierarchical clustering with regard to applying restrictions.
- III- Do evaluation on result matrix and clustering

$\forall\ i\ 1 \le i \le n : c_i := \{x_i\}$
$C := C' := \{c_1, \ldots, c_n\}$
$E := \varnothing$
$j := n+1$
matrix([n]*[n])
while ($|C'|>1$) **do**
if (rule=true) **then**
$c_j = c_{u'} \cup c_{v'}$
$C' = C' \cup \{ c_j \} - \{ c_{u'}, c_{v'} \}$
$C = C \cup \{ c_j \}$
$E = E \cup \{(c_{u'}, c_j), (c_{v'}, c_j)\}$
End if
C'=clustering(C)
end while
validation(C')
distance(matrix)
outlier(C')
correlation(matrix)

The matrix which we used is two-dimensional, i.e. instead of one aspect it considers several features. In the '*while'* condition, clusters are merged or deleted based on considered rules [18] and, the result is that the cluster will have to satisfy the rules. Similar to this method is used in [18], however, the merge and delete are not related to rules and, if the cluster is larger than 1, the merge and Elimination can be done everywhere.

Clustering function does hierarchical clustering while the ontology-based semantic clustering before applying this function must be applied. This operation is performed by a semantic function. This method is described below:

2.2. Proposed Ontology-Based Semantic Clustering Algorithm

The used algorithm is that in addition of the mentioned features, semantic features according to [18] is applied. However, the ontology used here has been changed.

The proposed ontology-based clustering is as following:

1- Create a matrix for two-dimensional properties

2- As long as, the cluster size is larger than one the following procedure could be done:

I- If law is true then do merge operation

II- add a certain number to ontology

III- Do Semantics

IV- Hierarchical clustering was performed with regard to applying law

V- 3- Do evaluation on clustering results and matrix

$\forall\ i\ 1 \le i \le n : c_i := \{x_i\}$
$C := C' := \{c_1, \ldots, c_n\}$
$E := \varnothing$
$O := \{o_1, o_2, ..., o_n\}\ \therefore$
$j := n+1$
matrix([n]*[n])
while ($|C'|>1$) **do**
if (rule=true) **then**
$c_j = c_{u'} \cup c_{v'}$
$C' = C' \cup \{ c_j \} - \{ c_{u'}, c_{v'} \}$
$C = C \cup \{ c_j \}$
$E = E \cup \{(c_{u'}, c_j), (c_{v'}, c_j)\}$
End if
$O = O \cup \{o_i, o_j\}$
C'= semantic($O \cup C$)
C'=clustering(C')
end while
validation(C')
distance(matrix)
outlier(C')
correlation(matrix)

In this part, features are clustered based on their meanings. These meanings can be done by installing an ontology. The ontology, attributes are based on concepts. Here create a matrix of features. Then the rules are applied to merge or delete in cluster Then the desired ontology are added to the number who have considered. Here are used WordNet and reasoning and accommodation ontology. And the semantics of clustering is done in a loop sequentially, The result is that we have a good correlation with the clustering.

3. IMPLEMENTATION AND EVALUATION

In this method, at first, the dataset is uploaded. This dataset will be a variety of features, some of the attributes of clusters usually are not considered. Therefore, we can apply the rules on clusters in a way that the clustering is limited to consideration of these desired features. In this section, an ontology-based semantic clustering are performed on the same dataset of Delta Visitors [18]. In this work, we have tried to collect data from Delta Natural Park, therefore, based on the available statistics in the literature and many of the resources, our view of the dataset are collected and created [18].

In spite of some probable differences in the original dataset, all the evaluation steps prove the properly progress in semantic clustering. Estimation is performed in two below parts: First, the evaluation of the simple clustering is done by rule applying. The assessment is performed via ontology-based clustering. The following group of similar features are considered which included four variables in order to characterize the tourist profile (Origin, age, social class and associated entities) and six variables to model the profile of travel (plan, the first coming, the second coming, accommodation, length of stay, interests or loyalty).

3.1. Implementation and Evaluation of a Simple Clustering by Rules

Since previous studies do not consider the application of intelligent data analysis [6], we first run the clustering by the traditional process of classified features. In this experiment, age, stage, and interests are

considered as numerical properties and the origin, with individuals, social classes, plan, accommodation and, first and second reasons are classified and considered.

Rapid-Miner software is used for simulation. In Figure 1, the dataset can be loaded in our view by the Retrieve procedure. Then we entered the Matrix. The data matrix can be loaded with several features and the two-dimensional space is achieved. To accomplish the clustering procedure as well as possible, creation of missing values should be avoided. That's why we use the Replace Missing Value Operator. Then a weight is given to each node. Finally, as it will be shown, the procedure is activated.

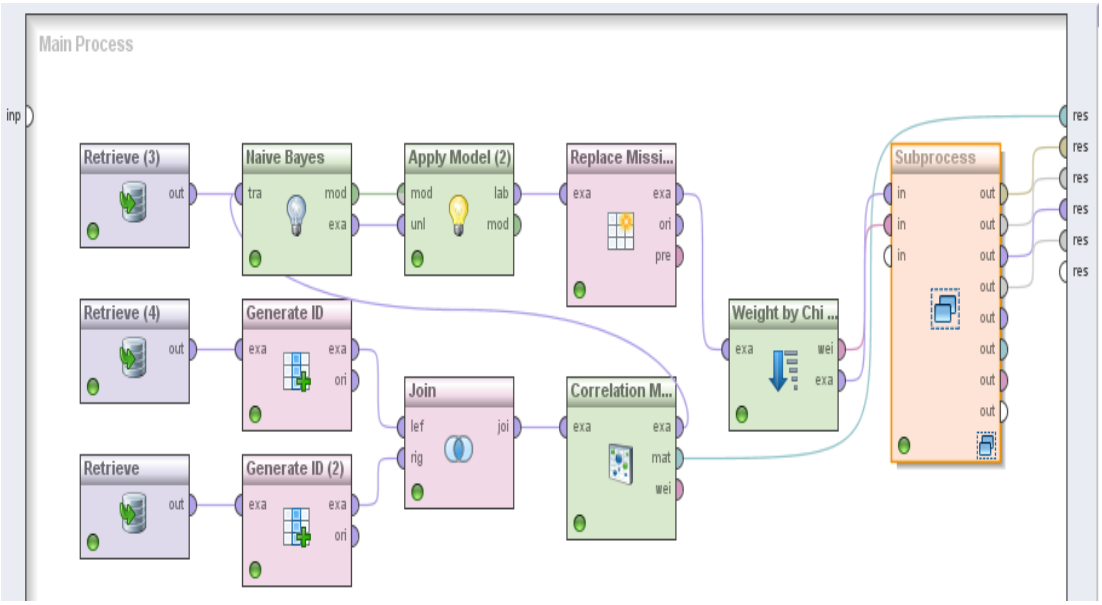

Figure 1. Clustering with applying rule

Here, properties of four religious buildings, office buildings and, the beach and gastronomy which were not considered before is combined with the original dataset. To improve clustering, we apply the rules which cause to contribute those features that were not considered in the formation of clusters. Therefore, we use the Apply Model which in Confidence and prediction rules should apply where the clustering is performed based on the values of these functions and rather than one feature multiple features join together. Data can be loaded in clustering operations in sub-procedure and traditional clustering operations are performed. Finally, validation and similarity, and calculation of the outlier are accomplished. Three sample classes in this clustering method are:

- Class 927: 47% of them are Catalanand have come to the excursion. 35% of Catalan people have come with their families and their travel reasons are various; 77% go to church and the rate of accommodation is heterogeneous. Prediction (religions building) rule for this cluster is equal to church. It means that most of them visit the church. However, Confidence rule takes various values for different parameters.
- Class 926: half of them came to innovation and improvisation who are often strangers and traveled with their family and 33% were for fun who are mostly Hispanic but the prediction rule (religions building) is same as the church for these clusters, most of them have one home, the length of stay is between zero and one.
- Class 907: Most of them are young. They are strangers and classmates and came for relaxation. Most of them hire camps. Prediction (religions building) rule for this cluster is same as the church. Using this clustering, there is some heterogeneity among clusters and outliers are about 0.155 and, the K-distance that is k = 3 in default also shown in Figure2.

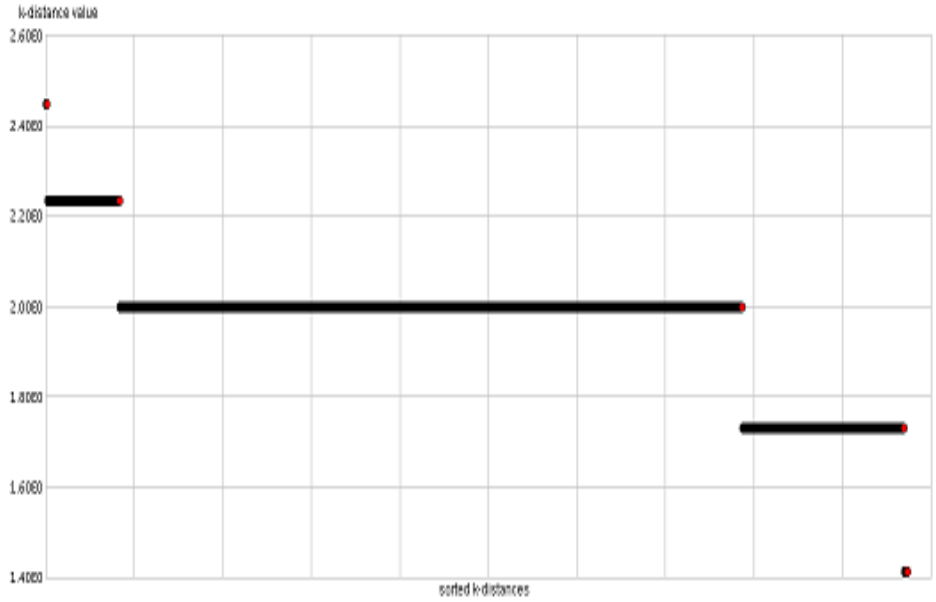

Figure 2. k-distance for a simple clustering

After normalizing k-distance, average distance becomes about 0.511 that according to the figures contained in the articles is a big distant. As a result, there is a low correlation between components.

Now that we got familiar with the operation of the clustering process we start the procedure. As running the procedure, the accuracy will be about 26/36% that is very low. There is a tree model validation which assesses the length of their stay according to age and whether they are traveling alone or with their friends or they are local or not and the level of their use of accommodation. It is determined by this evaluating that those number of tourists whose duration stay is less than 2 days usually use hotels and camps in comparison with others and they usually reserved and they are strangers from the middle/high class. The outliers is about 0.155 that can be improved. However, the correlation matrix is formed based on the values entered by the Apply Model.

The obtained profile here is more complete than a simple clustering because the applied rules improve the strength of clusters.

3.2. Assessment of Ontology-Based Clustering

In this part of the procedure, the ontology must be loaded. Thus, first we explain the used ontology with its construction. Ontology is used for the knowledge about some interested domains. Ontology describes the concepts in that domain and maintains the relationships between those concepts. In this work, the protégé application software is used to build the ontology. In this dataset, there are two important columns that we want to define them as the classes in the ontology. These columns are the type of the accommodation and the travel reasons [19]. In the classifications of the ontology, accommodation and camps and hotels and cottages are located in the simple sub-class and Houses and rental homes and apartments are located the fixed sub-class. The reason of travel, biking, fishing, eating, photography and gastronomy are in the sub-class of entertainment and the sport is in unknown sub-class, beach, mountains, flora, landscape and nature are in the environment sub-class and culture, family, holidays, sightseeing and relaxation are in priority sub-class and trade, ecology, education, drills, loyalty, accommodation are in principle sub-class and offer, chance, visit are in random sub-class.

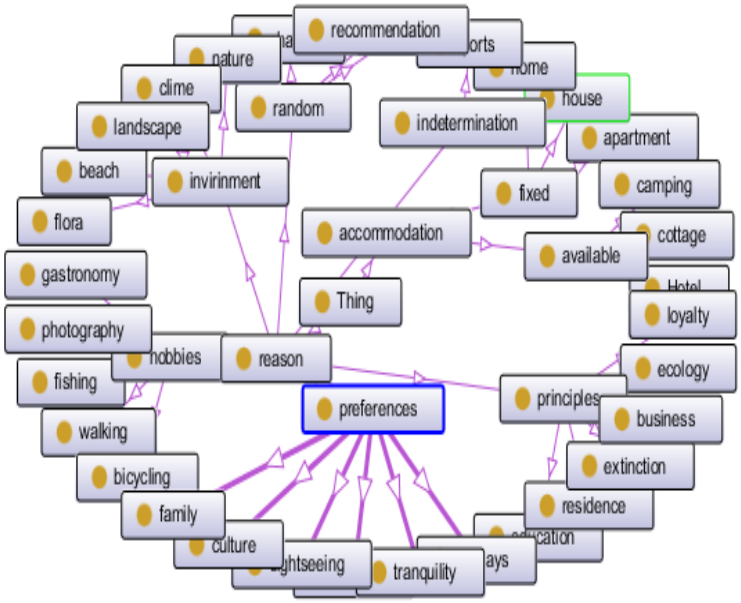

Figure 3. Ontograph

Then the ontology in Rapid Miner software should be loaded.

3.2.1. Designing of Clustering Using Ontology and Rules

Like the previous work with the following differences:

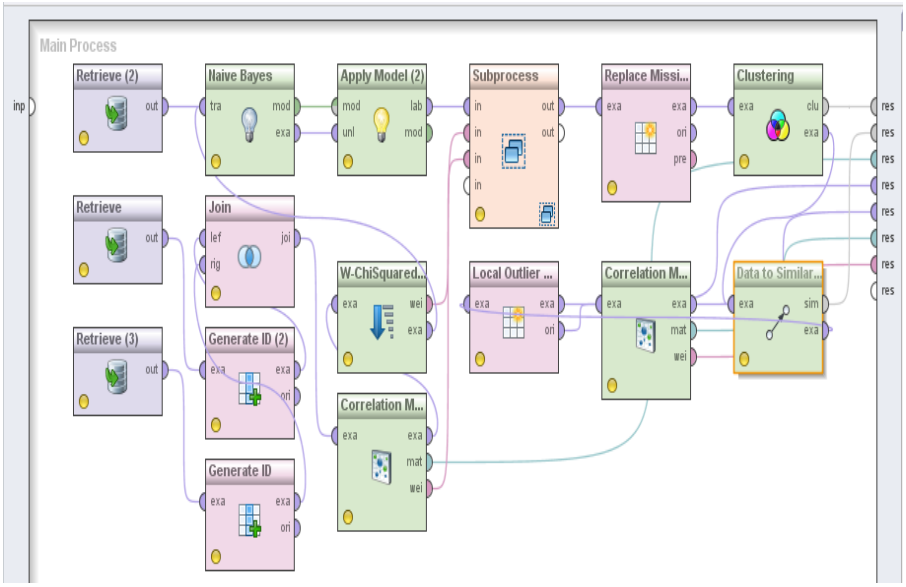

Figure 4. Ontology-based clustering with applying rule

Here the data is loaded with the matrix in the sub-procedure. The semantic k-mediod, an input is our main input of procedures and the other input is the ontology which is loaded in it. Then results can be returned by Apply Model to the procedure. First, in the basic procedure, the missing value problem is solved. Then clustering operations are performed. Finally, the results are evaluated, distance and outliers and correlation matrix are specified.

The obtained Dendrogram is:

- Class 993: Most of them are Spanish and have come to visit, by innovation and improvisation. They all own houses, and their reason is going to the beach or have other reasons. All of them are young and have come with their families.
- Class 989: All are Catalan, they have come with their classmates, education is their main reason and, most of them are from low- middle class and hire camps.
- Class 942: Most of them are foreigners and some of them are Catalan and have come for excursion. 73% of them have come with their classmates and others with their friends. Their social level is high or high - medium. Their reasons are 26% for education and 30% by chance.
- Class 943: 53% of them have come for excursion and 31% by improvisation or innovation. 53 % of them are Catalan, and 31% are Hispanic. Half of them own a house. Their reasons are nature, landscape and relaxation.
- Class 977: 77% of them have come by innovation and are Catalan. They use apartments and have come for relaxation.
- Class 960: Most of them have come for excursion and are Catalan. They have come with their families and have a second home or hotel.

It can be found by observing the clustering that heterogeneity among clusters is almost gone and outliers are about 119/0. K-distance is also shown in Figure 5 which is as default k = 3:

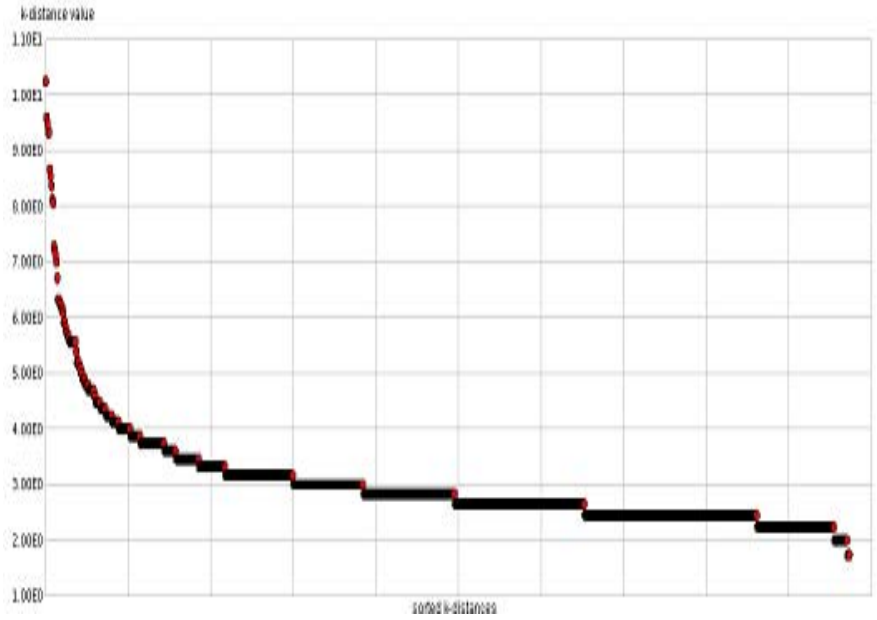

Figure 5. k-distance for ontology-based clustering

After normalizing the k-distance, the average of distance becomes about 246/0 which has been significantly reduced. Correlation among components in the matrix is increased significantly.

Four properties, beach, religious buildings, official buildings and, gastronomy which were not considered before are combined with the original dataset. To improve the clustering and to add other features in the clusters in form of two-dimensional, the Apply Model and rules on the cluster are applied. Such rules cause the features that have been considered, contribute to the formation of clusters.

In this procedure, the dataset can be loaded in the "join" then we form the matrix and do the weighting by chi-Square. As a result, all of these can be loaded in" Naïve bayes", output along with the columns that we want to enforce the rule on them go to the "Apply Model", so the input of the sub-procedure is loaded. The ontology is loaded in sub-procedure and at the end, clustering is performed on the input and evaluation is performed. Dendrogram is identical to previous method, the difference is that clustering have the matrix of properties and the rule is applied on the clusters and more complete profile will be available for visitors. There are no differences in outliers and k-distance. However, its correlation matrix is changed which gives additional information over the applied rule and their correlation.

Here, the correlation is increased, the maximum correlation is the 0.128 and, the frequency rose to 65. As a result, with this semantic clustering that uses the ontology to search for more information and a more homogeneous dataset, clusters with a more complete profile of the tourists are available and more features are included.

4. CONCLUSION AND FUTURE WORK

We developed a dataset with three different methods of tourist data. This dataset contains numerical, categorical and semantic of features. Several features are included which allow to expand our operations in the semantic context. Comparison of this model with prior conventional models results shows that outliers and the k-distance in the model are reduced. To evaluate the results, "apply model" and "correlation" tools are used. In the previous types of clustering [1, 12], only a portion of the data within the clusters are identical and clustering is done based on it. In this way, clusters have more consistent data. The quantitative results of this ontology-based model in comparison with other models proves that in this case study, the k-distance is 0.128 which is dropped and correlation is risen to 75 percent, outliers are reduced by the use of rule enforcement (0.119).

Table 1. Evaluation of the results of ontology-based clustering using applicable rules via simple method

Kind of clustering	k-distance	outlier	correlation
Simple clustering	0.511	0.155	37.5
Clustering with applying rule	0.246	0.155	37.5
Ontology-base clustering with applying rule	0.128	0.119	75

This research is performed based on the dataset of Delta Natural Park. However, it can be done over Iran's touristic areas and the development of tourism in these areas can be estimated. As an example, in the case of Ganjnameh-a historical place in Hamadan City of Iran- visitors can be asked to fill in a detailed questionnaire to create a dataset based on the tourist profile. Among this answer, the main reason for visiting this place and the level of satisfaction and other factors can be examined. Another suggestion is that the ontology can be modified so that it can change the clustering to consider more complete profiles of visitors or to define more rules so that heterogeneous clusters can be broken and homogeneous clusters can be merged together.

REFERENCES

[1] Han J and M Kamber. *Data Mining: Concepts and Techniques*. Morgan, Kaufmann. 2000.
[2] Mirkin B. *Clustering for data mining: a data recovery approach*. London, Chapman & Hall/CRC. 2005.
[3] Cao L, PS Yu, C Zhang and Y Zhao. *Domain Driven Data Mining*. Springer. 2010.
[4] Fan B. "A hybrid spatial data clustering method for site selection: The data driven approach of GIS mining". Expert Systems with Applications. 2009; 36(2): 3923-3936.
[5] Xu R and D Wunsch. "*Survey of clustering algorithms*". IEEE Transactions on Neural Networks. 2005; 16(3): 645-678.
[6] Gibert K and U Cortés. "*Weighing quantitative and qualitative variables in clustering methods*". Mathware and Soft Computing. 1997; 4(3): 251-266.
[7] Walesiak M. "*Walesiak*". Argumenta Oeconomica. 1999; 2(8): 167-173.
[8] Jajuga K, M Walesiak and A Bak. *On the general distance measure. Exploratory Data Analysis in Empirical Research*. M Schwaiger and O Opitz. 2003.
[9] Gibert K, R Nonell, JM Velarde and MM Colillas. "*Knowledge* Discovery with clustering: impact of metrics and reporting phase by using KLASS". Neural Network World. 2005; 15(4): 319-326.
[10] Studer R, VR Benjamins and D Fensel. "*Knowledge Engineering: Principles and Methods*". Data and Knowledge Engineering. 1998; 25(1-2)(1-2): 161-197.
[11] Etzioni O, M Cafarella, D Downey, A Popescu, T Shaked, S Soderland, D Weld and A Yates. "*Unsupervised named-entity extraction form the Web: An experimental study*". Artificial Intelligence. 2005; 165: 91-134.
[12] Landauer T and S Dumais. "*A Solution to Plato's Problem: The Latent Semantic Analysis Theory of the Acquisition, Induction, and Representation of Knowledge*". Psychological Review. 1997; 104: 211-240.
[13] Resnik P. *Using Information Content to Evalutate Semantic Similarity in a Taxonomy.* 14th International Joint Conference on Artificial Intelligence, IJCAI 1995, Montreal, Quebec, Canada, Morgan Kaufmann Publishers Inc. 1995: 448-453.

[14] Leacock C and M Chodorow. *Combining local context and WordNet similarity for word sense identification.* WordNet: An electronic lexical database, MIT Press. 1998: 265-283.
[15] Rada R, H Mili, E Bichnell and M Blettner. "*Development and application of a metric on semantic nets*". IEEE Transactions on Systems, Man, and Cybernetics. 1989; 9(1): 17-30.
[16] Wu Z and M Palmer. *Verb semantics and lexical selection*. 32nd annual Meeting of the Association for Computational Linguistics, Las Cruces, New Mexico, Association for Computational Linguistics. 1994: 133-138.
[17] Gibert K. "*The use of symbolic information in automation of statistical treatment for ill-structured domains*". AI Communications. 1996; 9(1): 36-37.
[18] Batet M, K Gibert and A Valls. *Semantic Clustering Based On Ontologies:An Application to the Study of Visitors in a Natural Reserve (in press)*. 3th International Conference on Agents and Artificial Intelligence, Rome, Italy. 2011.