Journal of Engineering Technology and Sciences
Volume 02, Issue 05, May 2025 *ISSN: 3048-913X*
www.jets.org.in *DOI:10.63590/jets.2025.v02.i05.pp13-18*

Improving Semantic Clustering Using Ontology and Rules

Li shenghui
Department of Computer Science, Fudan University, China
Corresponding Author: shenghuicse@gmail.com

To Cite this Article
Li Shenghui, "Improving Semantic Clustering Using Ontology and Rules", Journal of Engineering Technology and Sciences, Vol. 02, Issue 05, May 2025, pp: 13-18, DOI: http://doi.org/10.63590/jets.2025.v02.i05.pp13-18
Submitted: 10-02-2025 *Accepted: 18-04-2025* *Published: 10-05-2025*

Abstract*: Large amounts of data must be accessible in order to be properly organised and grouped so that access to the data is made easier; clustering algorithms are now doing this for us. Semantic data clustering, which is necessary for semantic interpretation of the input data, has received particular attention in recent years. Three modified clustering approaches are applied in this article, and their outcomes are assessed. On the basis of this, a method is initially created that applies a few principles to avoid confusion inside clusters. The provided data can be subjected to a rule-based clustering. The following method then applies ontology-based semantics to carry out these rules. Finally, the fundamental method modifies the assumed ontology before applying rules to clusters. The outcome demonstrates that the clusters formed from the data included within them were both highly similar and highly dissimilar from one another. Additionally, there was a notable decrease in the k-distance of these clusters, and the correlation was raised.*

Keywords*: Ontology, Protege, Clustering, Rapid mine, Retrieve, and Semantic*

I. Introduction

Given that several clustering strategies have been examined in the literature, selecting the best clustering for data mining is a crucial component of the data processing discussion. A method for ontology-based clustering has recently emerged since the large range of clustering approaches does not indicate which clustering is advantageous to produce good results of current data in a dataset that contains both numeric and semantic data values. Numerous clustering approaches have been developed to form clusters, with the goal of having the most differences between clusters and the most comparable data within a cluster. Clustering is the process of grouping data according to numerical characteristics. It is evident that the data inside clusters is quite varied and that text data is not taken into account in this form. As a result, a method that focusses on the finding of knowledge gleaned from data through clustering has been used [1]. A structure for a complex dataset is given as a result of some work that has been done in the community objects that has led to certain generalizations [2].

One concept, known as Domain-Driven Data Mining (DDDM), was the use of intelligent data analysis techniques to enhance semantic knowledge [3]. One of the relevant works completed in the new field of DDDM [5] is [4]. As a result, the clustering algorithm divides the internal data into a predetermined number of clusters. We employ a data matrix for this clustering, with the rows representing attributes of the characteristics that are available in [6], which are either numerical or a comparison of categorical features according to their equal or unequal values [7, 8]. Numerous studies have been conducted in the realm of numerical and categorical variables [9]. Based on the evidence already available, a source of knowledge can evaluate semantic association and similarity [10]. There are several ways to determine the semantic similarity of terms and concepts that have different distribution data from their original sources. There is a technique that uses an IC and an ontology to create a structured view based on similarity calculations [13]. An alternative approach develops an ontology based on a graph model. Lastly, a hierarchical clustering technique is used to determine the similarity or distance between phrases in order to interpret the semantic properties. We have also employed this strategy. Traditional AI techniques can be used to manage non-numerical data.

Batet has recently created a semantic clustering technique that is based on ontologies. Furthermore, the clustering result and numerical features will be more effective. Based on the current ontology, semantic features are taken into consideration, and clustering is carried out using all of the features. This work aims at enhancing ontology-based clustering. Batet did not take into account the features employed in this method. The ontology has been modified to provide a more regular dataset profile, and rules have been developed to ensure that they satisfy in their clusters. This work applies ontology and cluster ordering principles to include semantic aspects. Lastly, we adapt an existing ontology to produce more consistent clusters within the data. A feature was added to this ontology that would impact the clustering outcome. A study of the suggested clustering algorithm and the employed techniques are provided in section 2 in a comprehensive theoretical manner without the usage of simulation. The results and final implementation are clearly described in section 3.

II. Clustering Algorithm Summary

Despite the fact that clustering is typically done simply, it appears that the results are improved by using matrices, ontologies, and clustering limitations[13]. Next, we provide an ontology-based clustering technique that employs a distinct ontology in addition to applying limitations. There will be ontology-based clustering for lodging and travel. The matrix we utilized is two-dimensional, meaning it takes into account multiple features rather than just one. Clusters are combined or removed in the "while" condition according to the rules that are taken into consideration, which means that the cluster must meet the requirements. Employs a similar technique, but the merging and deletion are not governed by rules, and if the cluster size is greater than 1, they can be carried out wherever. The ontology-based semantic clustering must be done before using the clustering function, which does hierarchical clustering. A semantic function carries out this action.

III. Application and Assessment

The dataset is first uploaded using this manner. There will be a range of properties in this dataset; often, cluster attributes are not taken into account. As a result, we can use the clustering criteria in a way that restricts the clustering to taking these desired traits into account. In this part, the same Delta Visitors dataset is subjected to an ontology-based semantic clustering. Since we attempted to get data from Delta Natural Park for this study, our perception of the dataset is gathered and produced using the statistics that are currently available in the literature as well as numerous other sites[13]. All of the evaluation procedures demonstrate the correct development in semantic clustering, despite some likely differences in the original dataset. The two steps of estimation are as follows: First, a rule is applied to evaluate the simple clustering. Ontology-based clustering is used for the evaluation. This group of similar features includes six variables to model the travel profile (plan, first and second arrival, lodging, duration of stay, interests, or loyalty) and four variables to describe the tourist profile (origin, age, social class, and related entities).

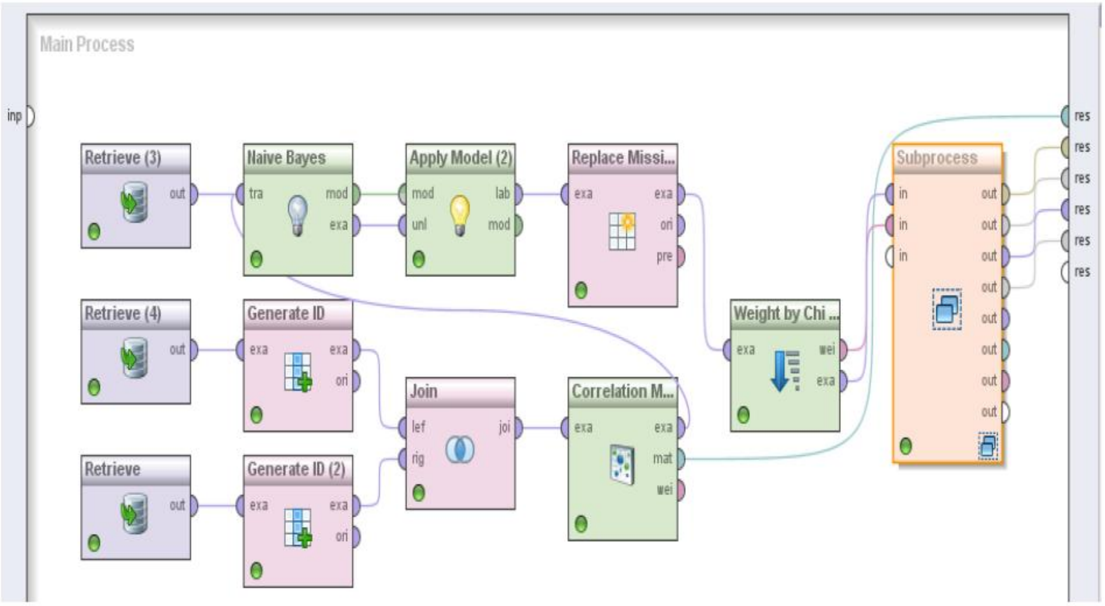

Fig 1: Clustering with applying rule

Here, the original dataset is coupled with previously unconsidered attributes of four places of worship, office buildings, the beach, and the gastronomy. We employ the principles that result in the contribution of features that were not taken into account during cluster creation in order to enhance clustering. As a result, we employ the Apply Model, which states that in confidence and prediction rules, clustering should be done based on the values of these functions and that numerous features should be joined together rather than just one feature. In the sub-procedure, data can be loaded for clustering operations, and conventional clustering activities are carried out. Lastly, the outlier is calculated and validation and similarity are completed.

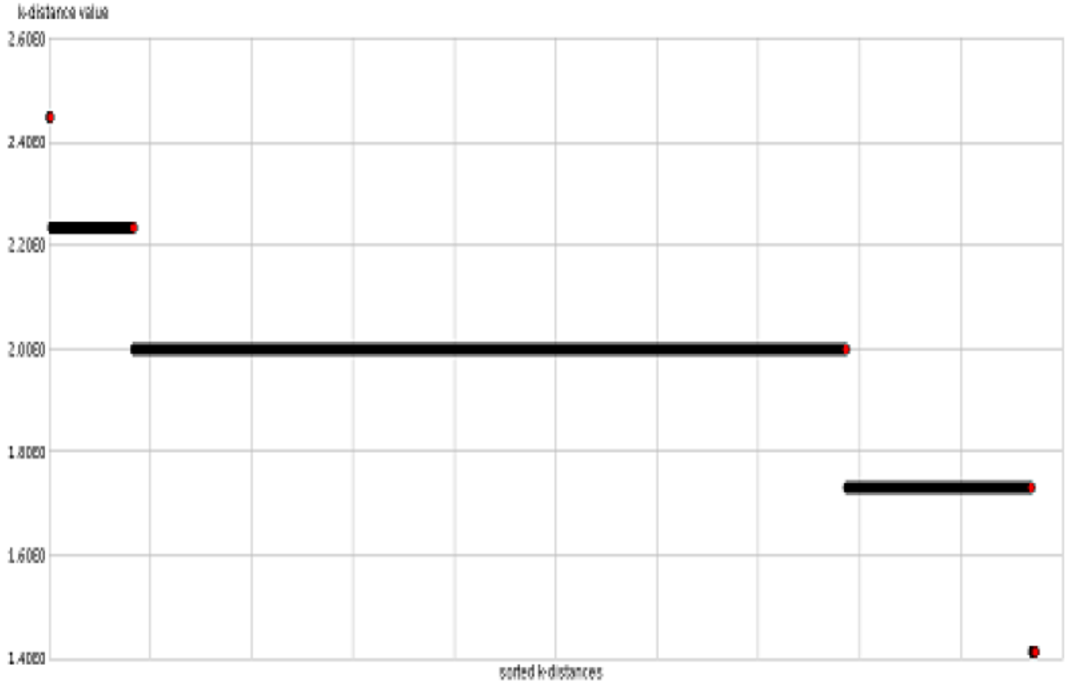

Fig 2: k-distance for a simple clustering

The average distance after normalizing k-distance is approximately 0.511, which is a significant distance based on the data in the articles. Consequently, the components have little association with one another. We now begin the process after becoming acquainted with how the clustering process works. The method will have an extremely low accuracy of roughly 26/36%. A tree model validation evaluates the duration of their stay based on age, whether they are travelling alone or with friends, whether they are local or not, and how much lodging they are using. According to this assessment, the majority of tourists who stay for less than two days typically book rooms in hotels and campers, are middle-class or upper-class strangers, and use these accommodations more often than others. There are roughly 0.155 outliers, which could be reduced. Nevertheless, the values that the Apply Model enters are used to create the correlation matrix. Because the applied rules increase the clusters' strength, the profile that is produced here is more comprehensive than one that would be acquired from basic clustering.

IV. Evaluation of Clustering Based on Ontologies

The ontology needs to be loaded at this step of the process. Therefore, we first describe the construction of the ontology that is being used. Ontology is used to learn about certain areas of interest. The concepts in that field are described by ontology, which also preserves the connections between them. In this work, the ontology is constructed using the protégé application software. We wish to designate two significant columns in this dataset as the classes in the ontology. The sort of lodging and the purpose of the trip are listed in these columns.

The simple subclass of the ontology's classifications includes lodging, camps, hotels, and cottages, whereas the fixed subclass includes houses, rental homes, and apartments. Travel, biking, fishing, eating, photography, and gastronomy fall under the entertainment subclass, while sports fall under the unknown subclass. The environment subclass includes beaches, mountains, flora, landscapes, and nature, while the culture, family, holidays, sightseeing, and relaxation subclass includes culture, family, and leisure. The principle subclass includes trade, ecology, education, drills, loyalty, and lodging, while the random subclass includes offer, chance, and visit.

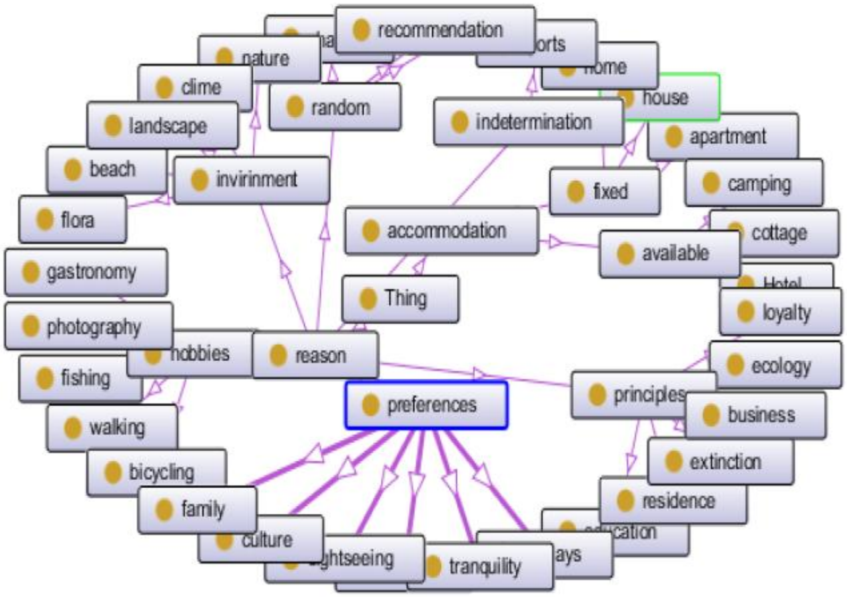

Fig 3: Ontograph

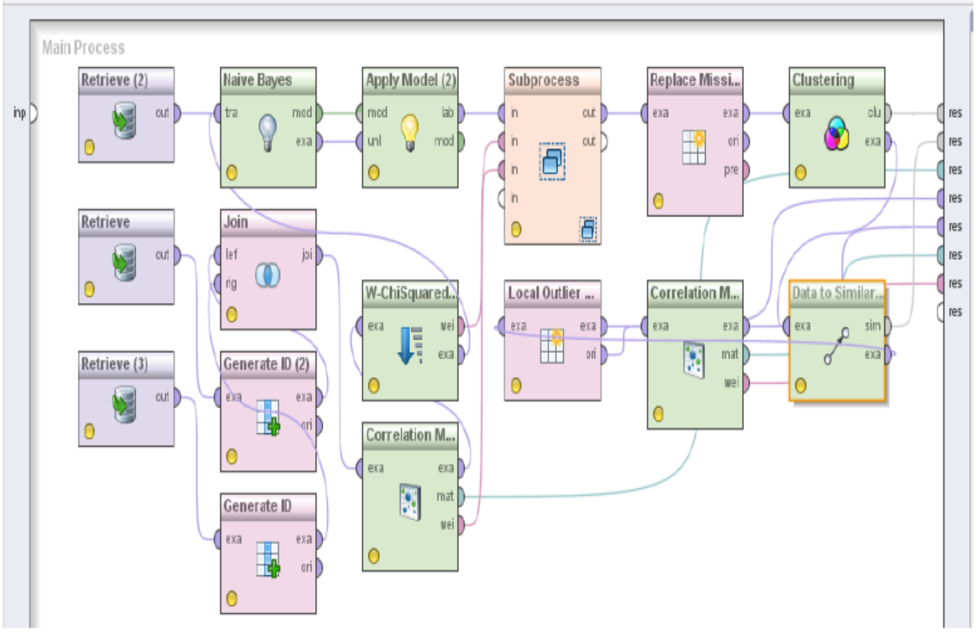

Fig 4: Ontology-based clustering with applying rule

The average distance after normalizing the k-distance is roughly 246/0, which is a significant decrease. There is a noticeable rise in correlation between the matrix's components. The original dataset is coupled with four houses, a beach, official buildings, places of worship, and gastronomy that were previously unconsidered. The Apply Model and cluster rules are used to enhance the clustering and add additional two-dimensional attributes to the clusters[13]. Cluster formation is facilitated by these rules, which result in the properties that have been examined. The dataset can be loaded into the "join" in this process, after which the matrix is formed and the weighting is done using chi-Square. All of these can therefore be fed into "Naïve Bayes," and the output and the columns on which we wish to apply the rule are sent to the "Apply Model," so loading the sub-procedure's input. Following the loading of the ontology in the sub-procedure, the input is clustered and evaluated. The dendrogram is the same as the previous way; the only difference is that clustering has a matrix of properties, and the rule is applied to the clusters, giving visitors access to a more comprehensive profile. The k-distance and outliers are identical. Its correlation matrix is altered, nevertheless, providing more details about the applied rule and its relationship.

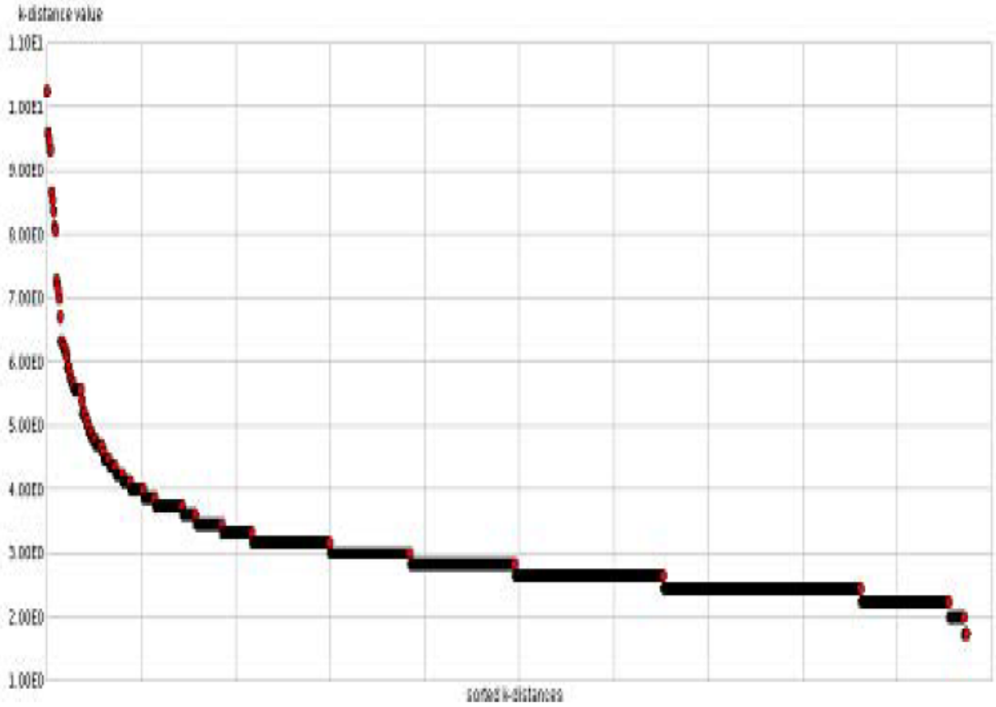

Fig 5: k-distance for ontology-based clustering

V. Conclusion

We created a dataset using three distinct approaches to gather information about tourists. There are numerical, category, and semantic features in this dataset. There are a number of aspects that enable us to extend our activities within the semantic environment. Outliers and the k-distance in the model are decreased when comparing the findings of this model with those of previous conventional models. The "apply model" and "correlation" tools are used to assess the outcomes. Only a subset of the data within the clusters is same in the earlier methods of clustering [1, 12], and clustering is carried out according to this[13]. Clusters have more consistent data in this way. When compared to other models, the quantitative findings of this ontology-based model demonstrate that in this case study, the k-distance is 0.128, which is eliminated, and correlation is increased to 75%. The application of rule enforcement also reduces outliers (0.119). The Delta Natural Park dataset served as the basis for this study. Nonetheless, it can be carried out over Iran's tourist destinations, and the growth of tourism in these regions can be calculated.

For instance, in order to generate a dataset based on the tourist profile, visitors to Ganjnameh, a historical site in Hamadan City, Iran, may be invited to complete a thorough questionnaire. The primary motivation for coming here, the degree of satisfaction, and other elements can be looked at among these responses. Another idea is to alter the ontology to alter the clustering to take into account more comprehensive visitor profiles or to specify additional rules that would allow homogeneous clusters to be combined and heterogeneous clusters to be broken.

References

[1] Han J and M Kamber. *Data Mining: Concepts and Techniques*. Morgan, Kaufmann. 2000.

[2] [2] Mirkin B. *Clustering for data mining: a data recovery approach*. London, Chapman & Hall/CRC. 2005.

[3] [3] Cao L, PS Yu, C Zhang and Y Zhao. *Domain Driven Data Mining*. Springer. 2010.

[4] [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] S, Samuel Johnson. "Design a Smart Active Filter for Solar Power System Using V2G." Journal of Science Engineering Technology and Management Sciences, vol. 2, no. 5, Apr. 2025, pp. 12–19. Crossref, https://doi.org/10.63590/jsetms.2025.v02.i05.pp12-19.

[14] 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.