ISSN:0975-9646

A Methodology for Incorporation of Domain Ontology in Knowledge Discovery Process for Interpretation and Improvement of Mining Results

Joachim Narzary[1]
Dept. Of Computer Science
M.Tech, CSE
Puducherry, India

Dr. S. Siva Sathya[2]
Dept. Of Computer Science Associate
Professor Puducherry, India

Minisrang Basumatary[3]
Dept. Of Computer Science
M.Tech, CSE
Puducherry, India

***Abstract*— We live in a world where vast amount of data are collected in every hour of the day. Thus, this era is usually called as data or information age. The analysis of this vast collection of data is of high necessity and utmost important for various decision making process. This necessity has led to the birth of data mining. Data Mining is the process of extracting potentially useful knowledge from raw data. However, the application of machine-understandable knowledge in data mining is not very prevalent and has been recognized as a gap in current traditional data mining practice. Ontology based data mining can play an important role in solving this issue of knowledge discovery. Millions of Indian schools are facing shortage of developmental materials and resources due to the fact that the Government is unaware of the school conditions. This paper aims to incorporate domain knowledge in the data mining process which can help the school authority to identify the urgent necessity of the various schools and make certain policies accordingly. Here, we considered the School Dataset of Assam as an input for the analysis of our methodology.**

***Keywords*— Data Mining, Domain Ontology, Simple K- Means, Cluster Analysis, Taxonomy Similarity, Classification**

I. INTRODUCTION

Ontology is a branch of artificial intelligence that represents the formal concepts of a particular domain and relationships amongst those concepts. It is a formal naming and definition of the types, properties, and interrelationship of the entities that fundamentally exist for a particular domain of discourse [20]. Its basic components are concepts, relationships, instances and the rules and axioms that constrain their interpretations [1]. Concept is a class of entities in a domain, like organ is a concept in the medicine domain. Relationships represent the interactions between the concepts or their properties (For example, disease p_ affects organ, here disease and organ are concepts and p_affects represents a relationship). There are two types of Relationships: taxonomy relationship and associative relationship. Taxonomy is a type of relationship that organize concepts into a hierarchical concept tree (For example, sickle_cell is a type of disease_linked_to_genes) whereas associative is a type of relationship that relate the concepts across the tree structure (Example, p_affects). Instances are instantiations of concepts which make up domain knowledge along with the taxonomies and relationships. For example, an instance can be represented as (sickle_cell p_affects spleen). Axioms are used to constrain the values for classes or instances (For example, a disease can affect number of organs and that particular organ can be affected by number of diseases). According to [11], there are two kinds of knowledge that are involved in a knowledge discovery process: data mining knowledge and domain knowledge. Data mining knowledge includes the knowledge about data mining algorithms, how they can be used, parameters tuning, and formats of input data and so on. Domain knowledge refers to the details of the dataset, how the attributes are related, their limits, etc, known as causal relations and so on [11]. The use of ontology has become increasingly popular, because of its ability to allow inferences and provide domains which are understandable by both human and application systems [11].

The following sections II, III, IV include the motivation, literature survey and proposed methodology respectively and finally the conclusion of the paper and future work is in section V.

II. MOTIVATION

Data Mining has become inevitable for variety of areas like Business, Government, Education, Learning Systems, Retrieval Systems, and Scientific Research etc. Well, different people and organizations have used various techniques to meet the purpose. However, those techniques do not consider the background knowledge of the data and therefore much of the knowledge remains undiscovered. The use of domain ontology in the mining process is the key to overcome this limitation of knowledge discovery. Ontology is referred to as the specification of a concept i.e. the description of the concepts and the relationships which exists for a domain or a community of domains. This methodology helps in interpretation of results of the traditional data mining algorithms. The work here focuses on the incorporation of

background knowledge of the data providing appropriate knowledge for the miner.

III. LITERATURE SURVEY

Ontology play an important role in data mining as it helps in quick discovery of knowledge and provides flexible inferences. Following are the three interesting reasons for which the ontology has been acquainted with knowledge discovery process [1]:

• To fill the semantic gap between the information, mining calculations, applications, and mining results.

• To utilize the former knowledge for the new or updated discovery of the knowledge and this lessens the time and cognitive task of the user.

• To provide a formal approach for the knowledge discovery process from information pre processing to mining results.

Several approaches have been developed for integrating user knowledge to solve the problem by representing existing domain knowledge using ontologies. Mor Peleg et.al in [11] proposed a method "Onto-clust" that focuses on the identification of groups of comorbidities for developmental disorders of children. The main objective of the proposed method is to identify systematically the developmental disorder groups and represent them in ontology. The developed methodology consists of two methods: (i) Patient data based on ontology which is a literature based on developmental disorder groups, and (ii) Identify and represent relationships between the developmental disorders by incorporating domain ontology using the patient clustered data.

In [14] Thangamani et.al proposed a technique to use ontology for distributed hierarchical clustering. The method is used for fuzzy document clustering and could address the problem of modularity, flexibility, and scalability.

Geo_ Macintyre et.al in [9] presented a methodology to identify gene co-expression with the help of the gene ontology in the clustering analysis. The algorithm could help in discovering more meaningful clustering results.

In [16] Abdelrahman Elsayed et.al proposed a technique using ontology for clustering documents. They introduced a distributed implementation using Map Reduction model of programming to bisect simple k-means clustering methodology result.

Andreas Hotho in [17] introduced a method for text document clustering that uses background knowledge in the process. At first the input data is pre processed applying a heuristics method for attribute selection based on the ontology. They have constructed a number of alternative text representations. Using Simple K Means various clustering results have been computed with respect to those representations. After that the result interpretation is done and explained by the selection of concepts in the ontology corresponding.

In [3] Hondjack et.al proposed a methodology "OntoDB" (ontology-based databases) to store the data and ontology in the same database data. For high performance of the proposed method they have used two representation schemes vertical and binary representations with a variant called hybrid.

Alexander et.al in [10] proposed an approach for clustering ontology-based metadata. The technique is considered for a set of similarity measures that allow computing similarities between ontology-based metadata along various dimensionalities. The similarity measures can be taken as the input of the hierarchical clustering algorithm. The very important contributions of this paper are the set of similarity measures defined to analyse study of an ontology application within a taxonomical clustering algorithm.

In [4] A. Hotho proposed a new approach for applying background knowledge during pre processing in order to improve the results of the clustering algorithm and to allow best selection of the results. They built variety of views based on the taxonomy text features on a hierarchy of concepts. Based on these aggregations, multiple clustering results are computed using K-Means.

Travis D in [2] proposed a method for clustering hierarchical information using ontology languages. The authors used the meaning features given in ontology languages and the methods such as keyword search to visualize and analysis the document clustering results at conceptually higher levels.

In [8] Paweł Lula presented a framework for ontology based cluster analysis. The framework concentrates to use various similarity calculations in the ontology environment between different objects. The framework is based on the ontology specification of two varieties of components: categories and objects descriptions.

Liping Jing et.al in [5] proposed a new scheme for clustering text that utilizes ontology and the distance measure. Before implementing clustering process, with the use of the Wordnet the mutual term matrix of information is calculated and some techniques of learning ontologies from textual data are considered. Traditional vector space model and mutual matrix information are combined, and then they design a new model of data for which the Euclidean distance measure may be used, and then run two k-means type clustering algorithms on the real-world text data.

Nadana Ravishankar et.al in [13] presented a clustering algorithm which is developed to improve the clustering results and visualization is based on decision tree, such that only the relevant results are given to the user.

Wei Wang et.al in [7] proposed a framework for Ontology Driven Subspace Clustering. They used the domain ontology to prune the unwanted rules that is obtained from the association rule mining. Further, the algorithm generates automatic interpretation of the clustering result by the relationships between the hierarchically organized clusters with efficient taxonomical improvement onto the ontology hierarchy.

Claudia Marinica et.al in [18] proposed an approach to prune and filter discovered rules. The proposed methodology is composed of three main parts: At first the authors used a basic mining process to extract a set of association rules. Secondly, they use a knowledge base that allows formalizing user knowledge and goals. Finally, several operators (i.e.

pruning) are applied in the post-processing step in order to extract the interesting rules.

Neves et.al in [19] proposed a methodology called SemPrune, which is built on domain ontology and it is intended for pre- and post-processing phases of data mining. The adopted technique filters the rules obtained from association rule mining with the use of generalization and specialization method and further, the coverage interest measure and confidence measure (CRm) indicators are considered.

IV. PROPOSED METHODOLOGY

This paper proposes a new approach to define a formal environment to incorporate background knowledge into the data mining process. It aims to cluster similar items using Simple K-Means algorithm and we reconstruct the cluster using split and join operation with the help of ontology. Further we classify them taxonomically with respect to similarity parameters. Finally, we find relevant patterns in each cluster using association rule mining. The proposed methodology shown in figure (i) composed of six different activities and discussed in the following section:

A. Data Pre processing and Feature Reduction

The dataset may contain numerous information for which we are not interested and that will not benefit our purpose. Those unwanted attributes should be removed. This can be done using different algorithms or manually for small set of data. The main objective of the feature reduction is to choose a subset of attributes and information of a particular dataset that will help to achieve the purpose and remove those attributes which will have little or no significant information. Data preparation deals with the understanding of the data and this in turn helps to use the right data miner algorithm. Therefore, the data preparation allows us to find faster and better mining results.

In our discussion we considered the school dataset of Assam which contains numerous information viz. Enrolment statistics, Teachers Information, Facility Details and Location Information. This paper deals with only the Facility Details of the school and therefore there are lot of pre processing to be done on the dataset. Following are the school features that we considered for our purpose.

BLDSTATUS, CLROOMS, TOILETB, TOILETG, MEALSINSCH, CAL_YN, ELECTRIC_YN, BNDRYWALL, LIBRARY_YN, PGROUND_YN, WATER, MEDCHK_YN, COMPUTER.

Finally, we converted the dataset in ARFF format which is acceptable in Weka Tool.

B. Development of School Ontology

Ontology is a branch of artificial intelligence that represents the formal concepts of a particular domain and relationships amongst those concepts. Its basic components are concepts, relationships, instances and the rules and axioms that constrain their interpretations [1]. Ontology can be developed using a tool called protégé which is a user friendly application. Protégé is an open source editor and framework for building intelligent systems. Protégé's plug in architecture can be used to build both complex and simple ontology based applications. Developers can use the output of protégé with other problem solver systems to construct a wide range of intelligent systems.

We used Protege Tool to develop the school ontology; fig (ii) shows a fragment of our ontology.

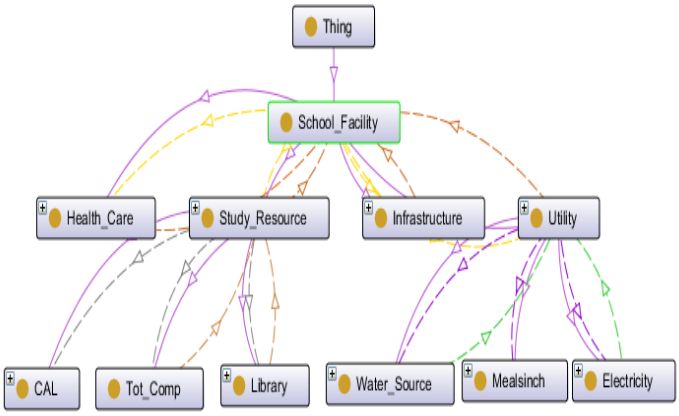

Fig (ii): A Fragment of School Ontology

C. Applying Simple K Means Clustering Algorithm

Cluster analysis refers to the grouping of dataset into different clusters or groups based on the similarity of the objects. The clustering is done such that (i) there is high similarity between the objects of same cluster (high intra cluster similarity) and (ii) low inter-cluster similarity. Simple K-Means is a clustering algorithm which can be used to cluster the data based on the feature or the attributes of the data. Given below is the pseudo code of the simple k means clustering [21]:

i. Randomly select "cn" cluster centres
ii. Calculate the distance between each data point and cluster centres using Euclidean Distance Formula

$$\text{Ed} = \sqrt{\sum_{i=1}^{n}(q_i - p_i)^2}.$$

iii. Assign the data point to that particular cluster whose distance from the centre of the cluster is minimum of all the cluster centres
iv. Recalculate the new cluster centres

$$v_i = (1/c_i) \sum_{j=1}^{c_i} x_i$$

Where "ci" represents the number of data points in ith cluster
v. Recalculate the distance between each data point and the new cluster centres obtained
vi. If no data point was reassigned then stop, otherwise repeat from step (iii)

We used the Weka Tool to cluster the school dataset using the Simple K Means Algorithm. The obtained result from the clustering algorithm is shown in figure (iii) .The number of clusters is set to four.

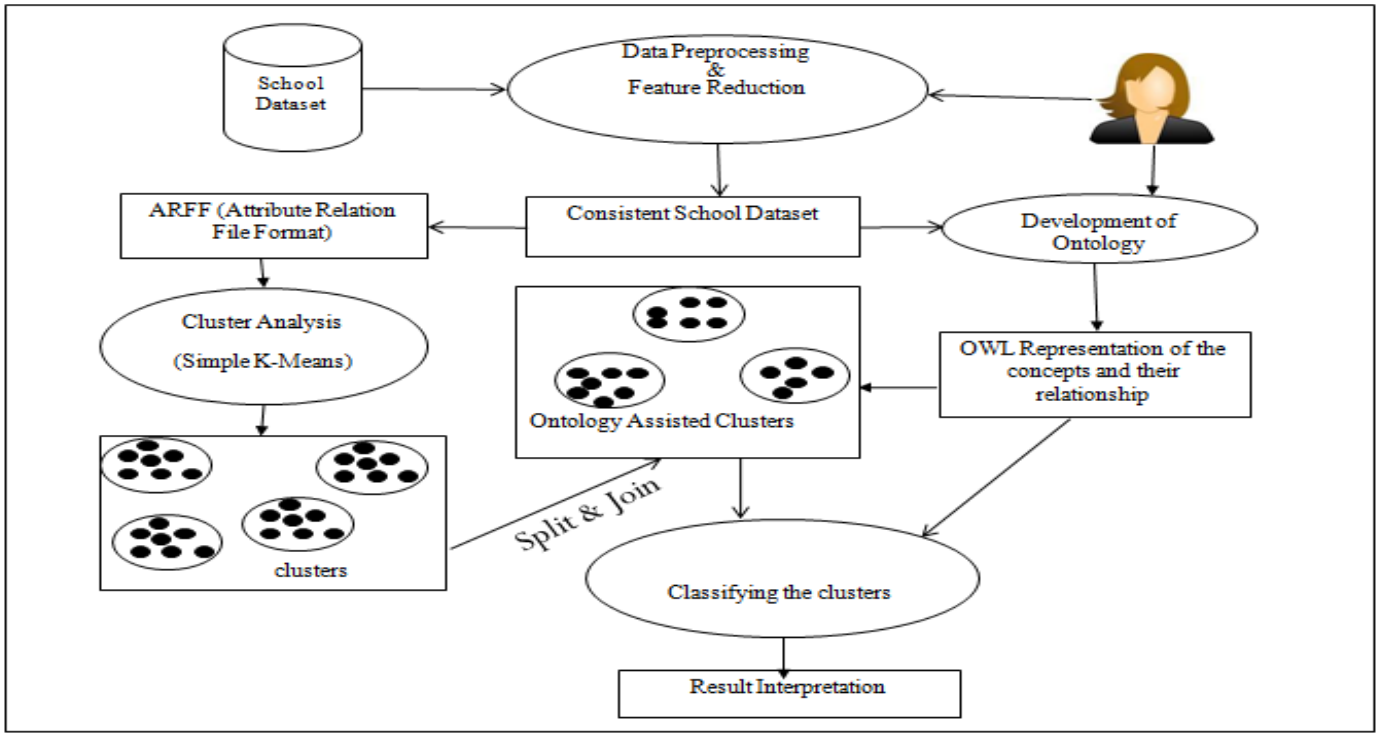

Fig (i): Ontology Assisted clustering framework

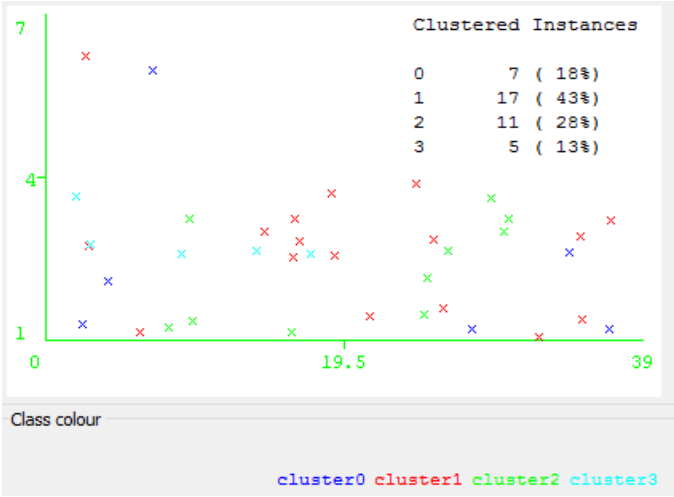

Fig (iii): Visualization of Cluster Result

D. Reconstruction of Cluster Using Ontology

After obtaining the clustering result we apply split and join operations using the relationship defined in the ontology. This helps to improve the clustering result [6]. Fig (iv), illustrates the split and join operations.

E. Classification of the Clusters

Once we obtain the precise clustering results, the next task is to classify the clusters with respect to the four parameters of the school ontology. The four parameters included in the school ontology are; Health Care, Infrastructure, Study Resource and Utility. Based on the taxonomy similarity we classify the clusters accordingly. The algorithm for the taxonomy similarity classification is given below:

i. Find the taxonomy concepts of each parameter T^c (as shown in D.I)
ii. Assign T^c weight to each parameter (as in Sec. D.I)
iii. Select an instance from a cluster Ci and find the taxonomy concepts (T^c) for it
   a. Find the taxonomy similarity of Ci with respect to all the parameters separately
   b. Continue (a) for all instances in Ci
   c. Calculate the average taxonomy similarity of all the instances of Ci
iv. Repeat (iii) for all clusters
v. Label the cluster with respect to the parameters taxonomy similarity values
vi. Stop

D.I. Analysis of the Taxonomy Similarity Classification Algorithm

1. Finding Taxonomy Concepts of each Parameter T^c

Let us consider the fragment of ontology shown in the fig (ii), as it is mentioned there are four parameters and the taxonomy concepts are given as:

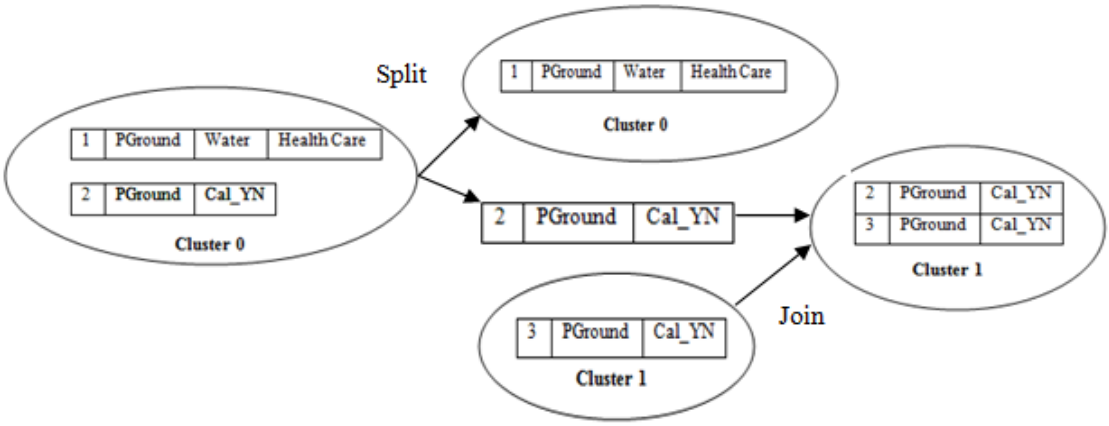

Fig (iv): In cluster0 the instance 2 is more similar with the instance in cluster1 than the instance of its own cluster. Therefore, the instances are divided into two halves and the instance2 is joined with the cluster1

T^c (Study Resource): {Cal, Tot_Comp, Library, School_Facility}

T^c (Utility): {Water_Source, Mealsinch, Electricity, Utility, School_Facility}

Likewise, we find the taxonomy concepts (T^c) of all the parameters.

2. Calculation of T^c Weight for each Parameter

The weight is calculated according to the concepts under each parameter. As it is mentioned that the ontology given in fig (ii) is partial therefore the explanation is just to understand the working behaviour. The T^c weight for the given parameters will be as follows:

Weight of Study_Resource will be 5 as it has four concepts related to it viz. {Cal, Tot_Comp, Library, School_Facility}, while Health_Care has 2 as its weight value.

3. Finding Taxonomy Similarity

Taxonomy Similarity can be calculated using the concept of intersection between concepts. The taxonomy similarity of Study_Resource with respect to Utility is given by:

Study_Resource (T^c) ∩Utility (T^c)

{Cal, Tot_Comp, Library, School_Facility} ∩ {Water_Source, Mealsinch, Electricity, Utility, School_Facility} =1

4. Label the cluster with respect to the parameters of taxonomy similarity values

Labeling is done according to the average taxonomy similarity with respect to the parameters concept weights. For example if the average taxonomy similarity of four clusters are 4, 5, 6, 7 respectively with respect to parameters and the parameters weight are given as 2, 5, 3, 6 then the similarity with respect to parameter weight will be calculated as follows: Normalized weight for parameters will be 15,6,10 and 5 respectively. Similarity of cluster0 with respect to four Parameters are:

Parameter 1: 15/30*4 =2
Parameter 2: 6/30*4=0.8
Parameter 3: 10/30*4=1.3
Parameter 4: 5/30*4=0.6

Therefore, cluster0 is more similar with parameter 1 and 3, 2, 4 respectively.

F. Finding Relevant Patterns

Finding relevant patterns using association rule mining is the final step of our methodology. Here we find the rules in each cluster separately.

We used weka tool for applying the Apriori algorithm. Following are the sample of rules obtained from the cluster0:

```
1. PGROUND_YN=1 16 ==> LIBRARY_YN=2 16    conf:(1)
2. LIBRARY_YN=2 16 ==> PGROUND_YN=1 16    conf:(1)
3. CAL_YN=2 15 ==> LIBRARY_YN=2 15    conf:(1)
4. CAL_YN=2 15 ==> PGROUND_YN=1 15    conf:(1)
5. CAL_YN=2 PGROUND_YN=1 15 ==> LIBRARY_YN=2 15    conf:(1)
6. CAL_YN=2 LIBRARY_YN=2 15 ==> PGROUND_YN=1 15    conf:(1)
7. CAL_YN=2 15 ==> LIBRARY_YN=2 PGROUND_YN=1 15    conf:(1)
8. LIBRARY_YN=2 16 ==> CAL_YN=2 15    conf:(0.94)
9. PGROUND_YN=1 16 ==> CAL_YN=2 15    conf:(0.94)
10. LIBRARY_YN=2 PGROUND_YN=1 16 ==> CAL_YN=2 15    conf:(0.94)
```

V. CONCLUSION

There is a growing necessity for the use of ontology or background knowledge in data mining process. Utilizing Ontologies scaffold the semantic hole between the information, applications, data mining calculations, and data mining results.

This paper presents a formal methodology for incorporation of background knowledge in the data mining process. The model aims to interpret and improve the mining results. Automating the methodology remains as the future work of this paper.

REFERENCES

[1] Bhagat, P. P. V. (2015). A Survey Paper on Ontology-Based Approaches for Semantic Data Mining, (April).

[2] Breaux, T. D., Carolina, N., & Reed, J. W. (2005).Hierarchical Information Clustering Using Ontology Languages.Proceedings of the 38 Th Hawaii International Conference on System Sciences (HICSS-38), 112b.

[3] Dehainsala, H., Pierra, G., &Bellatreche, L. (2007).Proc .of Database Systems for Advanced Applications (DASFAA ' 2007), OntoDB : An Ontology-Based Database for Data Intensive Applications, 02706.

[4] Hotho, A., &Staab, S. (n.d.). Ontology-based Text Clustering University of Karlsruhe The reference function with.

[5] Jing, L., Zhou, L., Ng, M. K. M. K., & Huang, J. Z. Z. (2006).Ontology-based distance measure for texclustering.Proceedings of SIAM SDM Workshop on Text Mining, Bethesda, Maryland, USA. Retrieved fromhttp://scholar.google.com/scholar?hl=en&btnG=Search&q=intitle:Ontologybased+Distance+Measure+for+Text+Clustering#0

[6] Kazi, A. (2015). An Ontology Based Approach to Data Mining An Ontology Based Approach to Data Mining, 1–5.

[7] Liu, J., Wang, W., & Yang, J. (2004).A framework for ontology-driven subspace clustering.Proceedings of the 2004 ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '04, 623. http://doi.org/10.1145/1014052.1014130

[8] Lula, P., &Paliwoda-Pękosz, G. (2008).An ontology-based cluster analysis framework.Proceedings of the First International Workshop on Ontology-Supported Business Intelligence - OBI '08, 1–6. http://doi.org/10.1145/1452567.1452574

[9] Macintyre, G., & Bailey, J. (2008). Gene ontology assisted exploratory microarray clustering and its application to cancer. Pattern Recognition in …, 1–12. Retrieved from http://link.springer.com/chapter/10.1007/978-3-540-88436-1_34

[10] Maedche, A., & Zacharias, V. (2002).Clustering Ontology-Based Metadata in the Semantic Web.Principles of Data Mining and Knowledge Discovery, 2431, 383–408. http://doi.org/10.1007/3-540-45681-3_29

[11] Peleg, M., Asbeh, N., Kuflik, T., & Schertz, M. (2009). Onto-clust—A methodology for combining clustering analysis and ontological methods for identifying groups of comorbidities for developmental disorders. Journal of Biomedical Informatics, 42(1), 165–175. http://doi.org/10.1016/j.jbi.2008.05.010

[12] Pyle, D. (1999). Data preparation for data mining.Applied Artificial Intelligence, 17(5), 375–381. http://doi.org/10.1080/08839510390219264

[13] Ravishankar, N. T., &Shriram, R. (2013). Ontology based Clustering Algorithm for Information Retrieval. 2013 Fourth International Conference on Computing, Communications and Networking Technologies (Icccnt), 2–5.

[14] Thangamani, M. (2010). Ontology Based Fuzzy Document Clustering Scheme, 4(7), 148–156.

[15] Zagoruiko, N. G., Gulyaevskii, S. E., &Kovalerchuk, B. Y. (2007).Ontology of the Data Mining Subject Domain.Pattern Recognition and Image Analysis, 17(3), 349–356. http://doi.org/10.1134/S1054661807030017

[16] Abdelrahman Elsayed 1, Hoda M. O. Mokhtar 2 and Osama Ismail3. ONTOLOGY BASED DOCUMENT CLUSTERING USING MAPREDUCE. International Journal of Database Management Systems (IJDMS) Vol.7, No.2, April 20

[17] Andreas Hotho and Alexander Maedche and Steffen Staab Institute AIFB, University of Karlsruhe, 76128 Karlsruhe, Germany {aho, ama, sst}@aifb.uni-karlsruhe.de http://www.aifb.uni-karlsruhe.de/WBS Ontology-based Text Document Clustering 15

[18] Ferraz, I. N., & Garcia, A. C. B. (2013). Ontology in association rules. SpringerPlus, 2(2006), 452. http://doi.org/10.1186/2193-1801-2-452

[19] Mansingh, G., Osei-Bryson, K. M., & Reichgelt, H. (2011). Using ontologies to facilitate post-processing of association rules by domain experts. Information Sciences, 181(3), 419–434. http://doi.org/10.1016/j.ins.2010.09.027

[20] https://en.wikipedia.org/wiki/Ontology_(information_science)

[21] https://sites.google.com/site/dataclusteringalgorithms/k-means-clustering-algorithm