Ontological Structure on Concept Clustering through Data Mining Techniques

P Anupama[1], Prof B Prajna[2], 1, 2 Department of Computer Science and Systems Engineering,
Andhra University College of Engineering, Andhra University, AP, INDIA

***Abstract*—Ontology is formal explicit specification of conceptualization of a domain that offers a platform for the sharing and reuse of knowledge across heterogeneous platforms. The technique of ontology finds its application in almost every area, some of which includes medicine, e-commerce, chemistry, education etc. Concept clustering is the foremost step in construction of ontology. Concept clustering is usually a manual process involves labor and time intensive task. Hence there is a need for automatic grouping of concepts for ontology construction. In this paper, automatic concept clustering is attempted through data mining clustering techniques. The clustering mechanisms should be fair enough to perform query executions and must also be responsible in categorizing things without any conflicts.**

In the proposed system we build an Ontology structure based on data integration of different inputs incurred by using clustering mechanisms which are used to generate patterns for upcoming evaluation when needed. The training set for the concepts formation of ontology structure is obtained from zoo dataset in UCI Machine Learning Repository. The clustering techniques are implemented through R 3.5.1, an open source data mining tool. Performance of clustering techniques viz., EM, Farthest First, K-Means, Density Based and Hierarchical are analyzed to yield best accuracy.

***Index Terms*—Clustering, Data Mining, Ontology, Concept Formation, Expectation Maximization (EM), Farthest First, K-Means, Density Based and Hierarchical clustering**

I. INTRODUCTION

Data mining is the process of discovering interesting knowledge, such as patterns, significant structures, from large amount of data stored in databases and other information repositories. Data mining techniques have always been a powerful tool for any type of categorization.Ontology is a technique for expressing formal specification. It is a conceptualization of domain and its terms and relationships. The technique represents entities, ideas and events, along with their properties and relations, as a form of knowledge representation. Though ontology has many-fold applications, construction of ontology structure is a complex and laborious task [2]. It includes domain and scope identification, Representation, FunctionTerm Identification, Hierarchy definition, classes definition, reasoner, attributes definition, inter relationship identification and annotation. Function term identification, Hierarchy definition, class definition, attribute definition and inter-relationship definition forms the main core of conceptualization in ontology building.

Data mining is the process of finding anomalies, patterns and correlations within large data sets to predict outcomes. Using a broad range of techniques, you can use this information to increase revenues, cut costs, improve customer relationships, and reduce risks and more. With the enormous amount of data stored in files, databases, and other repositories, it is increasingly important, if not necessary, to develop powerful means for analysis and perhaps interpretation of such data and for the extraction of interesting knowledge that could help in decision-making. Data mining is defined as a process used to extract usable data from a larger set of any raw data. It implies analysing data patterns in large batches of data using one or more software. Data mining has applications in multiple fields, like science and research. As an application of data mining, businesses can learn more about their customers and develop more effective strategies related to various business functions and in turn leverage resources in a more optimal and insightful manner. This helps businesses be closer to their objective and make better decisions. Data mining involves effective data collection and warehousing as well as computer processing. For segmenting the data and evaluating the probability of future events, data mining uses sophisticated mathematical algorithms. Data mining is also known as Knowledge Discovery in Data (KDD).

Clustering is a process which partitions a given data set into homogeneous groups based on given features such that similar objects are kept in a group whereas dissimilar objects are in different groups. It is the most important unsupervised learning problem. It deals with finding structure in a collection of unlabeled data. Clustering algorithm must be able to deal with different types of attributes. Clustering algorithm must be able to find clustered data with the arbitrary shape. Clustering algorithm must be insensitive to noise and outliers.

Concept grouping is the main basis for conceptualization. This demands advice from the domain experts. Human intervention makes it time, labor and resource intensive task. Automatic concept grouping will be of great help in constructing ontology in more efficient way. Only a few attempts have been made to automatically group the concepts, some of which are summarized below.

In this paper, attempt has been made to automatically group concepts through data mining clustering techniques. Data mining techniques [9-17] have always been a powerful tool for any type of categorization. Data mining techniques include supervised and unsupervised learning. Supervised learning requires training data with the label of the classes while unsupervised learning does not require the labels apriori. For the current research work, since the classes are not known earlier, unsupervised clustering techniques are sought for clustering techniques [18] viz. EM, Farthest First, KMeans, Density Based and Hierarchical were implemented and attempted for grouping of the concepts for the zoo ontology dataset obtained from UCI Machine Learning Repository [19].

The paper is organized as follows: Section II gives details on literature survey. Section III is on the problem description. Section IV details on the proposed methodology for automatic clustering of concepts for ontology construction. Section V presents the performance of the different clustering algorithms. Section VI concludes the paper and also gives direction on the future research in this area.

A. Challenges in Data mining:

- Algorithms must be highly scalable to handle huge datasets.
- Micro-array may have tens of thousands of dimensions.
- Poor-quality data such as: dirty data, missing values, inadequate data size, and poor representation in data sampling.
- Lack of understanding/lack of diffusion of data mining techniques in academic arenas.
- Data variety - trying to accommodate data that comes from different sources and in a variety of different forms (images, geo data, text, social, numeric, etc.).
- Data velocity - online machine learning requires models to be constantly updated with new, incoming data.

II. Literature Review

Literature survey is the basic step in preparing the new methodology for the particular area of subject. Many researchers have been made their work on the latest advancements in the technology for the improvements in the conversion of unstructured data into structured format. This conversion will help the user to take decisions by applying the queries on the structured data. Objectives of the literature survey are presented in the section:

Document Clustering [21] is used as traditional techniques for clustering the document are mostly based on the number of occurrences and the existence of keywords. COBWEB clustering algorithm was adopted by software agents to automatically generate concepts for music domain [3]. Structured knowledge was created for geneproduct using iterative statistical information extraction in combination with nearest neighbor clustering [4]. Formal Concept Analysis was used to formally abstract data as conceptual structures [5].

A further refinement to Formal Concept Analysis was made in [6] by incorporating fuzzy logic in it to deal with uncertainties in data and interpret the concept hierarchy. The fuzzy formal concept analysis was used in automatic generation of ontology for scholarly semantic web. TextOntEx constructs ontology from natural domain text using semantic pattern-based approach, and analyze natural domain text to extract candidate relations, and map them into meaning representation to facilitate ontology representation [7]. Based on the data mining outputs from rule sets and decision trees, ontologies were built automatically. RDF, RDF-S and DAML+OIL were used for defining ontologies [8].

III. Problem Description

Patterns that are generated are not able to act as a repository for the data collected through objects so there is a serious need to implement system which can create various hidden patterns. So our goal is to cluster the domain data and group that data based on attributes and instances by using different clustering techniques.

IV. Design and Methodology

The following figure 1 shows Data Mining Architecture containing six components. That is a data source, data warehouse server, data mining engine, and knowledge base. We can say it is a process of extracting interesting knowledge from large amounts of data. That is stored in many data sources. Such as file systems, databases, data warehouses. Also, knowledge used to contributes a lot of benefits to business and individual.

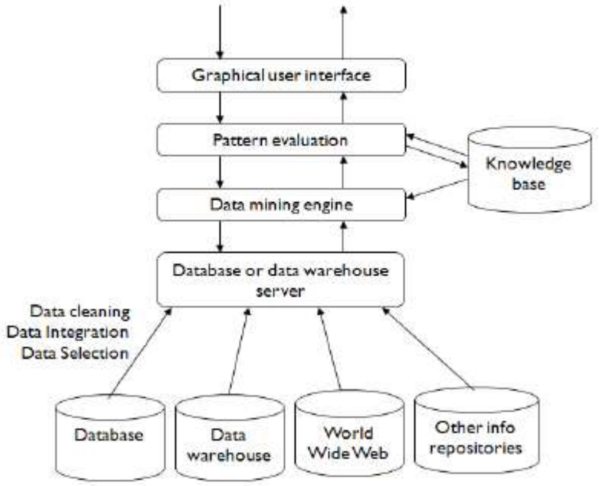

Figure 1: Data Mining Architecture

A. Major Components of Data Mining

- **Data Sources:** There are so many documents present. That is a database, data warehouse, World Wide Web (WWW). That is the actual sources of data. Sometimes, data may reside even in plain text files or spreadsheets. World Wide Web or the Internet is another big source of data.
- **Database or Data Warehouse Server:** The database server contains the actual data that is ready to be processed. Hence, the server handles retrieving the relevant data. That is based on the data mining request of the user.
  - **Data Mining Engine:** In data mining system data mining engine is the core component, as it consists a number of modules. That we used to perform data mining tasks. That includes association, classification, characterization, clustering, prediction, etc.
  - **Pattern Evaluation Modules:** This module is mainly responsible for the measure of interestingness of the pattern. For this, we use a threshold value. Also, it interacts with the data mining engine. That's main focus is to search towards interesting patterns.
  - **Graphical User Interface:** We use this interface to communicate between the user and the data mining system. Also, this module helps the user use the system easily and efficiently. They don't know the real complexity of the process. When the user specifies a query, this module interacts with the data mining system. Thus, displays the result in an easily understandable manner.
  - **Knowledge Base:** In whole data mining process, the knowledge base is beneficial. We use it to guiding the

search for the result patterns. The knowledge base might even contain user beliefs and data from user experiences. That can be useful in the process of data mining. The data mining engine might get inputs from the knowledge. That is the base to make the result more accurate and reliable. The pattern evaluation module interacts with the knowledge base. That is on a regular basis to get inputs and also to update it.

Conceptualization and hence concept formation is the most important step in construction of ontology. In this paper, automatic concept formation through clustering techniques is proposed. The proposed framework consists of domain identification, data pre-processing, Clustering, performance evaluation, cluster pattern formation and Inference engine. The proposed framework is depicted in Figure 2.

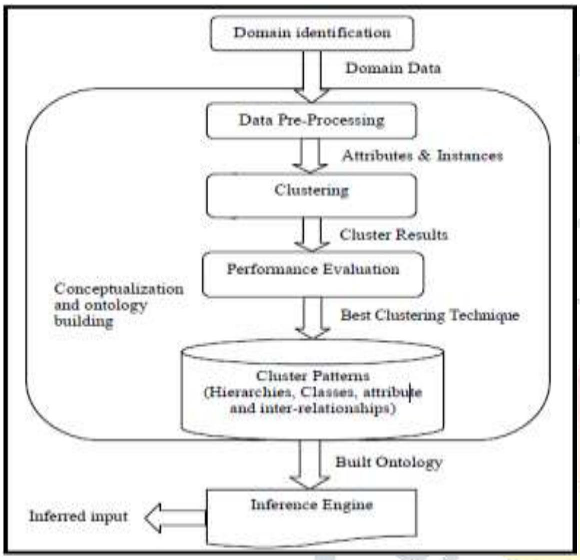

Figure 2:Automatic concept formation through clustering technique

B. *Domain Identification*

For the construction of the current ontology framework, the domain is chosen as zoo environment. The dataset used for experimentation is zoo dataset from UCI machine learning repository [19]. The attributes in the dataset include Hair (indicating its presence or absence), Feathers (indicating its presence or absence), Eggs (Giving birth through eggs or some other means), Milk (ability or inability to secrete milk), Airborne (ability or inability to cause a disease), Aquatic (lives in water or other), Predator (feeds through hunting or other means), Toothed(indicating presence or absence of tooth), Backbone (indicates presence or absence of backbone), Breathe (breathe through nose or some other means), Venomous (indicating poisonous or non-poisonous), Fins (presence or absence of fins), Legs (number of legs), Tail(indicates the presence or absence of tail), Domestic (tamed or untamed by humans) and Cat size.

C. *Data Pre-Processing*

The domain data is given as input to the data pre-processing step. Data preprocessing is a data mining technique that involves transforming raw data into an understandable format. Real-world data is often incomplete, inconsistent, and/or lacking in certain behaviors or trends, and is likely to contain many errors. Then the pre-proceesed data is now available for clustering.

D. *Clustering*

Clustering is the task of grouping a set of data in such a way that data in the same group are more similar to each other than to those in other groups. Different clustering algorithms attempt to group the data in a different way. The common clustering techniques include Expectation Maximization, K-Means, Farthest First, Density Based and Hierarchical algorithms.

1. *Expectation Maximization*:Expectation Maximization (EM) [18] algorithm is used to classify each point into the most likely Gaussian and estimate the parameters of each distribution.
2. *K-Means Clustering*:K-Means clustering [18] finds the cluster centers and assigns the objects to the nearest cluster center, such that the squared distances from the cluster are minimized. It is an optimization problem.
3. *Farthest First Clustering*:Farthest First [18] is a variant of K-Means that differs in the initial centroid assignment. It places each cluster centre in turn at the point furthest from the existing cluster centres. This point must lie within the data area. This greatly speeds up the clustering in most cases since lesser assignment and adjustment is needed.
4. *Density Based Clustering*:Density based clustering algorithm [18] has played a vital role in finding nonlinear shapes structure based on the density. Density-Based Spatial Clustering of Applications with Noise (DBSCAN) is most widely used density based algorithm. It uses the concept of density reachability and density connectivity.
5. *Hierarchical Clustering*: The Hierarchical clustering algorithm (HCA) [18] is also called as connectivity based clustering, which is mainly based on the core idea of objects that are being more relative to the nearby objects than to the objects far away. It is a method of cluster analysis which seeks to build a hierarchy of clusters. Its result is usually presented in a dendrogram. It is generally classified as Agglomerative and Divisive methods that depended upon how the hierarchies are formed.

- Agglomerative: It is a bottom up approach. It starts by placing each object in its own cluster. Then merges these minute clusters into larger and larger clusters, until all of the objects are in a single cluster or until certain termination conditions are satisfied. Its complexity is $O(n^3)$ which makes then too slow for large data sets.
- Divisive: It is a top down approach. It starting with all objects in one cluster. Then splits are performed recursively as one move down the hierarchy. Its complexity is O(2n) which is worse. These algorithms join the objects and form clusters by measuring their distance. These algorithms cannot provide a particular partitioning in the dataset, but they provide a widespread hierarchy of clusters that are merged with each other at accurate distance.

E. *Performance Evaluation*

Performance evaluation is based on class to cluster evaluation. Classes are assigned to the clusters based on the majority value of the class attribute within each cluster. Accuracy [20] is defined as the number of correctly clustered instances to the total number of instances. Performance of the three clustering techniques is evaluated using this metric.

F. Cluster Patterns

The best clustered patterns are those which have the least misclassification (grouping into the wrong clusters). The cluster patterns provide information on function terms, hierarchies, classes, attributes and inter-relationships. Hence the cluster patterns automatically conceptualize the ontology from which building of ontology can be easily done.

G. Inference Engine

Inference Engine contains facts and the best clusters. Any domain query can be solved using the inference engine. The query becomes the input and solution to it inferred by the inference engine from the built ontology forms the output. This work thus attempts to automatically group the concepts through the proposed framework. Experimental results and performance comparison of the clustering techniques are presented in the next section.

V. Results

Automatic concept clustering helps in effective ontology construction. The case study for experimentation is zoo dataset from UCI machine repository. Different clustering algorithms are evaluated using R 3.5.1, an open source tool. The instances and the associated classes are shown in Table I. In this Result Mammal, Bird, Reptile, Fish, Amphibian, Insect and Invertebrate representsType1,2,3,4,5,6 and 7 respectively.

TABLE I
INSTANCES AND ASSOCIATED CLASSES

Animals	Group
Aardvark,antelope,bear,boar,buffalo,calf,cavy,cheetah,deer,dolphin,elephant,fruitbat,giraffe,girl,goat,gorilla,hamster,hare,leopard,lion,lynx,mink,mole,mongoose,opossum,oryx,platypus,polecat,pony,porpoise,puma,pussycat,raccoon,,reindeer,seal,sealion,squirrel,vampire,vole,wallaby,wolf	Mammal
Chicken,crow,dove,duck,flamingo,gull,hawk,kiwi,lark,ostrich,parakeet,penguin,pheasant,rhea,skimmer,skua,sparrow,swan,vulture,wren	Bird
Pitviper,seasnake,slowworm,tortoise,tuatara	Reptile
Bass,carp,catfish,chub,dogfish,haddock,herring,pike,piranha,seahorse,sole,stingray,tuna	Fish
Frog,frog,newt,toad	Amphibian
Flea,gnat,honeybee,housefly,ladybird,moth,termite,Wasp	Insect
Clam,crab,crayfish,lobster,octopus,scorpion,seawasp,slug, starfish,worm	Invertebrate

The number of animals and the associated classes are shown in Figure 3.They represents of groups of mammals, birds, reptiles, fish, amphibian, insect, invertebrate of count 41, 20, 5, 13, 4, 8, 10 respectively.

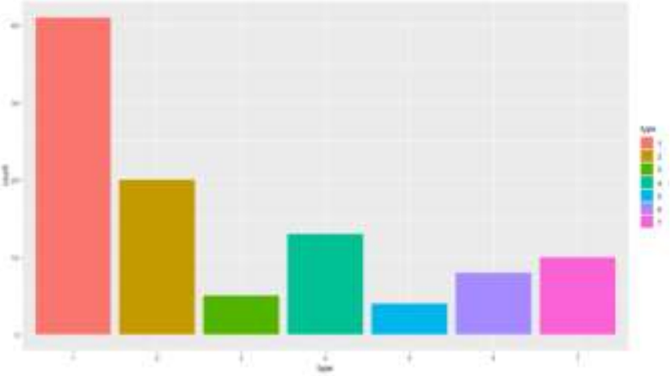

Figure 3: Number of animals for each Group

The scatter plot of attributes like animal name, hair, feathers, eggs, milk, airborne, aquatic, predator, toothed, backbone, breathes, venomous, fins, legs, tail, domestic, catsize, type of animals are depicted in Figure 4.

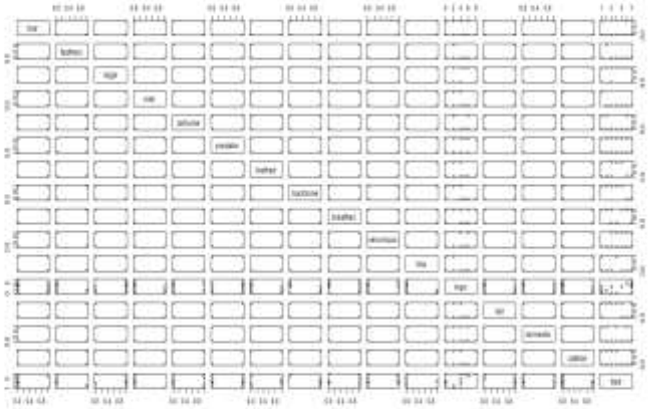

Figure 4: Scatter plot of Attributes of animals

The number of legs to type of animal is depicted in Figure 5.

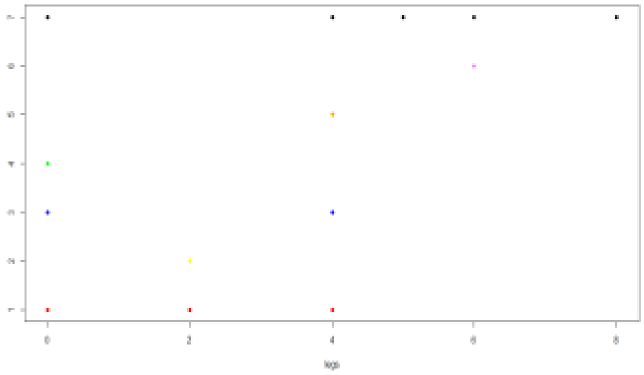

Figure 5: Number of legs to Type of animal

K-Means Clustering: We can see and notice the suitable number of clusters is 5. K-means algorithm is used and each cluster indicates to different type of animal. We circled different clusters with different colors. The cluster plot is depicted below in Figure 6.

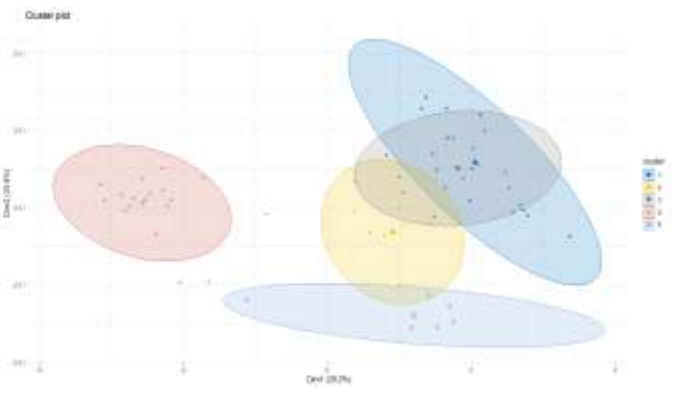

Figure 6: K-Means Cluster Plot

Hierarchical Clustering: We clustered the animals using the Group-average clustering for agglomerative clustering. We also applied the Complete-link clustering and Single-link clustering but we found that the Group-average clustering gives the best result among three methods. The dendrogram is depicted below in Figure 7 and Figure 8.

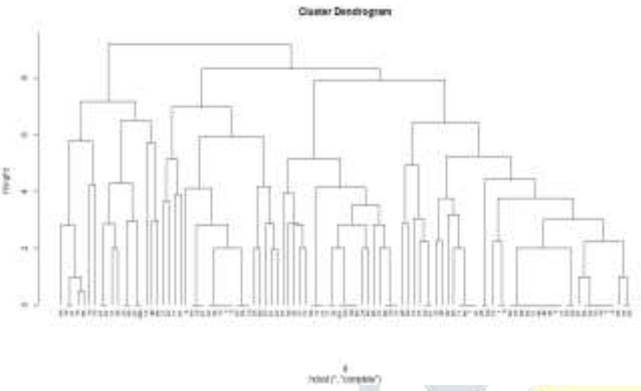

Figure 7:Dendrogram Plot

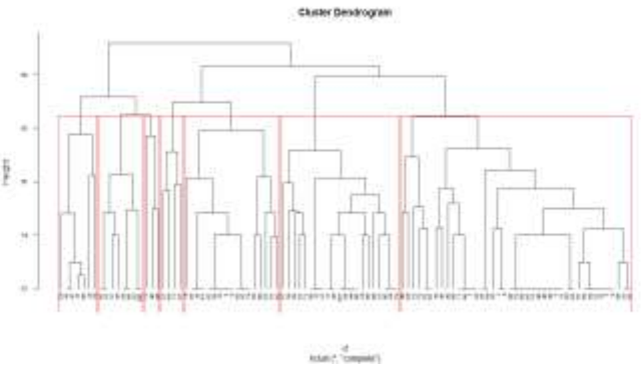

Figure 8: A Dendrogram of Hierarchical Clustering

Expectation Maximization(EM) Clustering: We can see and notice the suitable number of clusters is 7. EM algorithm is used and each cluster indicates to different type of animal. We circled different clusters with different colors. The cluster plot is depicted below in Figure 9.

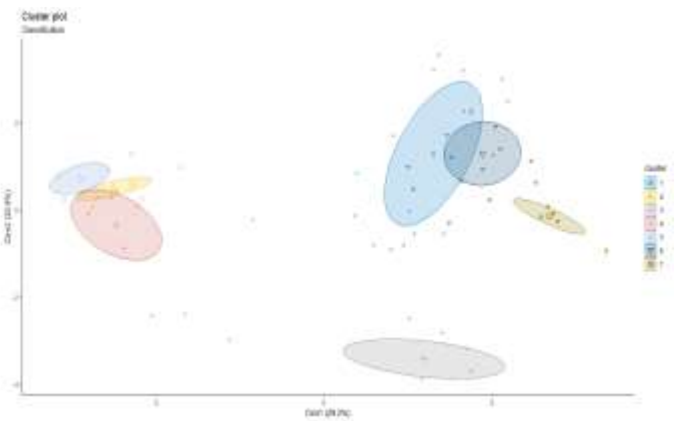

Figure 9: EM Cluster Plot

DBSCAN: We can see and notice the suitable number of clusters is 7. DBSCAN algorithm is used and each cluster indicates to different type of animal. We circled different clusters with different colors. The cluster plot is depicted below in Figure 10.

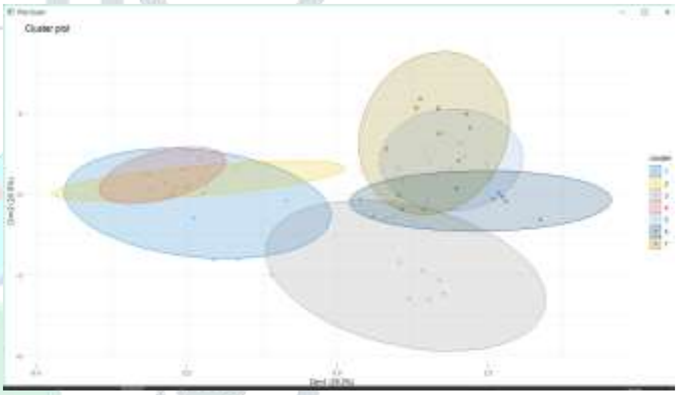

Figure 10: DBSCAN Cluster Plot

Farthest First Clustering: We can see and notice the suitable number of clusters is 5. Farthest First algorithm is used and each cluster indicates to different type of animal. We circled different clusters with different colors. The cluster plot is depicted below in Figure 11.

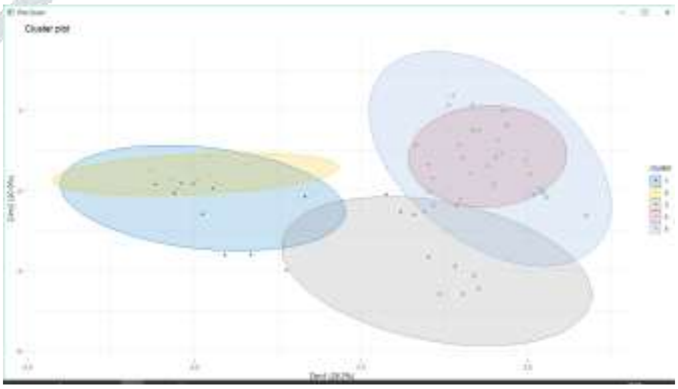

Figure 11: Farthest-First Cluster Plot

The performance comparison of the five clustering algorithms is presented in Table II.

TABLE II

Performance of Clustering Algorithms on Concept Formation for Zoo Data

Algorithm	Accuracy
K-Means	88%
Hierarchical	51%
Expectation Maximization	92%
DBSCAN	43%
Farthest First	72%

VI. Conclusion

In this project, we have reviewed data mining techniques and their potential use in ontology building. The algorithms that are based on these techniques are the basis for automated ontology building. The complex relations and concept structures give understanding on what information can be contained within ontology for use in expert systems. Project on automated ontology building describe how exactly data mining techniques are used for this task. The automatic concept clustering yields cluster patterns that provide information on function terms, hierarchies, classes, attributes and interrelations. The proposed work can be extended further by analyzing on the various hidden patterns that the clusters hold.

References

[1] Asunción Gómez-Pérez, Mariano Fernández-López,OscarCorcho,Ontological Engineering: With Examples from theAreas of Knowledge Management, E-commerce and the SemanticWeb, Springer, 2004,ch. 1,pp.5-10.

[2] Natalya F. Noy, Deborah L. McGuinness, “*Ontology Development101: A Guide to Creating Your First Ontology*”, StanfordUniversity, Stanford, CA, 2001.

[3] Clerkin, P., Cunningham, P., and Hayes, C., “Ontology Discoveryfor the Semantic Web Using Hierarchical Clustering,” TrinityCollegeDublin,Ireland, 2002.

[4] Blaschke, C., & Valencia, A., “Automatic Ontology Constructionfrom the Literature”, *Genome Informatics*,vol. 13, pp 201–213,2002.

[5] Quan, T. T., Hui, S. C., Fong, A. C. M., and Cao, T. H. Automaticgeneration of ontology for scholarly semantic Web. In: LectureNotes in Computer Science. vol. 3298.pp. 726–740,2004.

[6] Ganter, B.; Stumme, G.; Wille, R. (Eds.) Formal Concept Analysis:Foundations and Applications. Lecture Notes in ArtificialIntelligence, , Springer-Verlag, no. 3626, 2005.

[7] Wuermli, O., Wrobel, A., Hui S. C. and Joller, J. M. “Data MiningFor Ontology Building: Semantic Web Overview”, DiplomaThesis–Dep. of Computer science , Nanyang TechnologicalUniversity.,2003.

[8] Dahab, M. Y. Hassan, H., and Rafea, A.., “TextOntoEx: Automaticontology construction from natural English text, Expert Systemswith Applications*” , Elsevier*.,vol .34 pp.1474-1480,2007.

[9] R. GeethaRamani, Lakshmi Balasubramanian ,ShomonaGraciaJacob, “Data Mining Method of Evaluating Classifier PredictionAccuracy in Retinal Data”, *Proceedings of IEEE InternationalConference on Computational Intelligence and ComputingResearch.*, pp.426-429, 2012.

[10] ShomonaGaciaJacob , R. GeethaRamani, “Mining ofclassification patterns in clinical data through data miningalgorithms”, *Proceedings of the International Conference on Advances in Computing, Communications and Informatics*,pp.997-1003, 2012.

[11] ShomonaGracia Jacob , R. GeethaRamani ,” Evolving EfficientClustering And Classification Patterns In LymphographyDatathrough Data Mining Techniques” ,*International Journal on SoftComputing*, vol.3,no.3,2012.

[12] R. GeethaRamani ,ShomonnaGracia Jacob , “Prediction of P53Mutants (Multiple Sites) Transcriptional Activity Based onStructural (2D&3D) Properties” , *PloS one.*, vol.8, no.2,e55401,2013.

[13] R. GeethaRamani ,ShomonaGracia Jacob “ Discovery ofKnowledge Patterns in Lymphographic Clinical Data throughData Mining Methods and Techniques “ *Advances in Computingand Information Technology.*, pp 129-140 , 2013.

[14] Nancy. P, R. GeethaRamani ,ShomonaGracia Jacob. “Discoveryof Gender Classification Rules for Social Network Data usingData Mining Algorithms” *Proceedings of the IEEE International Conference on Computational Intelligence and ComputingResearch*., pp 808-812, 2011.

[15] A. Shanthi , R. GeethaRamani, “Classification of VehicleCollision Patterns in Road Accidents using Data MiningAlgorithms” *International Journal of Computer Applications.*,vol.35,no. 12,pp. 30-37, 2011.

[16] ShomonaGracia Jacob, R.GeethaRamani, Nancy.P, “EfficientClassifier for Classification of Hepatitis C Virus Clinical Datathrough Data Mining Algorithms and Techniques”, *Proceedingsof the International Conference on Computer Applications,Pondicherry.*,pp.27-31, Jan.2012.

[17] ShomonaGracia Jacob, R.GeethaRamani, “Evolving Efficientclassification rules from Cardiotocography data through datamining methods and techniques”, *European Journal of ScientificResearch* ,vol.78. no.3 pp. 468-480, June .2012.

[18] Nancy .P, R. GeethaRamani, “Discovery of Patterns andEvaluation of Clustering algorithms in Social Network Data(Facebook 100 Universities) through Data mining methods andtechniques”, *International Journal of Data Mining andKnowledge Management Process*.,vol.2, no.5,Sep.2012.

[19] UCI Machine Learning Repository, [Online]. Available:http://archive.ics.uci.edu/ml/datasets/Zoo.

[20] R.GeethaRamani, Lakshmi Balasubramanian, ShomonaGraciaJacob, “Automatic Prediction of Diabetic Retinopathy andGlaucoma through Image processing and Data MiningTechniques”. Proceedings of International Conference onMachine Vision and Image Processing., pp.163-167, 2012.

[21] SapnaGupta , Prof. Vikrant Chole,” Document Clustering Using Concept Weight”.