IJCSE International Journal of Computer Science and Engineering Open Access
Technical Paper Volume-2, Issue-2 E-ISSN: 2347-2693

An Overview of Ontology Based Text Document Clustering Algorithms

Anuradha Awachar[1*], Rajashree Bairagi[2], Vijayalaxmi Hegade[3] and Mahadev Khandagale[4]

[1*]Computer Department, PCCOE, Pune University, India, anu2awachar@gmail.com
[2]Computer Department, PCCOE, Pune University, India, rajashree.bairagi@gmail.com
[3]Computer Department, PCCOE, Pune University, India, vijayalaxmivhegade@gmail.com
[4]Computer Department, PCCOE, Pune University, India, mahadevkkh@gmail.com

www.ijcaonline.org

Received: 08 Feb 2014 Revised: 14 Feb 2014 Accepted: 26 Feb 2014 Published: 28 Feb 2014

Abstract- Text document clustering is an important activity in data mining. It is emerged from text retrieval, and had important application in establishing information retrieval, knowledge management system. Clustering can help to get solutions for many problems associated with real time applications such as, in commercial; in biotechnology; in geography; in the banking sector; in the insurance industry; in the Internet etc. Hence It is important to know different ways available to implement clustering. In Text based clustering approach using title of document it is found out to which cluster this document belongs But It doesn't give better results because it may possible that same document is renamed with two different names, as content of both documents are similar it is expected that the document should go to the same cluster but depending on the name of the document it may go to two different clusters . A new approach called semantic based text clustering [1] comes into picture in which entire document is parsed and depending on its content it is clustered. Ontology based text clustering [2] is a way to implement semantic based clustering. In this paper we discussed about different Ontology based algorithms like K-means, DBScan, SOM etc.

Keywords- Term-Clustering, k-means, Single-Linkage, DBSCAN, Self-Organizing Maps, F1Measure

INTRODUCTION

Clustering can help to get solutions for many problems associated with real time applications such as, in commercial, cluster analysis was used to find the different customer groups, and summarize different customer group characteristics through the buying habits; in biotechnology, cluster analysis was used to categorized animal and plant population according to population and to obtain the latent structure of knowledge; in geography, clustering can help biologists to determinate the relationship of the different species and different geographical climate; in the banking sector, by using cluster analysis to bank customers to refine a user group; in the insurance industry, according to the type of residence, around the business district, the geographical location, cluster analysis can be used to complete a automatic grouping of regional real estate, to reduce the manpower cost and insurance company industry risk; in the Internet, cluster analysis was used for document classification and information retrieval etc. In this paper information about ontology based text document clustering algorithms is given .

English text is composed of words, but other language may have new semantic unit. So far, Figure 1 shows Pretreatment processes for text clustering algorithms. the natural language text, we still can't use strict syntax and semantic rules to analyze text semantic that classified the text comparative intelligently. Due to the nature of the English text itself, English words are separated by Spaces, and basically every word can express the meaning independently. Therefore, the key word for extraction is space, all kinds of punctuation is the limits of the document words.

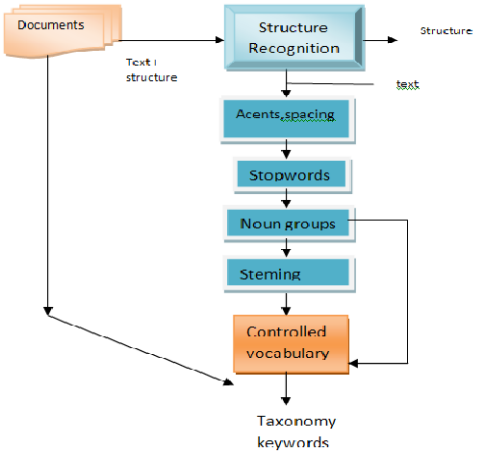

Figure 1: shows Pretreatment processes for text clustering algorithms

Ontology based text clustering (OTC):
Ontology is a knowledge repository in which concepts and terms are defined as well as relationships between these concepts. It consists of a set of concepts, axioms, and relationships that describe a domain of interests and

Corresponding Author: *Anuradha Awachar*

represents an agreed-upon conceptualization of the domain's "real-world" setting. Implicit knowledge for humans is made explicit for computers by ontology. Thus, ontology can automate information processing and can facilitate text mining in a specific domain(such as research project selection). The remainder of this paper is organized as follows. Section 2 reviews an ontology based K-means algorithm, DBScan. Section 3 gives detailed information about SOM algorithm. Section 4 gives comparative study of all above. Section 5 provides the conclusion.

REVIEW OF K-MEANS, DBSCAN ALGORITHM

K-Means:
K-means clustering is a method of cluster analysis which aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean. It is one of the simplest unsupervised learning algorithm that solve the well known clustering problem. The k-means algorithm is an evolutionary algorithm that gains its name from its method of operation. The algorithm clusters observations into k groups, where k is provided as an input parameter. It then assigns each observation to clusters based upon the observation's proximity to the mean of the cluster. The cluster's mean is then recomputed and the process begins again.

1. Begin with a decision on the value of k= number of clusters.

2. Put any initial partition that classifies the data into k clusters.

- Take the first k data as single-element clusters
- Assign each of the remaining (N-k) data to the cluster with the nearest centroid. After each assignment, recomputed the centroid of the gaining cluster.

3. Take each data in sequence and compute its distance from the centroid of each of the clusters. If a sample is not currently in the cluster with the closest centroid, switch this sample to that cluster and update the centroid of the cluster gaining the new sample and the cluster losing the sample. The distance metric used is Euclidean distance.

4. Repeat step 3 until convergence is achieved, that is, until a pass through the training sample causes no new assignments.

Figure 2: K Means Algorithm

Time Complexity: O (n * K * I * d)
Where n = number of points, K = number of clusters, I = number of iterations, d = number of attributes.

Advantages and disadvantages:
The main advantages of this algorithm are its simplicity and speed which allows it to run on large datasets. Its disadvantage is that it does not yield the same result with each run, since the resulting clusters depend on the initial random assignments (the k-means++ algorithm addresses this problem by seeking to choose better starting clusters). It minimizes intra-cluster variance, but does not ensure that the result has a global minimum of variance. Another disadvantage is the requirement for the concept of a mean to be definable which the case is not always. For such datasets the k-medoids variants is appropriate. An alternative, using a different criterion for which points are best assigned to which centre is k-medians clustering.

DBScan algorithm:
Density based spatial clustering of applications with noise, DBSCAN; rely on a density-based notion of clusters, which is designed to discover clusters of arbitrary shape and also have ability to handle noise. The main task of this algorithm is class identification, i.e. the grouping of objects into meaningful subclasses. Two global parameters of DBSCAN algorithms are: If p is a core point, then a cluster is formed. If p is a border point, no points are density- reachable from p and DBSCAN visits the next point of the database.

- DBSCAN(D, eps, MinPts)
- 1. C = 0
- 2. for each unvisited point P in dataset D
- mark P as visited
- 3. N = getNeighbors (P, eps)
- if sizeof(N) < MinPtsmark P as NOISE
- else
- C = next cluster expandCluster(P, N, C, eps, MinPts)
- 5. expandCluster(P, N, C, eps, MinPts)
- 6. add P to cluster C
- 7. for each point P' in N
- if P' is not visited
- mark P' as visited
- 8. N' = getNeighbors(P', eps)
- if sizeof(N') >= MinPts
- N = N joined with N'
- if P' is not yet member of any cluster
- add P' to cluster C

Figure 3: DBScan Algorithm

EXPERIMENTAL RESULTS

The performance of the two algorithms, ontology based k-means algorithm and ontology-based DBScan algorithms,

were compared using two text corpora, namely, ModApte a popular variant of Reuters 21578 and 20 Newsgroup. The performance of the two algorithms was analyzed using the precision, recall, F measure and Accuracy. The F-measure is calculated from two measures, precision and recall, which are derived from four values, namely, true positive (TP), true negative (TN), false positive (FP) and false negative (FN) during analysis of performance.(table1).

	Same category	Different category
Same cluster	TP	FP
Different cluster	FN	TN

Table1

The equation used to calculate precision (p) and recall (r) are given in Equ 4 and 5.

Pij= Nij/Nj ……Equ 4

Rij= Nij/N ……Equ 5

where Nij is the number of objects of class „i" in cluster „j". Nj is the number of objects in cluster „j", Ni is the number of objects of class „i'.

The F-measure is calculated using Equ 6. It is always desired to obtain a large F-measure, which indicates better clustering performance. In general, a larger F-measure value indicates better clustering result. The accuracy is calculated using Equ (7).

Fij=2(Pij)(Rij)/Pij+Rij ……..Equ(6)

Accuracy=(TP+TN) / (TP+TN+FP+FN) ……..Equ (7)

In the results, OKM and ODB refer to the ontological-based K-Means and ontological-based DBScan algorithms. The precision and recall values obtained for the two algorithms are shown in Table 2.

Dataset	FMeasure	FMeasure	Accuracy	Accuracy
	OKM	ODB	OKM	ODB
20-News group	0.665	0.791	92.43	93.96
Reuters -21578	0.636	0.702	90.16	91.66

Table 2: Precision and Recall

From the tabulated results again the ontological DBScan proves to provide better clustering both in terms of F Measure and accuracy than ontological K-means with both the text datasets. The results from the various experiments show that the clustering algorithm that uses semantics of the documents for term weighting and DBScan for clustering produces significant difference in results when compared with the ontological K-means algorithm and has improved the process of clustering [4].

SELF-ORGANIZING ALGORITHM STUDY

In 1975 Teuvo Kohonen introduced new type of neural network that uses competitive, unsupervised learning. This approach is based on WTA (Winner Takes All) and WTM (Winner Takes Most) algorithms. Therefore, these algorithms will be explained here briefly. The most basic competitive learning algorithm is WTA. When input vector (a pattern) is presented, a distance to each neuron's synaptic weights is calculated. The neuron whose weights are most correlated to current input vector is the winner. Correlation is equal to scalar product of input vector and considered synaptic weights. Only the winning neuron modifies it's synaptic weights to the point presented by input pattern. Synaptic weights of other neurons do not change. The learning process can be described by the following equation:

$$\| x - wc \| = \min_j \{\| x - wi \|\}$$

$$wc(t+1) = wc(t) + [x(t) - wc(t)]$$

where i [0..*numer of neurons*] , *Wi* represents all synaptic weights of the winning neuron, is learning rate in the interval [0, 1] that linearly proportional with t inverse reduced and shows total weights attached to the winning cell and x stands for current input vector. In this section WTM strategy describe that is a extension of WTA strategy. The difference between those two algorithms is that many neurons in WTM strategy adapt their synaptic weights in one learning iteration. In this case not only the winner, but also its neighborhood adapts. The further the neighboring neuron is from the winner, the smaller the modification which is applied to its weights. This adaptation process can be described as:

$$Wi+1 = Wi + \eta * K(i,x) * (x - W1)$$

For all neurons i that belong to winner's neighborhood. Wi stands for synaptic weights of neuron i and x is current input vector. η stands for learning rate and N(i,*x)* is a function that defines neighborhood. Where shows the weights attached to the cells and cells located in the neighborhood of winning. X vector is input pattern and learning rate that have a positive value smaller than the unit. K (i,x) is Neighborhood function that is Gaussian kernel that was a descent function and with a way of win cell and time decreases. And thus the cell in farthest neighborhood will have low change in weights. Neighborhood function can be described as:

$$N(1,x) = \exp(-\|x - w1\|^2, t) \quad \text{for } w1 \in \lambda(1,x)$$

otherwise
In order to train SOM network the Euclidean distance between input vector and weight vectors of all cells should be computed. The cell which has the lowest distance with input vector, in other words the cell which has the most similarity to input pattern is selected as a winner and its adjoined weights change in order to approach input pattern. In addition, adjacent cells are selected and according to their distance to winner cell their weights are modified in the same orientation. The movement of cells and the number of mobile cells is high in the beginning of algorithm and they reach their minimum value due to reducing the rate of learning and neighbor radius. This algorithm draws input vector on one line (in two-dimensional topological state).

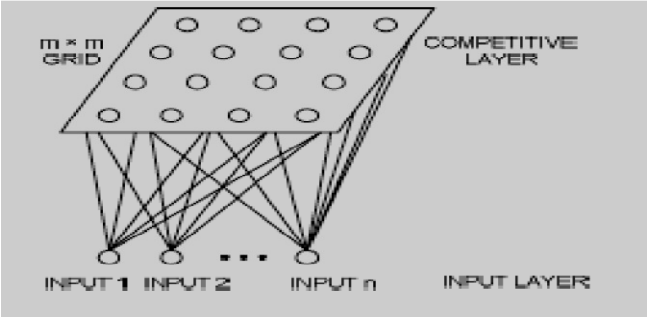

Figure 4: SOM with the neighborhood of two dimensional input vectors

Input patterns that are similar to each other, that have minimum Euclidean distance from each other, are also after mapped are placed together. In 1-D network each cell has 2 neighbors, a neighbor on the left and the other cell in the right placed. Two-dimensional network in each cell has four neighbors, which is on the left, right, top and bottom cell are placed.

SOM algorithm can be summarized as follows:

- Choose weight of all the cells randomly.
- Apply input pattern to network.
- Find win cells.
- Select the neighbor cells.
- Correct weights attached to the cells and the winner of the neighbor cells according to their Euclidean distance, learning rate and neighborhood radius.
- Repeat stages 2 to 5 for the number of distinct and pre-determined periods.

Figure 5: SOM algorithm

IV. Result & Discussion/Experimental/ Analysis/Implementation

Method	Used_algo	Num_Of_Sys_cluster	num_of_correct_clusters	Num_of_common_CLusters	Precision	Recall	F1 measure
Mapping to 2d space	K-means	27	38	19	0.7	0.58	0.58
	DBSCAN	23	38	10	0.43	0.26	0.33
	SOM	28	38	22	0.785	0.578	0.66
Use_Average	K-means	17	38	12	0.44	0.31	0.36
	DBSCAN	26	38	8	0.23	0.15	0.19
	SOM	22	38	12	0.54	0.31	0.39

Table3

V. CONCLUSION

The purpose of this study is to give an overview of ontology based clustering algorithms. In this paper we studied three algorithms K-Means, DBSCAN and SOM. According to the results, SOM algorithm can give the effective results in semantic based clustering and removal of duplications. It can also be used to handle the large databases.

REFERENCES

[1]. "Ontology-based Semantic clustering" by dr. Aida alls and dr. Karina gibert computer science and mathematics sanroma aPh.d. thesis supervised by department of Tarragona.
[2]. "Ontology-based Text Document Clustering" by Andreas Hotho and Alexander Maedche and teffen Staab Institute AIFB, University of Karlsruhe, 76128 Karlsruhe, Germany.
[3]. "Ontology-based Text Clustering" by A. Hotho and S. Staab A. Maedche.
[4]. " Survey of Clustering Algorithms " by Rui Xu, Student Member, IEEE and Donald Wunsch II, Fellow, IEEE.
[5]. "Support Vector Clustering" by Asa Ben-Hur asa ,Raymond and Beverly, Nello Critianini, John Shawe-Taylor and Bob Williamson.
[6]. "Modern Information Retrieval " a book written by yates & neto.

AUTHORS PROFILE

Anuradha Anil Awachar

B.E.(Computer Science),
Pimpri Chinchwad College of Engineering,
Nigadi,Pune.

Vijayalaxmi Vilas Hegade

B.E.(Computer Science)
Pimpri Chinchwad College
of Engineering,
Nigadi,Pune.

Rajashree Babudas Bairagi
B.E.(Computer Science)

Pimpri Chinchwad College
of Engineering,
Nigadi,Pune.

Mahadev Khandu Khandagale
B.E.(Computer Science)
Pimpri Chinchwad College
of Engineering,
Nigadi,Pune.