ISSN No. 0976-5697

Volume 5, No. 7, September-October 2014

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

Comparison between Clustering Algorithms Based On Ontology Based Text Mining Techniques

S.Suguna, B.Gomathi
Department of Computer Science
Sri Meenakshi Govt College for Women
Madurai-625 002, India
2

Abstract: Large number of documents is grouped according to their similarities. The Text Mining methods have been proposed to solve the problem by automatically classifying text documents, mainly in English. But this method has some limitations when dealing with non-English language texts, e.g., Chinese research proposals. An ontology based text MINING approach is to cluster the documents based on their similarities. For ontology making in text documents, the clustering algorithm is need to be applied. This paper focuses on two clustering techniques SOM and DVA for clustering documents based on their similarities. SOM and DVA algorithms are applied after the preprocessing process with documents such as Stop Word Removal, Stemming. DVA includes Vector Dimension Reduction also. The proposed Ontology based text mining technique is efficient and effective in terms of time, reliability and quantity. DVA out performs than SOM in terms of all the factors.

Keywords: Text Mining, Ontology, OTMM, Clustering, SOM and DVA Algorithm

I. INTRODUCTION

In the fast developing information explosion era, much of the knowledge available is stored as text. It is not surprising, therefore, that data mining (DM) and information retrieval (IR) from text collections (text mining) has become an active and exciting research area in computer applications [1]. The purpose of Text Mining is to process unstructured (textual) information, extract meaningful numeric indices from the text, and, thus, make the information contained in the text accessible to the various data mining (statistical and machine learning) algorithms. Information can be extracted to derive summaries for the words contained in the documents or to compute summaries for the documents based on the words contained in them. Hence, we can analyze words, clusters of words used in documents, etc., or one could analyze documents and determine similarities between them or how they are related to other variables of interest in the data mining project. Clustering or segmentation of data is a fundamental data analysis step that has been widely studied across multiple disciplines for over 40 years. Clustering text documents into different category groups is an important step in indexing, retrieval, management and mining of abundant text data on the Web or in corporate information systems.

In section 2 Literature survey is discussed. Our proposed algorithm is explained in section 3. Section 4 deals with performance analysis using line chart with SOM and DVA algorithms. Conclusion is in section 5.

II. LITERATURE SURVEY

An Ontology based text mining method discussed in [2]. The R&D project selection [3], how the project selection and development can be made is described. The text mining application is discussed in [4]. The Information Retrieval and Information Extraction [5] is to extract the documents for users need.

Different methods are used for measure the performance of ontology based information extraction. The Balanced Distance Metric (BDM) is described[6]. The survey performed here is [7] different component based information extraction rules are categorized. The extraction from the large amount of data using RDBMS is discussed in [1]. The Ontology based text mining framework is used to increase the effectiveness of the project, which is described in [8].

The Ontology based text mining method and Information Extraction is used for biological domain is explained in [9]. In pattern discovery for text mining [10] and the Ontology based concept weighing [11], different concepts are compared. The Ontology based text mining techniques uses different clustering methods. The clustering methods to cluster the text documents are discussed in [12]. The ontology based applications such as network based P2P Fuzzy logic techniques and the approach of automatic construction of hypertexts in text mining are explained in [13, 14].The different aspects of semantic web[15] and the learning features are discussed. Like the previous one, [16] this book also teach the researchers and developers who are in ontology based applications development and in E-Learning. Several techniques in text categorization and the learning rule algorithms are discussed [17].

III. PROPOSED ALGORITHM

To solve the aforementioned problems, an Ontology-based TMM (OTMM) is proposed. An ontology is a knowledge repository in which concepts and terms are defined as well as relationships between these concepts[18].

Some of the reasons to develop ontology

a. To share common understanding of the structure of information among people or software agents
b. To enable reuse of domain knowledge
c. To make domain assumptions explicit
d. To separate domain knowledge from the operational knowledge

e. To share common understanding of the structure of information among people or software agents

f. To share common understanding of the structure of information among people or software agents

This section is discussed as follows. In Phase1 text documents collection process is explained. In Phase 2 preprocessing process such as stop word removal and stemming process is discussed. In Phase 3 grouping process is explained. In Phase 4 and 5 the importance of Ontology making and clustering by SOM and DVA algorithms are discussed. For algorithm making the parameter usage [19] is discussed. The proposed architecture is shown in figure 1.

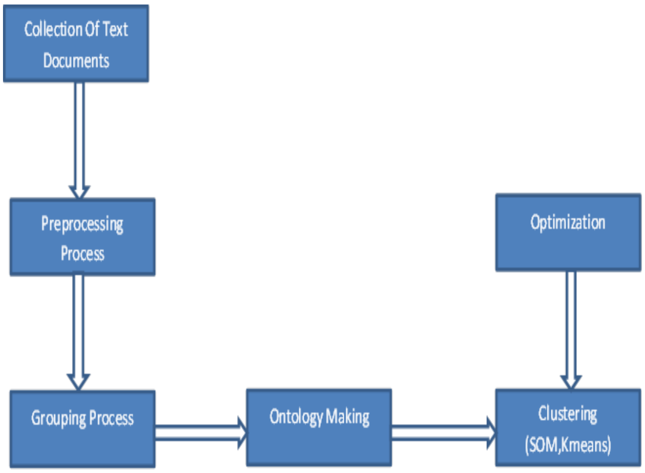


Figure 1. Ontology Based Mining Process

A. Phase 1: Collection Of Text Documents:

The text documents are collected for the mining process. There is no limitation for collecting the documents. They are placed in a particular directory. Because every time we choose the file path and select any document for text mining process.

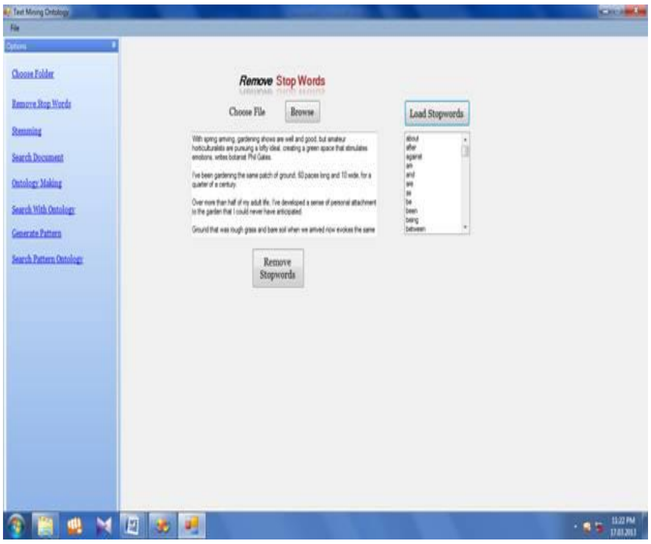
B. Phase 2: Preprocessing Process:

Step1: Stop word removal

A stop word is a commonly used word (such as "the") that a search engine has been programmed to ignore, both when indexing entries for searching and when retrieving them as the result of a search query. When building the index, most engines are programmed to remove certain words from any index entry. The list of words that are not to be added is called a stop list. Stop words are deemed irrelevant for searching purposes because they occur frequently in the language for which the indexing engine has been tuned. In order to save both space and time, these words are dropped at indexing time and then ignored at search time. Table 1 shows the sample stop words, and Figure 2 shows the sample stop word removal process. After choosing the file the list of stop words removed are shown in the list box.

Table 1. The List of sample stop words

a	an	and	are	as	at	be	by	for	from
has	he	in	is	it	its	of	on	that	the
to	was	were	will	with					



With spring arriving, gardening shows are well and good, but amateur horticulturalists are pursuing a lofty ideal, creating a green space that stimulates emotions, writes botanist Phil Gates.

I've been gardening the same patch of ground, 60 paces long and 10 wide, for a quarter of a century.

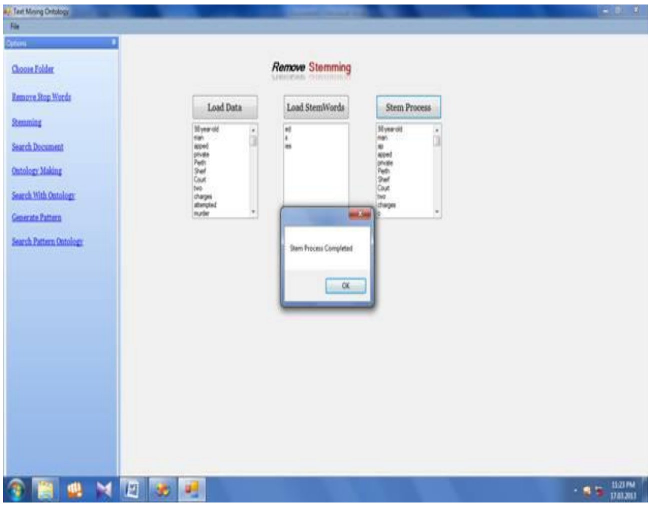
Over more than half of my adult life, I've developed a sense of personal attachment to the garden that I could never have anticipated.

Ground that was rough grass and bare soil when we arrived now evokes the same

Figure 2. Removal of Stop Words

Step 2: Stemming Process:

Stemmers used to reduce a word to its stem or root form. Thus, the key terms of a query or document are represented by stems rather than by the original words. This not only means that different variants of a term can be conflated to a single representative form – it also reduces the dictionary size, that is, the number of distinct terms needed for representing a set of documents. A smaller dictionary size results in a saving of storage space and processing time. In Figure 3 sample stemming process is shown.



Stem Process Completed

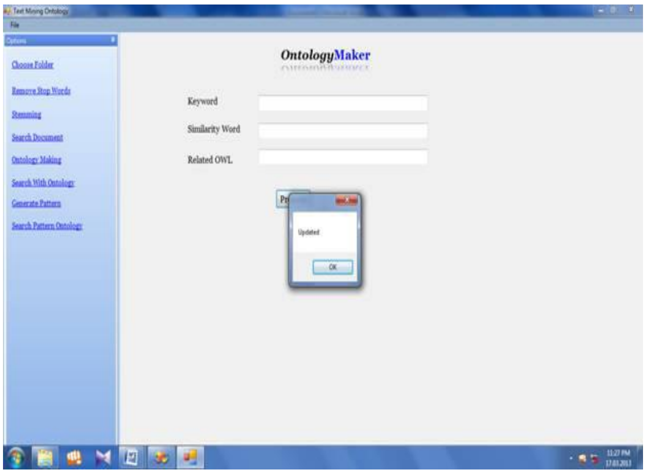
Figure 3. Stemming Process

C. Phase 3: Grouping Process:

After completing the preprocessing steps such as stop word removal and stemming process the words are grouped according to their similarity.

D. Phase 4: Ontology Making:

The creation of domain Ontology's is also fundamental to the definition and use of an enterprise architecture framework. Pragmatically, a common ontology defines the vocabulary with which queries and assertions are exchanged among agents. Ontological commitments are agreements to use the shared vocabulary in a coherent and consistent manner. The agents sharing a vocabulary need not share a knowledge base; each knows things the other does not, and an agent that commits to Ontology is not required to answer all queries that can be formulated in the shared vocabulary. The Ontology maker is shown in Fig. 4.

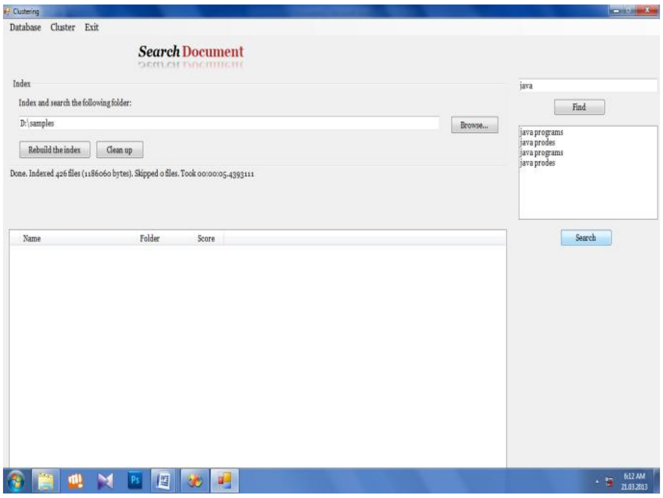


Updated

Figure 4. Ontology Making

E. *Phase 5: Clustering:*

Clustering usually groups keywords of documents into clusters and constructs ontology by selecting representative concepts from each cluster. Cluster analysis as such is not an automatic task, but an iterative process of knowledge discovery or interactive multi-objective optimization that involves trial and failure. It will often be necessary to modify data preprocessing and model parameters until the result achieves the desired properties. Sample clustering of documents is shown in Fig.5.



Index and search the following folder: D:\samples

Done. Indexed 426 files (1186060 bytes). Skipped 0 files. Took 00:00:05.4393111

java

java programs
java prodes
java programs
java prodes

Figure 5. Clustering the Documents

The SOM and DVA Clustering Algorithms are used in our work . The SOM and the DVA algorithms performances are compared and analyzed. The Algorithms are given below.

a. *SOM Algorithm:*

Step 1: Initialize network weight vectors $w_{i,}$ initialize learning rate parameter µ, define topological neighborhood function and initialize parameter Nq, set k=0.

Step 2: Check Stopping condition. If false, continue; If true,stop.

Step 3: For each training vector x, perform Steps 4 to 7.

Step 4: Compute the best match of a weight vector with the input

$Q(x)= \max sim(x,w_i)\ v_i$

Where sim can be calculated as cosine value of the angle between the vector.

Step 5: For all the units in the specified neighborhood i € N_q (k), where q is the winning neuron, update the weight vectors according to:

$\{wi(k)+\pi(k)[x(k)-w_i(k)]$ i € $N_q(k)$ w_i $(k+1)\} = \{wi(k)$ i € $N_q(k)\}$

where 0 < µ(k) < 1 (the learning rate parameter).

Step 6: Adjust the learning rate parameter

Step 7: Appropriately reduce the topological neighborhood $N_q(k)$

Step 8: Set k = k+1; then go to step 2.

b. *DVA Algorithm (KMeans):*

The optimization problem itself is known to be NP-hard, and thus the common approach is to search only for approximate solutions. A particularly well known approximative method is Lloyd's algorithm, often actually referred to as "*k-means algorithm*". It does however only find a local optimum, and is commonly run multiple times with different random initializations. Variations of k-means often include such optimizations as choosing the best of multiple runs, but also restricting the centroids to members of the data set (k-medoids), choosing medians (k-medians clustering), choosing the initial centers less randomly (K-means++) or allowing a fuzzy cluster assignment (Fuzzy c-means).

The steps for DVA algorithms are given below.

a) Document Vector Computation
b) Document Clustering
c) Document Collection
d) Class Document Collection

```
{
public  List<String> DocumentList { get; set; }
}
Calculates TF-IDF weight for each term t in document d
private static float FindTFIDF(string document, string term)
{
float tf = FindTermFrequency(document, term);
float idf = FindInverseDocumentFrequency(term);
return tf * idf;
}
```

IV. PERFORMANCE EVALUATION

The Core i5 processor is used. For system configuration, windows 7 operating system is used. The problem is done using the application server Visual Studio 2007.Tthe front end C#.net and the Back end SQL Server is used.
In this work, the SOM and DVA algorithms performances are compared against time based frequency and reliability of retrieved documents.

A. *Time Based Frequency:*

Here, the time needed to process the number of documents by SOM and DVA algorithms are analyzed and the results are shown in Figure 6. For analyzing the quantitative data[20], if the large amount of data from the two different algorithms are used the chart and the axis defined values are explained.

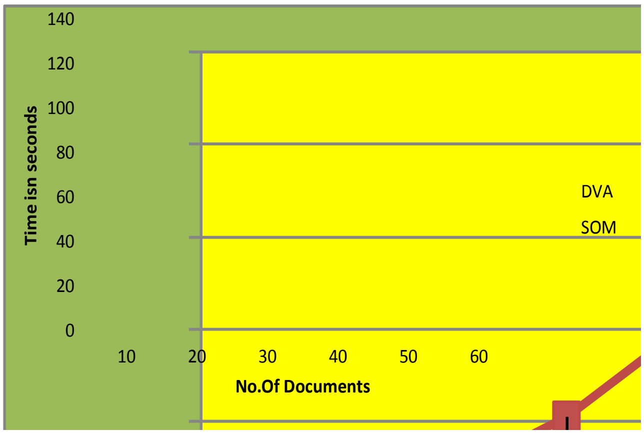


Figure 6. Time Based Frequency

B. *Reliable Documents:*

Reliability of retrieved documents can be made by using the number documents and the frequency measurement. The SOM and DVA algorithms are compared using the line chart and shown in Figure 7. In this Figure 7 the reliability of retrieved documents are analyzed.

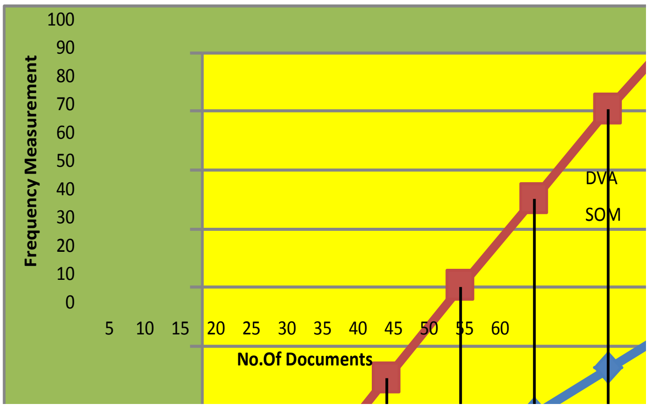


Figure 7. Reliable Documents

C. *SOM And DVA Factors:*

Table 2 shows the factors according to which the SOM and DVA algorithms performances are compared in our work.

Table 2. Performance Evaluation Analysis

	SOM	DVA(K means)
The Size of Data Set	Small	Huge
Number of Clusters	Minimum	Maximum
Types of Dataset	Ideal	Random
Time	Maximum	Minimum
Reliability	Low	High

V. CONCLUSION

Thus an Ontology based text mining approach is to cluster the documents based on their similarities. This method is systematic, efficient and effective for clustering the documents. In this work the collected documents are preprocessed by stopword removal and stemming process. The preprocessed details are grouped then the documents are clustered using SOM and DVA algorithms. The SOM and DVA algorithms performances are compared against time based frequency and reliability of retrieved documents. We proved that DVA outperformances then SOM. Future work is planned to compare the performance of document clustering when various similarity measures and clustering algorithms are combined with ontology features of documents. The Future work suggested that there is a need to compare the results of manual classification to Systematic classification.

VI. REFERENCES

[1]. Luis Tari, Jo¨ rg Hakenberg, Yi Chen, Tran Cao Son, Graciela Gonzalez, and Chitta Baral , "Incremental Information Extraction Using Relational Databases" IEEE Transactions On Knowledge And Data Engineering, Vol. 24, No. 1, January 2012.

[2]. Jian Ma, Wei Xu, Yong-Hong Sun, Efraim Turban, Shouyang Wang, and Ou LiuAn "Ontology-Based Text-Mining Method to Cluster Proposals for Research Project Selection" IEEE Transactions On Systems, Man, And Cybernetics – Part A:Systems And Humans, Vol, 42, No.3, May 2012.

[3]. Q.Tian,J Ma, and O.Liu, "A hybrid knowledge and model system for R & D project selection" Expert Syst. Appl., vol. 23, no. 3,pp. 265 – 271, Oct.2002.

[4]. M. Konchady, "Text Mining Application Programming" Boston, MA:Charles River Media, 2006.

[5]. Daya C. Wimalasuriya and Dejing Dou, "Ontology-based information extraction: An introduction and a survey of current approaches".

[6]. Diana Maynard, Wim Peters, Yaoyong Li, "Evaluating Evaluation Metrics for Ontology-Based Applications: Infinite Reflection".

[7]. Daya C. Wimalasuriya, Dejing Dou, "Components for Information Extraction: Ontology-Based Information Extractors and Generic Platforms".

[8]. S. Bloehdorn , P. Cimiano1, A. Hotho, and S.Staab"An Ontology-based Framework for Text Mining", July 28, 2004 .

[9]. Khaled Khelif, Rose Dieng-Kuntz, and Pascal Barbry "An Ontology-based Approach to Support Text Mining and Information Retrieval in the Biological Domain", Journal of Universal Computer Science, vol. 13, no. 12 (2007), 1881-1907.

[10]. Ning Zhong, Yuefeng Li, and Sheng-Tang Wu," Effective Pattern Discovery for Text Mining" IEEE Transactions On Knowledge And Data Engineering, Vol. 24, No. 1, January 2012.

[11]. Hmway Hmway Tar , Thi Thi Soe Nyaunt "Ontology-based Concept Weighting for Text Documents" World Academy of Science, Engineering and Technology 57 ,2011.

[12]. S.C. Punitha, K. Mugunthadevi, and M. Punithavalli ,"Impact of Ontology based Approach on Document Clustering" International Journal of Computer Applications (0975 – 8887) Volume 22– No.2, May 2011.

[13]. Thangamani . M, Dr. P.Thangaraj "Ontology Based Fuzzy Document Clustering Scheme for Distributed P2P Network" Global Journal of Computer Science and Technology Volume 11 Issue 5 Version 1.0 April 2011

[14]. H. C. Yang and C. H. Lee, "A text mining approach for automatic construction of hypertexts," *Expert Syst. Appl.*, vol. 29, no. 4, pp. 723–734, Nov. 2005.

[15]. V.Devedzic, "Semantic Web and Education", Springer, 2006.

[16]. D.Dicheva, R.Mizoguchi, and J. Greer,"Semantic Web Technologies for E-Learning", eds IOS Press, 2009

[17]. K. Aas and L.Eikvil,"Text categorization: A Survey,"Technical Report Raport NR 941,Norweigian Computing Center,1999.

[18]. Yao-Tang Yu,Chien-Chang Hsu, "A Structured Ontology Construction By Using Data Clustering And Pattern Tree Mining",pro.2011 INT'l Conf. On Machine Learning and Cybernetics,Guilin,10-13 July,2011.

[19]. R.Agarwal and R.Srikant, "Fast algorithms forming Association Rules in larged Databases", Proc.20 th INT'l Conf.very large databases, pp 478-499,1994.

[20]. N.Blaikie, "Analysing Quantitative Data", Sage Publications, 2003