Asian Journal of Convergence in Technology *Volume II, Issue III*
Issn No.:2350-1146, I.F-2.71

Relative analysis of Ontology based mining and mining using side-information.

Atiya Kazi,
Department of Information Technology
Finolex Academy of Mgmt. and Technology, Ratnagiri
atiyakazi@gmail.com

Priyanka Bandagale
Department of Information Technology
Finolex Academy of Mgmt. and Technology, Ratnagiri
priyabandagle@gmail.com

***Abstract* — Text mining applications generally disregard the side-information contained within the text document, which can enhance the overall clustering process. To overcome this deficiency, the proposed algorithm will work in two phases. In the first phase, it will perform clustering of data along with the side information, by combining classical partitioning algorithms with probabilistic models. This will automatically boost the efficacy of clustering. The clusters thus generated, can also be used as a training model to promote the solution of the classification problem. In the second phase, a similarity based distance calculation algorithm, which makes use of two shared word spaces from the DISCO ontology, is employed to perk up the clustering approach. This pre-clustering technique will calculate the similarity between terms based on the cosine distance method, and will generate the clusters based on a threshold. This inclusion of ontology in the pre-clustering phase will generate more coherent clusters by inducing ontology along with side-information.**

Index Terms – Clustering, Ontology, Side-information.

I. INTRODUCTION

There are several attributes in a text document that carry side-information for clustering purposes. Such side-information may be of different kinds, such as document provenance information, the links in the document, user-access behavior from web logs, or other non-textual attributes which are embedded into the text document. Such attributes may contain a tremendous amount of information for clustering purposes. But, an optimized way is necessary enable the mining process, so that the side information is correctly utilized. This probabilistic approach of mining can be also extended to the classification problem. Along with it an existing ontological schema can be added to the clustering process at compile time and its effects on the generated output could be analyzed. The current work statement can be put down as, Developing a novel Clustering approach for mining raw text data along with its side information, and comparing it with Ontology based clustering that provides semantically enhanced clusters. Traditionally, data mining comprises of clustering and classification on text based data, numeric data and web based data. In many application domains, a remarkable amount of side information is also available along with the documents which is not considered during pure text based clustering [8]. Clustering text collections has been scrutinized under Data mining in [13]. Some efficient streaming techniques use clustering algorithms that are adaptive to data streams, by introducing a forgetting factor that applies exponential decay to historical data [9]. Normally, text documents typically contain a large amount of Meta information which may be helpful to enhance the clustering process. While such side-information can improve the quality of the clustering process, it is essential to make sure that the side-information is not noisy in nature. In some cases, it can hamper the eminence of the mining process. Therefore, one needs an approach which, carefully perceives the consistency of the clustering distinctiveness of the side information, along with the text content. The core approach is to determine a clustering process where text attributes along with the additional side-information provide comparable hints regarding the temperament of the basic clusters, as well as, they ignore conflicting aspects. In recent times, Ontologies have become a vital part of fabricating knowledge, so as to create knowledge-rich systems. An ontology is formally defined as an explicit formal hypothesis of some domain of interest which helps in the interpretation of concepts and their associations for that particular domain [2]. To create any ontology, one needs a data mining expert who can analyze different domain concepts, domain hierarchies and the relationships between them for any specialized domain. A similar approach is proposed in [5], which uses domain based, schema based, constraint based and user preference based ontologies for enhancing the test clustering process. The current work focuses on generating clusters, by incorporating a similarity- based distance measurement scheme, using the DISCO ontology, during the pre-clustering phase of the data mining process. This ontology makes use of two SIM word spaces as explained below in Table 1. Each of them contains the word spaces, together with the word vector and the most similar words for each word.

II. Related Work

The major work in the field of data mining looks upon scalable clustering of spatial data, data with Boolean attributes, identifying clusters with non-spherical shapes and clustering for large databases[7]. Several general clustering algorithms are discussed in [3]. An efficient clustering algorithm for large databases, known as CURE, has been covered in [14]. The scatter-gather technique, which uses clustering as its primitive operation by including liner time clustering is explained in [16]. Two techniques which develop the cost of distance calculations, and speed up clustering automatically affecting the quality of the resulting clusters are studied in [10]. An Expectation Maximization (EM) method, which has been around ages for, text clustering has been studied in [12]. It selects relevant words from the document, which can be a part of the clustering process in future. An iterative EM method helps in refining the clusters thus generated. In topic-modeling, and text-categorization, a method has been proposed in [11] which makes use of, a mathematical model for defining each category. Keyword extraction methods for text clustering are discussed in [10]. The data stream clustering problem for text and categorical data domains is discussed in [8]. Speeding up the clustering process can be achieved by, speeding up the distance calculations for document clustering routines as discussed in [15]. They also improve the quality of the resulting clusters. However, none of the above mentioned works with the combination of text-data with other auxiliary attributes. The previous work dealing with network-based linkage information is depicted in [6], [7], but it is not applicable to the general side information attributes. The current approach uses additional attributes from side information in conjunction with text clustering. This is especially useful, when the Side- information can regulate the creation of more consistent clusters. There are three forms of extending the process of knowledge discovery, with respect to their related ontologies, which are categorized as follows [4],

- Using on hand ontologies for knowledge discovery, during data mining.
- Construction of ontologies through knowledge discovery from mined results.
- Constructing and extending ontologies through knowledge discovery via existing ontologies.

The combination of the first two plays a major role in the methodology of the current work of interest.

III. System Architecture

There are three major modules in the system as depicted using Fig.1, the first is the Preprocessing module, the second one is clustering module and the last is the Classification module. Each of them works in tandem to achieve the prime aim of knowledge gain from raw textual data.

A. Preprocessing Module

Documents from the datasets are stored within the corpus. In the preprocessing module, extracted documents from the repository are preprocessed. Preprocessing technique includes tokenizing the word, removing stop words, stemming the word and other preprocessing tasks such as calculating the Term Frequency for each word.

B. Clustering Module

The role of this module is the creation of clusters which are according to the content of the document. The system uses either COATES algorithm or an Ontology based method to generate the clusters. In the ontology based module, document similarity is usually measured by a pair-wise similarity function such as a cosine function, which reflects the similarity between two documents.

C. Classification

The classification engine is powered by an ontology of similarity indices that categorizes the input document with respect to the clusters generated using DISCO ontology. This ontology can be extended dynamically to allow classification without recompiling the system.

D. DISCO API

DISCO stands for extracting distributional related words using co-occurrences. It is a Java application which helps in regaining the semantic parallel between unreliable words and phrases. These similarities are generated on the basis of numerical analysis of very large text collections. The DISCO Java API provides methods for extracting the semantically most similar words for an input word, e.g. shy = (timid, quiet, soft-spoken, and gentle). It also works in the assessment of the semantic similarity between two input key words or phrases. The fundamental principles on which the method for knowledge discovery is based on says that the knowledge discovery process is dominated by pre-existing data and the ontologies relevant to the considered domain. Both data and ontologies evolve over a period of time by interacting with each other. The ontologies are enriched with knowledge from the patterns extracted with the help of the data mining tools, while the data is enriched

through new inferences which are derived from the ontologies. An excellent style manual for science writers is [7]. Data mining techniques are used to produce suitable patterns that can be filtered out and selected on the basis of their integration with the ontologies. Ontologies are used to select the input of the data mining techniques, based on their common relevance. New ontological models help in abstracting and validating the existing ones on their consistency. They help in consolidating the available data leading to multiple versions of ontologies and data. They can branch over multiple iterations. The proposed data mining system framework helps in supporting the system's intelligence by incorporating ontologies in the data mining framework. It includes the characteristics of a data-warehouse schema, along with the user preference based ontologies

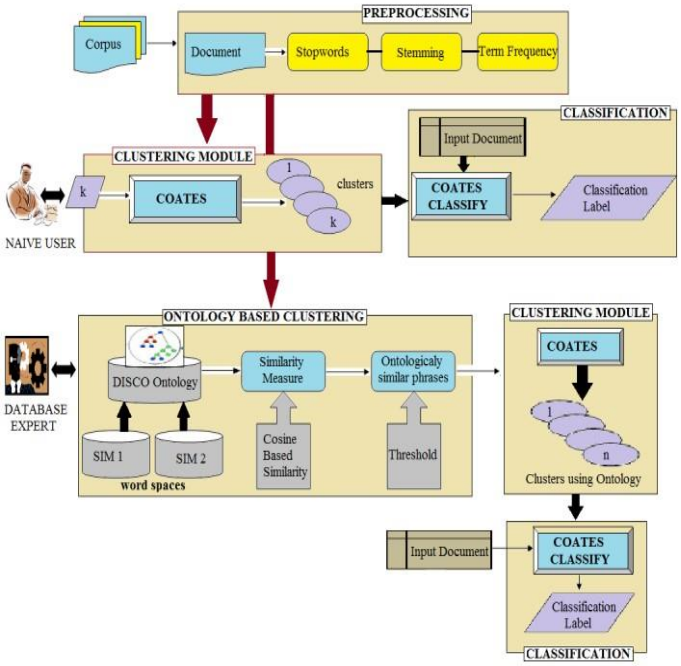


FIGURE 1. SYSTEM ARCHITECTURE

IV. ALGORITHM WORKING

The algorithm used for mining using side-information is referred to as COATES, which corresponds to Content and Auxiliary attribute based Text clustering algorithm [1]. The input to the algorithm is any cluster value k. Before the clustering begins, it is mandatory to segregate the stop-words and perform stemming for finding the root words. In each content-based phase, a document will be clustered using a closest seed centroid by making use of a cosine similarity function. This is followed by an auxiliary phase which generates a probabilistic model. It combines the attribute probabilities with the cluster-membership probabilities, by including the clusters created in the previous text-based phase. This determines the coherence of the text clustering by including side-information. The proposed algorithm for enhancing the clustering phase is an ontology based similarity distance measurement algorithm as shown in Fig.2. It uses cosine based distance calculation to find the similarity distance of two concepts denoted by C1 and C2. This algorithm is executed before the clustering phase begins. It will generate clusters using the SemDis(C1,C2) measure for two concepts C1 and C2 taken from the text within the dataset, with threshold above 0.4. The value $w_{c,}$ refers to the weight allocation function calculated using (1), while depth(C) presents the depth of concept C from the root concept to node C in ontology hierarchy, k is a predefined factor larger than 1 indicating the rate at which the weight values decrease along the ontology hierarchy.

Algorithm 1 Semantic distance solving algorithm

INPUT: Two concepts C1,C2
OUTPUT: Semantic Distance: SemDis(C1,C2)
METHOD:

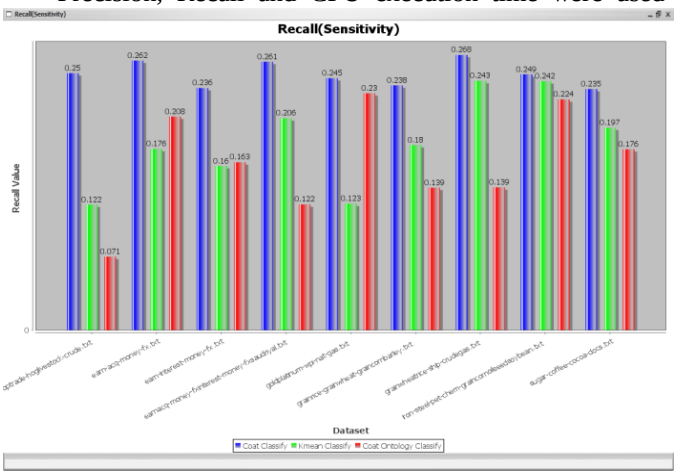
Begin
for all ontological concepts C1,C2 do
If C1, C2 are same concept
SemDis(C1,C2)=0
Else if there exists direct path relation,
SemDis(C1,C2)=w[sub(C1,C2)]
Else if there exists indirect path relation,
$SemDis(C1,C2) = \Sigma_{\alpha SPATH(C1,C2)} w_c[sub(C1,C2)]$
where SPATH is shortest distance in C1 and C2.
Else
SemDis(C1,C2)=minSemDis(C1,C0) + minSemDis(C2,C0)
end for

FIGURE 2. ONTOLOGY ALGORITHM FOR DISTANCE MEASUREMENT

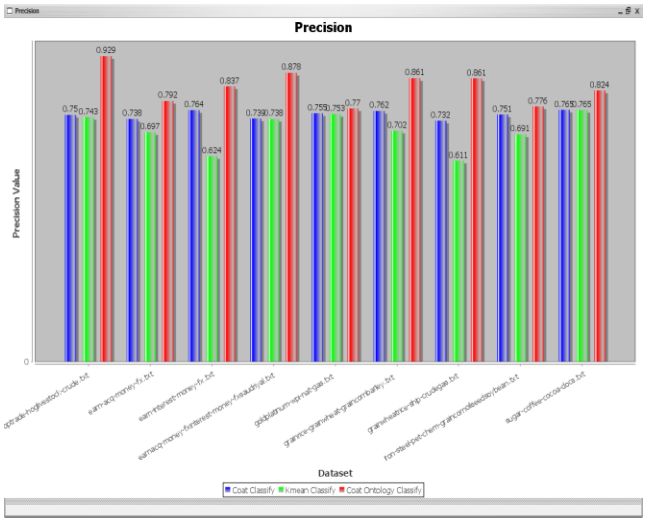
V. RESULT ANALYSIS

This section reports the experimental results generated while applying the similarity based distance calculation algorithm using ontology, to cluster documents. During

experimentation, eight datasets were collected from different fields of interest ranging from a sugar plantation to an iron and steel plant related data. To ascertain the performance of the proposed ontology based clustering scheme, several experiments were conducted using an Intel Core 2 duo machine with 2GB RAM. Four performance metrics, namely, Accuracy of generated cluster output, Precision, Recall and CPU execution time were used to



measure the cluster purity for different datasets. The results were compared with the existing COATES-CLASSIFY algorithm and K-means clustering algorithm. The overall results obtained after comparative analysis of these three algorithms for different number of clusters are depicted using tables and graphs in this section. As observed from the results, the Precision and Accuracy are highest for the proposed work as compared to existing k-means and COAT CLASSIFY method. On the other hand, the proposed algorithm takes more execution time than both baseline k means and the existing COATES algorithm. The metrics depend upon the True Positive and False Negative values gained after applying classification techniques to the existing datasets. The graphical analysis module generates the output graphs as revealed in Fig. 3 to Fig. 5.



A. *Precision and Recall*

To evaluate the accuracy of the proposed clustering algorithm, we have used the Recall and Precision performance metrics. In the current scenario, the precision and recall values are calculated for existing algorithm with values generated from classification module. This makes use of side information for generation of clusters and an ontology based approach which performs clustering based on semantic distance calculation. Fig. 3 and Fig. 4 shows the precision and recall value comparison for the current and proposed techniques with baseline k means method. As seen from the graphs, the ontology based approach has higher precision and lower recall. This fulfils the criteria of creation of more lucid clusters.

FIGURE 3. PRECISION VALUE COMPARISON

B. *Time Requirement*

The time requirements shown in Fig. 5 and Table I reveal that, the proposed algorithm takes more time than existing k means as well as COATES algorithm. This is because of the gigantic size of ontology repository, due to which the similar

FIGURE 4. RECALL VALUE COMPARISON

distance matching process may take more time. To avoid this, one needs a hierarchical structure with semantic interpretation of data. This is included in the future scope of the project.

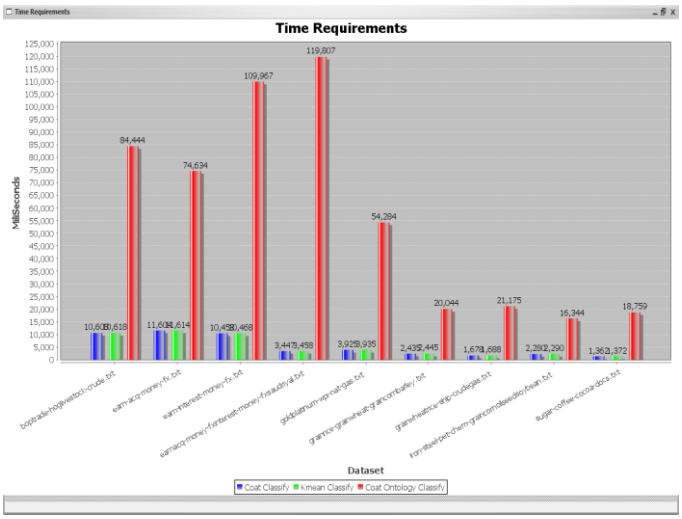


FIGURE 5. TIME REQUIREMENT COMPARISON

TABLE I. COMPARISON OF TIME OF EXISTING COATES ALGORITHM AND PROPOSED ONTOLOGY BASED ALGORITHM.

Dataset Name	Time-COATES(ms)	Time-Ontology(ms)
boptradehoglivestock-crude.txt	24482	32721
earn-acq-money-fx.txt	19144	33769
earnacq-money-fxinterest-money-fxsaudriyal.txt	23322	34983
goldplatinum-wpi-nat-gas.txt	10935	18331
grainrice-grainwheat-graincornbarley.txt	21533	17771
grainwheatrice-ship-crudegas.txt	8502	35103
iron-steel-pet-chem-graincornoilseed soybean.txt	11944	20626
sugar-coffee-cocoa-docs.txt	7009	15694

VI. CONCLUSION

The primary goal was to study the clustering problem, where auxiliary information is available with text and compare it with ontology based clustering. There is also an extension to include text classification using ontology, which proves that, the incorporation of side information and ontology enhance the classification process. The generated results have proved how the use of ontology elevates the quality of text clustering and classification, while maintaining a high level of efficiency. It was also observed that applying the ontologies before the phase of clustering minimally partitions the documents into coherent, clustered branches. The simple process of clustering and indexing documents by their ontological relationships puts ordered implication to the meaning of documents. While classification hierarchies only suggest, "What a document is about," ontological knowledge assigns richer significance to documents. Clustering algorithms that rely exclusively on probabilistic techniques may not help in uncovering the more complex semantic significance, endorsed to text document collections, by more affluent ontologies.

REFERENCES

[1] C. C. Aggarwal et al, "On the use of side-information for mining text data", IEEE Trans. Knowl. Data Eng, vol 26, pp. 1415-1429, June 2014.

[2] Henrihs Gorskis, Yuri Chizhov, "Ontology Building Using Data Mining Techniques", Information technology and management science, vol 15, pp 183-188, 2013.

[3] C. C. Aggarwal and C.-X. Zhai, "A survey of text classification algorithms," in Mining Text Data. New York, NY, USA: Springer, 2012.

[4] Mathieu d'Aquina, Gabriel Kronbergerb, and Mari Carmen Suárez-Figueroa, "Combining Data Mining and Ontology Engineering to enrich Ontologies and Linked Data", Proc. first International workshop on knowledge discovery and Data Mining , pp 19-24, 2012.

[5] Chin-Ang Wu et al., "Toward Intelligent Data Warehouse Mining: An Ontology-Integrated Approach for Multi-Dimensional Association Mining", Information Technology and Management Science, Expert Systems with applications, volume 38, Issue 9, pp 11011-11023, sept- 2011.

[6] J. Chang and D. Blei, "Relational topic models for document networks", in Proc. AISTASIS, Clearwater, FL, USA, 2009, pp. 81-88.

[7] R. Angelova and S. Siersdorfer, "A neighborhood-based approach for clustering of linked document collections", in Proc. CIKM Conf., New York, NY, USA, 2006, pp. 778–779.

[8] C. C. Aggarwal and P. S. Yu, "A framework for clustering massive text and categorical data streams", in Proc. SIAM Conf. Data Mining, 2006, pp. 477-481.

[9] S. Zhong, "Efficient streaming text clustering", Neural Netw., vol. 18, no. 5–6, pp. 790–798, 2005.

[10] Y. Zhao and G. Karypis, "Topic-driven clustering for document datasets", in Proc. SIAM Conf. Data Mining, 2005, pp. 358-369.

[11] C. C. Aggarwal, S. C. Gates, and P. S. Yu, "On using partial supervision for text categorization," IEEE Trans. Knowl. Data Eng.,vol. 16, no. 2, pp. 245–255, Feb. 2004.

[12] T. Liu, S. Liu, Z. Chen, and W.-Y. Ma, "An evaluation of feature selection for text clustering", in Proc. ICML Conf., Washington, DC, USA, 2003, pp. 488-495.

[13] M. Steinbach, G. Karypis, and V. Kumar, "A comparison of document clustering techniques," in Proc. Text Mining Workshop KDD, 2000, pp. 109–110.

[14] S. Guha, R. Rastogi, and K. Shim, "CURE: An efficient clustering algorithm for large databases", in Proc. ACM SIGMOD Conf., New York, NY, USA, 1998, pp. 73-84.

[15] H. Schutze and C. Silverstein, "Projections for efficient document clustering", in Proc. ACM SIGIR Conf., New York, NY, USA, 1997, pp. 74-81.

[16] D. Cutting, D. Karger, J. Pedersen, and J. Tukey, "Scatter/Gather: A cluster-based approach to browsing large document collections", in Proc. ACM SIGIR Conf., New York, NY, USA, 1992, pp. 318-329.