Paper ID: CSEIT08

ENHANCING TEXT MINING USING ONTOLOGY BASED SIMILARITY DISTANCE MEASURE

Atiya Kazi, Priyanka Bandagale
FAMT (Ratnagiri)

Abstract— Generally, Text mining applications 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. But, an optimized way is necessary enable the mining process, so that the side information is correctly utilized. The probabilistic approach for 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 the 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, each of them containing multiple word spaces, together with the word vector and the most similar words.

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 make 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 expressed in [4]. Their combinations play a major role in the methodology of the current work of interest.

The paper is organized as follows; related work speaks about several text mining techniques for clustering huge and spatial databases. This is followed by the proposed architecture with a brief working of existing and proposed algorithm. The next section talks about the results generated using different performance metrics such as precision and recall. Finally the conclusion part highlights how the proposed work enhances the quality, accuracy and effectiveness of the clustered output.

III. SYSTEM METHODOLOGY

There are three major modules which perform clustering and classification with ontology. The first is the Preprocessing module, the second one is clustering module and the last is the Classification module which works in tandem with DISCO ontology word-spaces. They complement each other to achieve the primary aim of knowledge gain from raw textual data and help in creation of more lucid clusters.

A. 0Preprocessing 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 Module

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.

IV. ALGORITHM WORKING

The algorithm used for mining using side-information is referred to as COATES, which stands for 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. On the other hand, the proposed algorithm for enhancing the clustering phase is an ontology based similarity distance measurement algorithm. It uses cosine-based distance calculation to find the similarity distance of two concepts denoted by C1 and C2. This algorithm described 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 wc, 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 one, indicating the rate at which the weight values decrease along the ontology hierarchy.

$$w[\mathrm{sub}\,(C1,C2)] = 1 + (1\ / k^{\mathrm{depth}C2}) \qquad (1)$$

The threshold for considering only relevant terms for further clustering procedure is calculated by the average distance value of 300 random words collected from the text documents and their synonyms collected from the respective word spaces using the cosine distance formula.

V. Result Analysis

In order to judge the performance of the proposed Ontological based mining approach, an application has been built for comparing the performances of data mining using side information and data mining using existing DISCO ontology. This type of mining finds distributional similar co-related words for the text contained in the data-set. Several experiments were conducted in the mining domain using test sets by performing initial pre-processing tasks such as stemming, tokenization, stop-word removal on the eight training sets. An example of the results generated is presented here. The user submits input conditions, such as the number of clusters (k) to be generated, to the mining application. The k output clusters are then grouped together according to their classification labels. The same process is repeated using the existing DISCO Ontology. In this approach, there is estimation of the cosine based similarity measure for two concepts without any k value input from the user. The estimated accuracy graph indicating the results, is shown in Fig. 1[17]. As seen, one can conclude that accuracy is higher for ontology based mining on an average of 0.2. On the other hand Fig. 2[17], shows that, Ontological based mining is more time consuming. The reason for this slowing down of the mining process is, Side-information mining is faster because it searches over common keywords based on their probabilities, as opposed to ontological similarity based distance search that searches the entire database by calculating word vectors. Ontological mining is also more precise since it incorporates domain knowledge discovered from existing DISCO Ontology. The actual values of the experiments, based on confusion matrix principles, for the eight datasets are revealed in Table I and II [17]. They show the highest gain for proposed algorithm.

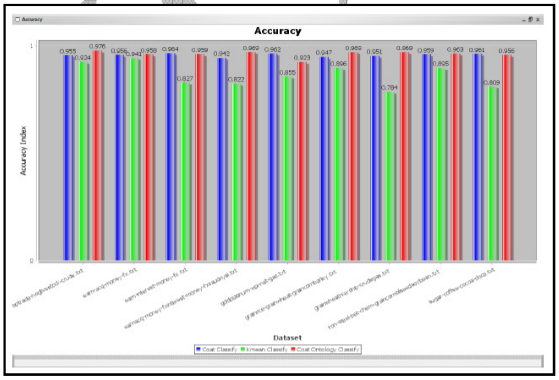

Figure 1. Accuracy Comparison for side-information and Ontology based clustering

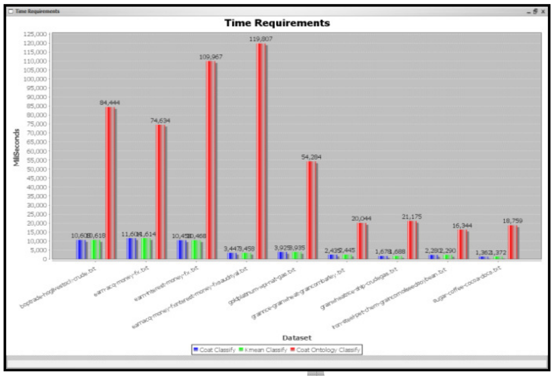

Figure 2. Time Requirement for side-information and ontology based approach

Table I. Accuracy comparison Table for eight datasets

Dataset Name	Accuracy-Side-Information	Accuracy-Ontology
Boptradehoglivestock- crude.txt	0.910756501	0.977124183
earn-acq-	0.905499889	0.942439024
earnacq-money-fxinterest-money-fxsaudriyal.txt	0.909304511	0.9359319
goldplatinum-wpi-nat- gas.txt	0.900083542	0.939327485
grainrice-grainwheat-graincornbarley.txt	0.910761155	0.953497942
grainwheatrice-ship- crudegas.txt	0.896700796	0.971380471
iron-steel-pet-chem-graincornoilseed soybean.txt	0.896117805	0.95308642
sugar-coffee-cocoa- docs.txt	0.894056848	0.951271186

Table II. Time-Requirement comparison Table for eight datasets

Dataset Name	Time-Side-Information (ms)	Time-Ontology(ms)
Boptradeho glivestock- 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 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. For future work, one can propose the idea of helping the naive user to acquire knowledge from the domain expert. The user will use the extracted knowledge as a guide in creating user defined Ontologies. The domain expert will corroborate the extracted knowledge, and retain information about the missed knowledge. One can also explore the tactic for building ontology from amorphous data such web pages and documents. This representation based, control based or domain specific ontology can tune the mining engine with the help of a Database expert.

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.

[17] Atiya Kazi, D.T.Kurian, Text Mining using Ontology based Similarity Measure, International journal of engineering and computer science, e-ISSN: 2319-7242, Vol.4, Issue 7, July-2015.