A Semantic Ontology-based Document Organizer to Cluster eLearning Documents

Sara Alaee
Electrical and Computer Engineering Department
University of Tehran
Tehran,Iran
sara.alaee@ut.ac.ir

Fattaneh Taghiyareh
Electrical and Computer Engineering Department
University of Tehran
Tehran,Iran
ftaghiyar@ut.ac.ir

***Abstract*—Document clustering is a useful technique to organize large sets of documents into meaningful groups. The usefulness is appreciated by labeling the clusters with relevant words that describe their associated documents. The traditional approach for document clustering, i.e. bag-of-words representation, often ignores the semantic relations between terms. Hence, ontology-based document clustering is proposed. In the context of e-Learning, the richer annotation of learning materials, via the use of appropriate ontologies is a way to deal with the reusability and remix of learning objects. Through providing a semantic infrastructure that will explicitly declare the semantics and relations between concepts used in labeling learning objects, the desired quality in the learning offer can be ensured. This paper proposes an ontology-based document clustering approach based on two-step clustering algorithm and compares its performance with the conventional clustering. Ontology is introduced through defining a weighting scheme that integrates traditional scheme, i.e. co-occurrences of words, with weights of relations between words in ontology. Our experimental evaluations are performed on ICVL (International Conference on Virtual Learning) paper collection as dataset with e-Learning domain ontology as the background knowledge. The ontology was implemented by us through a different research. The results show that inclusion of ontology increases the clustering quality.**

Index Terms—*Document Clustering; Ontology-based Clustering; eLearning; Ontology Generation; Semantic Relation; eLearning Concept*

I. INTRODUCTION

Increase in number of text documents requires efficient techniques to retrieve and search documents from large corpora. Document clustering is a method that organizes documents into meaningful groups such that all the documents in the same cluster have high similarity and the documents between clusters have low similarity [1]. Document clustering has applications in: search engines where relevant responses to user queries is important, personalized recommender systems where main focus is on the process of recognizing, accumulating and classifying information with respect to users' favorites or interests, and any organization or institution which requires efficient assortment of documents and storing them in large databases.

Traditional document clustering techniques were based on frequency and co-occurrence of words in a document [2]. This means that these techniques only consider the documents as bag of words, thus ignore the possible semantic relationships between the words. Since clustering is an unsupervised task, the quality of the results may not be fully optimal due to lack of guidance about which documents actually belong to the same category. To overcome this problem, new approaches on text clustering are mainly focused on identifying the background knowledge; tacit or explicit. Experts' opinions, wiki contents, structure, and links, search engines results and ontologies are additional knowledge used in text clustering process [3].

Ontology is an explicit representation of a set of concepts and their relationships. The main components of ontology are classes, attributes and relations. Classes represent concepts in broader sense. Attributes represent properties of each concept, and relations represent the association between concepts. Domain-specific ontologies are ontologies for a certain domain of interest. Particular terms or concepts applied to that domain are provided by domain ontology. Domain ontologies are one of the useful tools for text clustering.

This study will discuss a clustering methodology based on ontology. Our goal is to group specialized texts from the domain of e-Learning into meaningful groups. The clustering results would be valuable to many systems such as education systems, content management systems, recommender systems, etc. Even automatic classification of conference tracks would be possible using data from such clustering. Our methodology, if integrated into the learning management systems and content management systems, will result in personalization and reusability of learning contents. Given the relatively high cost of content generation, the classification, reusability and remix of learning objects leads to better management of available resources. Also, using the proposed scheme, the recommender modules in learning systems will be able to serve more appropriate results to users. Generally, our approach can be applied to any repository with archiving capability in which classification of contents based on users' needs is the main purpose. This will increase the retrieval efficiency and accessibility. While, our contribution in this paper is to develop and utilize an e-Learning domain ontology to organize the learning contents, the same approach can be applied to any other domain as well.

The e-Learning domain ontology was generated according to a large set of papers from a famous e-Learning conference. This ontology was then used to cluster documents (papers) from another e-Learning conference. It is worth noting that our proposed ontology is a dynamic structure which allows to constantly updating the directory of nodes and their corresponding relations with the introduction of new concepts. Using domain ontologies, the main concepts of a text (i.e. the concepts that exist in the ontology) can be identified and weighted based on their identity and/or synonym relationships in the ontology. Then a clustering algorithm is used to group these documents based on their similarity.

The structure of paper is as follows: In section 2 a review of the related works on document clustering is presented. Specifically, Section 2.1 proposes a brief summary of ontology generation approaches in the literature and, section 2.2 outlines a brief summary of related works on clustering based on ontologies. The proposed methodology is presented in section 3. Section 3.1 and section 3.2 lay out the ontology generation approach and ontology-based clustering approach, respectively. The experimental results is discussed in section 4. Section 4.1, 4.2 and 4.3 lay out the dataset, evaluation methodology and statistical analysis, respectively. The concluding remarks and future research directions are imparted in section 5.

II. RELATED WORKS

A. Ontology Generation

Ontology generation is the process of manual, automatic or semi-automatic creation of ontologies. Manual construction of ontologies demands a lot of time and effort while relying on advancements in machine learning techniques, automatic ontology construction have been much facilitated. However, the automatically constructed ontology could be too deficient since it relies only on pure data and not on human judgments, while the manually generated ontology is much more precise and reliable. This subsection will provide a brief overview of the (semi-) automatic techniques:

Bedini and Nguyen group ontology generation in four main categories [4]: 1. *Conversion* or *translation* which is only applicable for those applications where an ontology is already defined. Using this approach researchers produce softwares that transform common knowledge representations such as XML to ontology format. 2. *Mining-based* which implements mining techniques to retrieve information to produce ontology. Most such techniques are focused on processing unstructured resources like text documents or web pages through Natural Language Processing (NLP) methods. 3. *External Knowledge-based* which builds or enriches an ontology using an external resource. External resources include existing ontologies, external dictionaries or a general knowledge resource such as WordNet. 4. *Frameworks* which provide a platform with different modules to assist ontology generation. One of these frameworks is called Protégé. Protégé is a free, open source, platform to design ontologies, developed by the Stanford Medical Informatics group (SMI) at the University of Stanford. It is one of the most widely used platforms for ontology development and training.

Most of the (semi-)automatic approaches fall into the second category, i.e. Mining-based approach. Wong, Liu and Bennamoun classify the techniques used by these systems into four categories [5]: *Statistics-based*, *Linguistics-based*, *Logic-based* and *Hybrid*.

Statistics-based techniques do not consider the underlying semantics and relations between the components of a text. The main idea behind these techniques is that the co-occurrence of lexical units in text often provides a reliable estimate about their semantic identity. Common methods include Clustering [6], Latent Semantic Analysis [7], Co-occurrence Analysis [8], Term Subsumption [9], Contrastive Analysis [10], and Association Rule Mining [11]. In *clustering*, some similarity measures are needed to assign terms into groups for discovering concepts or constructing hierarchy. *Latent semantic analysis* and other approaches based on dimension-reduction techniques are applied on term-document matrices to overcome the problem of data sparseness in clustering. *Co-occurrence analysis* attempts to identify lexical units that occur together to discover implicit relations between concepts. In *term subsumption*, the conditional probabilities of the occurrence of terms in documents are calculated to discover hierarchical relations between them. The higher the subsumption value of a term, the more general it is with respect to another term. The extent to which two terms occur together in a document and in text corpora is employed for relevance analysis. One of the most common relevance measures in information retrieval includes Term Frequency-Inverse Document Frequency (TF-IDF) [12]. *Association rule mining* is applied to find the associations between the concepts at the appropriate level of abstraction.

Linguistics-based techniques are mainly dependent on natural language processing tools. Some of the techniques include part-of-speech tagging, sentence parsing, syntactic structure analysis, and dependency analysis. In *part of speech tagging* and *sentence parsing*, the syntactic and dependency structures are extracted and used for further linguistic analysis. Some of the famous parsers include Principar [13], Minipar [14], and Stanford Parser [15] which is placed under linguistic-based category but is also built on a statistical parsing system. *Syntactic* and *dependency analysis techniques* extract syntactic and dependency information from the text to uncover relations between terms in sentence level. In *syntactic analysis*, structures such as noun phrases, verb phrases and propositional phrases are analyzed to discover relations. In *dependency analysis*, grammatical relations such as subjects, objects and adjuncts are analyzed to discover more complex relations.

Logic-based techniques are the least common in ontology learning and are mainly adopted for complex tasks involving relations and axioms. Some of the techniques in this category include inductive logical programming [16], [17] and logical inference [18]. In *inductive logic programming* rules are extracted from the collection of concepts and their relations. The rules prove only the positive statements. In *logical inference*, hidden relations are extracted from the existing ones using transitivity or inheritance rules. The potential problem with this method is the possibility of introducing invalid or conflicting relations.

Finally, Hybrid techniques use the combination of the above-mentioned approaches for ontology learning. In reality, this technique is the most common. Shamsfard and Abdollahzadeh Barforoush proposed an automatic hybrid ontology building approach which starts from a small ontology kernel and then extends it through text processing [18]. Their introduced test bed is called Hasti. Hasti is an ontology learning system that learns lexical and ontological knowledge from Persian texts using a combination of logical, linguistic and semantic analysis methods.

B. Clustering based on ontology

Most related literature to ontology-based clustering focus on defining efficient weighting schemes.

Yang, et al. implements word clustering by calculating word relativity and defines some factors, i.e. property similarity, semantic distance and hierarchy path to measure similarity [19]. These similarity values are then used to cluster words and their corresponding documents into appropriate groups. Sureka and Punitha calculate concepts weights by considering the correlation coefficient of the words, i.e. the presence and absence of the words in ontology, and probability of the concept in the document [20]. The system ranks the concepts and selects the ones with bigger weights. These concepts are then used for clustering using KMeans [21] and DBScan algorithms [22].

Logeswari and Premalatha introduces another weighting scheme which considers the semantic aggregation of all concepts close to one concept in the ontology [23]. Four types of semantic relationships are defined: identity, synonym, hypernymy and meronymy. The number of relations a word has with other words in the ontology and also the weight of each relation is considered when calculating the weight. KMeans clustering algorithm is then applied to the weights. Zhang and et al. use three types of semantic similarity measures with a term reweighting method [25], [26] to cluster documents from a Medical dataset [24]. Term reweighting includes discounting general key words which do not belong to the domain and emphasizing on core words. Through this scheme, the weights of general words do not change considerably but the weights of semantically related concepts are increased. Each term's weight will be first represented by a certain value such as TF-IDF. The three types of semantic similarity measures include: path-based [27], [28], [29], [30] information content-based [31], [25] and feature-based [32]. Path-based similarity measure utilizes the information of the shortest path between two concepts. Information content-based measure associates probabilities of concepts in ontology to the information they contain. The more information two terms share, the more similar they are. Feature-based measure describes each term by a set of terms assumed to be its features. For example, all ancestor nodes of a certain concept might be regarded as its feature set. The more common features two terms have, the more similar they are.

III. METHODOLOGY

In this section, we present a framework for clustering eLearning documents based on its domain ontology. Our approach takes the document collections as input and provides the corresponding clusters as output. Fig. 1 shows the proposed approach.

The framework starts with two sets of papers as inputs: One for creating ontology and the other for clustering using the generated ontology. To prevent overtraining, we selected different datasets for each one. The clustering dataset could also be used to further refine the ontology.

It should be noted that our datasets comprise papers with various formats and structures. Therefore, we have collections of unstructured text data. To avoid unstructured text processing which is beyond the scope of our research, we only considered titles and abstracts as input data. Considering the fact that abstracts are the essence of the papers and draw information from all of the other sections of the paper, while titles are compact versions of abstracts, we believe we have obtained all the important information from the papers through these parts.

A. Ontology Generation

In e-Learning realm, like many other fields of research, ontology can easily manage the knowledge domain and allow a more detailed organization of the concepts. Since the introduction of semantic web and its various tools and technologies several research projects has been conducted to develop ontologies for different fields, such as biomedicine. However, an expressive well-defined ontology for many other fields, such as e-Learning, is still nonexistent. In our attempt to resolve this matter, we created an automatic e-Learning ontology using the specialized texts and documents from this domain. Domain ontology is built through the following steps:

Step1: Preprocessing

In this step, all textual data from the documents that are useful for mining procedures will be extracted. The objective is to obtain key terms or features that will best describe each document. The general procedures include: case folding, removing stop-words, stemming the key words, and merging synonyms and complex words.

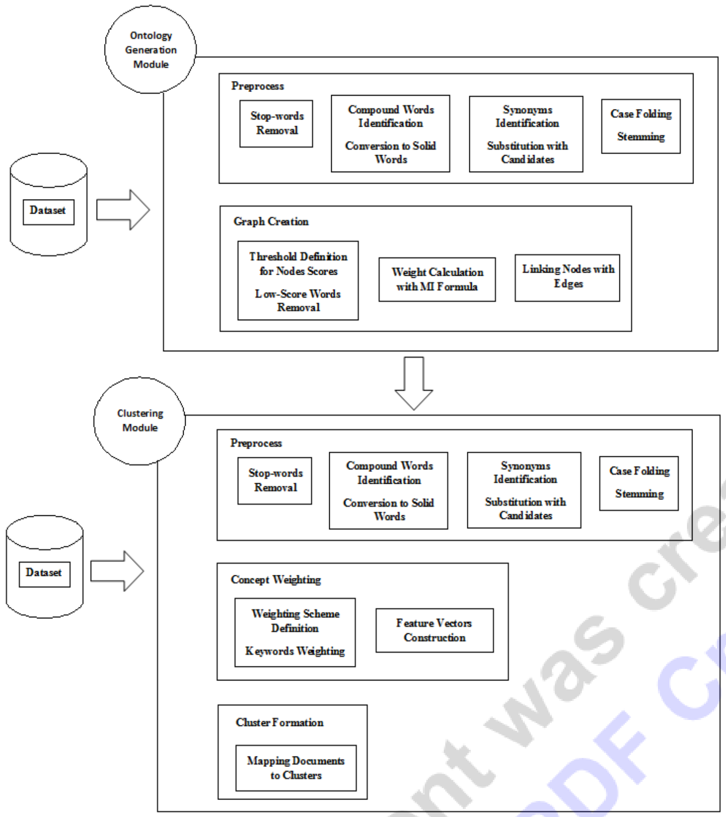
Ontology Generation Module
Dataset
Preprocess
Stop-words Removal
Compound Words Identification
Conversion to Solid Words
Synonyms Identification
Substitution with Candidates
Case Folding
Stemming
Graph Creation
Threshold Definition for Nodes Scores
Low-Score Words Removal
Weight Calculation with MI Formula
Linking Nodes with Edges
Clustering Module
Dataset
Preprocess
Stop-words Removal
Compound Words Identification
Conversion to Solid Words
Synonyms Identification
Substitution with Candidates
Case Folding
Stemming
Concept Weighting
Weighting Scheme Definition
Keywords Weighting
Feature Vectors Construction
Cluster Formation
Mapping Documents to Clusters

Fig. 1. Ontology-based Document Clustering Framework

a. In case folding, all characters of the words will be converted to either lower or upper case. Here, we converted them to lower case.

b. A document may contain many unnecessary words which are irrelevant to the main subject, namely verbs (such as "is", "am" or many other verbs such as "provide", "introduce", etc.), identifiers (such as "the", "a", …) and propositions such as "at", "in", "on", etc. These are called stop-words. We removed them using a list of common English words that we had prepared.

c. Stemming is the process for reducing words to their root forms. This is necessary since words with the same roots should be considered as same words in our proposed method. There are several algorithms for stemming English words. We used Porter's Stemmer Algorithm [33].

d. Semantic importance of nodes and their corresponding relations in ontology are represented through scores and weights we assign to them. In order to correctly calculate these scores, we need to first unify all similar words together. This is performed by manually removing hyphens or spaces between each of words that are to be considered as part of a complex word and representing them as one. For example, "e-Learning" and "eLearning" are the same words. Yet with a hyphen, an automatic tool might recognize them as different words. Therefore, they must be unified. It is worth noting that unification process could also be performed automatically (using techniques such as association rule mining). However for the following reasons we did it manually (i.e. by expert judgments):

- Different representations of one word. As mentioned earlier, similar words might be represented in different ways, for example "e-Learning" might be written as. "eLearning",
- Word summarization. For example, "eLearning" vs. "electronic learning",
- Word Omitting. Some words might be omitted from the complex word. For example, "educational data mining" might be written as "educational mining",
- Corpus size limit. To obtain reasonable confidence and support, we need a huge corpus.

We also merged synonyms by replacing words that have the same meaning (such as "teach" and "train") or are used in place of each other (such as "student" and "learner") with the same candidate.

Step 2: Creating graph

If we assume ontology as a graph of words, each concept would be represented as a node and the corresponding relations with other words would be the edges. Fig. 2 shows a representation of such graph.

All words obtained from the previous step are potentially regarded as key words representing the core concepts of the documents. However, we only placed some of them, i.e. title words, as nodes in our ontology. This is because title words are semantically more related to the domain and thus better represent subjects of the documents as whole. Other words are either low frequent (e.g. newly introduced concepts) or technically nonrelated to the base ontology.

As mentioned earlier, edge weights represent relations between nodes in ontology. The normalized frequency with which each pair of words in the titles appears in the abstracts would be regarded as the weight of the edge between them. In other words, if two nodes x_i and x_j have mutual information, M_{ij}, based on Eq. 1, then they will have an edge with weight equal to M_{ij}:

$$M_{ij} = \frac{f(x_i, x_j)}{f(x_i) * f(x_j)} \qquad (1)$$

In this equation, x_i is a node in the ontology, $f(x_i)$ is the frequency of occurrences of x_i in all abstracts, and $f(x_i, x_j)$ is the frequency of occurrences of x_i and x_j together in all abstracts.

B. *Ontology-based Clustering*

Our proposed document clustering approach combines concept weighting and a clustering method, namely two-step clustering [34]. The system performs clustering in three steps: preprocessing, concept weighting and clustering based on concept weights. The followings describe each of these steps.

Step1: Preprocessing

Each document is represented by its key words which are then used for clustering the document. Therefore, this part is similar to the one described in the ontology generation module (steps a, b, c and d).

Step2: Concept Weighting

Like the Ontology Generation Module, words obtained from the preprocessing step are the key words representing the core concepts of the document. To perform clustering, these words should be assigned weights and thus each document should be converted to a vector of key word weights. In this research, we used TF-IDF along with some information from the domain ontology (the e-Learning domain ontology, since our test set is an e-Learning conference dataset) to define the weighting scheme. The reason to use ontology is that TF-IDF only considers the frequency of words while other factors, such as relationships between words and semantics, might be ignored.

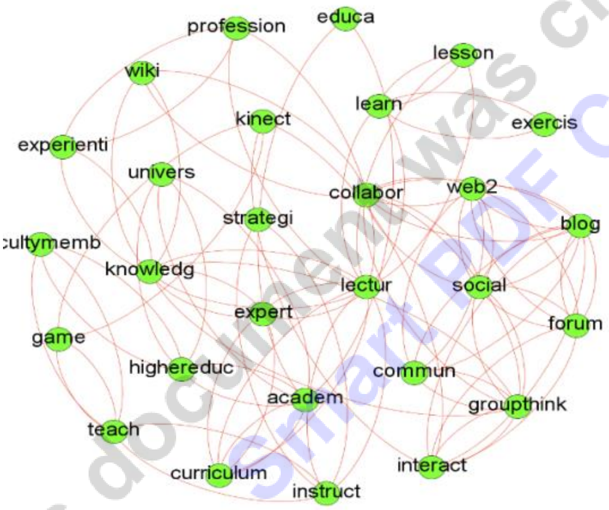

Fig. 2. Ontology Graph Representation

The proposed weighting scheme is defined as follows:

$$W_i' = W_i + \sum_j \left[-Log_{10}\left(E_{ij}\right) * W_j \right] \tag{2}$$

$$W_i = \frac{f_i}{N_d} * log_{10}^{\left(\sum_{d,i} \frac{N_d}{df_i}\right)} \quad ; \quad df_i = No. of\ documents\ that\ contain\ i \tag{3}$$

Where W_i' is the weight of word i after reweighting by ontology, W_i is the value of TF-IDF for word i (i.e. the weight of word i before reweighting), E_{ij} is the weight of the edge from i to j in the ontology (i.e. the mutual information between these two words in the proposed ontology). If there's no edge between two words in the test set and their corresponding nodes in the ontology, then the second part would be zero and only TF-IDF would be considered. This is reasonable since words that appear in the ontology are semantically more important than other words in the document and thus should have more effect in clustering that document. We used logarithm to increase the effect of ontology weights on the final weights.

Step3: Clustering based on concept weights

To cluster documents we used Two-step clustering [34]. The algorithm is based on a two-stage approach. In the first stage, a procedure similar to K-means is applied on the input data. The inputs are vectors of concept weights calculated using Eq. 2 in step 2. In the second stage, a hierarchical agglomerative clustering procedure is conducted on pre-clusters to form homogeneous clusters.

IV. Experimental Resutls

A. Dataset

Since our domain of interest is e-Learning, we selected papers from an international e-Learning conference, namely ICALT (International Conference on Advanced Learning Technologies), as our desired data set for generating base ontology. We also used papers from another international conference, namely ICVL (International Conference on Virtual Learning) for the purpose of clustering. ICALT repository includes 1270 papers published between years 2009 and 2014. ICVL repository includes 118 papers published between years 2012 and 2013.

B. Evaluation Methodology

Clustering quality for two approaches, i.e. base-line two-step clustering and ontology-based two-step clustering, was compared using ICVL's 2 years published papers. (Base-line is the representation with TF-IDF weighting and without considering the semantics.) The performance was analyzed using the normalized precision, recall and F-measure. The equation used to calculate precision and recall are given in Eq. 4 and Eq. 5:

$$Precision_i = \frac{N_i - \sum_j N_{ij}}{N_i} \tag{4}$$

$$Recall_i = \frac{N_i - \sum_j N_{ij}}{N_i - \sum_j N_{ij} + \sum_k N_{ki}} \tag{5}$$

Where $Precision_i$ is the precision value for cluster i, N_i is the total number of objects in cluster i, N_{ij} is the number of objects related to cluster j but wrongly placed in cluster i, and N_{ki} is the number of objects related to cluster i and wrongly placed in cluster k.

The normalized precision and recall is calculated using Eq. 6 and Eq. 7:

$$Normalized\ Precision = \sum_i \frac{N_i}{N} * Precision_i \tag{6}$$

$$Normalized\ Recall = \sum_i \frac{N_i}{N} * Recall_i \tag{7}$$

N is the total number of items. The F-measure is calculated using Eq. 8:

$$F - measure_{\ i} = 2 * \frac{Precision_i * Recall_i}{Precision_i + Recall_i} \tag{8}$$

According to the previous formula, the normalized F-measure is calculated using Eq. 9:

$$Normalized\ F - measure = \sum_i \frac{N_i}{N} * F - measure_{\ i} \tag{9}$$

The F-measure is the harmonic mean of precision and recall measures with equal weights.

C. *Results Analysis*

This section presents the experimental results of document clustering on ICVL dataset. OntTS and TFIDFTS refer to the ontology-based two-step clustering and TF-IDF-based two-step clustering algorithms, respectively. The values obtained for these three measures from both algorithms are shown in Table I:

TABLE I. DOCUMENT CLUSTERING RESULTS WITH AND WITHOUT ONTOLOGY

DataSet	**Precision**		**Recall**		**F-measure**	
ICVL 2-years papers	***OntTS***	***TFIDFTS***	***OntTS***	***TFIDFTS***	***OntTS***	***TFIDFTS***
	0.63	0.51	0.66	0.60	0.64	0.54

From the experimental results shown in Table I, it is obvious that the ontology-based clustering algorithm shows better performance. Generally, we could say that the clustering approach that uses the semantics of the documents for term weighting produces better results than the approach without semantics. The low values of precision, recall and thus F-measure from the above table might indicate that the reweighting scheme, as a method of integrating domain ontology to clustering, is not very effective. However, since our experiment was performed on a specific research domain and effective clustering of its documents requires going into more details about concepts' meanings and relations, it seems reasonable that we did not obtain very high results. This is besides the fact that overall accuracy in the baseline method was also relatively low. It is expected that if the same procedure is performed on the documents from a general domain (e.g. news articles) using a general-purpose ontology, the results would show better performance.

This research was only focused on a simple application of domain ontology for the purpose of clustering. The clustering algorithm was also based on a simple two-step methodology. Using more advanced procedures for applying ontology on clustering will further enhance the results. This will be our future work.

V. CONCLUSION AND FUTURE WORKS

This paper introduced an ontology-based approach for e-Learning documents clustering. The proposed method used term re-weighting and a clustering algorithm, namely two-step algorithm on ICVL papers set. We first generated an ontology based on documents from a famous international e-Learning conference (i.e. ICALT). Then in the clustering procedure, after stop words removal, stemming and merging synonyms and complex words unification, concept weights were calculated by taking into consideration the TF-IDF of the words and the corresponding weights of relations between each word and the other words in ontology. The term weights vectors for these documents were then clustered using the clustering algorithm. The experimental results showed that re-weighting has positive effects on document clustering. In detail, we found that term re-weighting based on ontology will enhance precision, recall and f-measure values on our desired dataset. Applying the same procedure on a general-domain corpus is expected to result in more significant improvements. We believe that integrating the proposed framework with learning and content management systems will help enrich the user's experience by avoiding redundant or repetitive contents. Therefore, future work would involve combining these systems with the current framework. Other future directions include finding methods that combine different features and semantics from the domain ontology with more advanced techniques for clustering, refining the current base ontology with regard to the existing papers, and extending the ontology by incorporating concepts from other domains.

REFERENCES

[1] A. Shawkat Ali, "K-means Clustering Adopting RBF-Kernel," in Data Mining and Knowledge Discovery Technologies, 2008, pp. 118-142.

[2] C. C. Aggarwal and C. Zhai, "A Survey of Text Clustering Algorithms," in Mining Text Data, Springer US, 2012, pp. 77-128.

[3] C. C. Aggarwal and C. X. Zhai, Mining Text Data, Dordrecht Heidelberg London : Springer Publishing Company, 2012.

[4] I. Bedini and B. Nguyen, "Automatic Ontology Generation: State of the Art," PRiSM Laboratory Technical Report, University of Versailles, Versailles, 2007.

[5] W. Wong, W. Liu and M. Bennamoun, "Ontology learning from text: A look back and into the future," ACM Computing Surveys (CSUR), vol. 44, no. 4, August 2012.

[6] W. Wong, W. Liu and M. Bennamoun, "Tree-Traversing Ant Algorithm for Term Clustering based on Featureless Similarities," Data Mining and Knowledge Discovery, vol. 15, no. 3, pp. 349-381, 2007.

[7] P. D. Turney, "Mining the Web for Synonyms: PMI-IR versus LSA on TOEFL," in Machine Learning: ECML 2001, Springer Berlin Heidelberg, 2001, pp. 491-502.

[8] A. Budanitsky, "Lexical Semantic Relatedness and Its Application in Natural Language Processing," Department of Computer Science; University ofToronto, 1999.

[9] H. Fotzo and P. Gallinari, "earning GeneralizationSpecialization Relations between Concepts - Application for Automatically Building Thematic Document Hierarchies," in Proceedings of the 7th International Conference on Computer-Assisted Information Retrieval (RIAO) , 2004.

[10] P. elardi, R. Navigli, A. Cucchiarelli and F. and Neri, "Evaluation of OntoLearn, a methodology for automatic learning of ontologies," in Ontology Learning from Text: Methods, Evaluation and Applications, IOS Press, Hershay, PA., 2005.

[11] R. Srikant and R. Agrawal, "Mining generalized