ISBN 978-981-14-1455-8
Proceedings of 2019 The 11th International Conference on Future Computer and Communication
WCSE_2019_SPRING
Yangon, Myanmar, February 27-March 1, 2019, pp. 35-41

Impressive Approach for Documents Clustering Using Semantics Relations in Feature Extraction

Wai Wai Lwin[1+]

University of Computer Studies, Yangon

Abstract. The Internet or World Wide Web (WWW) is awful spread today, therefore to navigate, summarize, and organize information effectively fast and high-quality web document clustering algorithms play an important role. In this area, dimensionality reduction and semantic relations are also of fantastic influence step in the data mining process. For computing the document similarity, it is used vector-space-model that represents several features present in document. In general, it cannot account for the words (noun) such as names of the people, countries and items. as features. They almost are ignored as irrelevant attributes. But some of these irrelevant terms are valuable in specific domain. Moreover traditional feature representation is not able to reflect the semantic contents of a document because of the synonym problem and polysemy problem. Motivation of these reasons, we proposed the domain ontology which represents the semantic relations of specific terms and semantic words like lexical database. It can increase the process of extraction of features in specific documents and reduce the dimensionality. As a result, the calculation of similarity measure will be more definite, and enhancing in the segmentation between clusters. In this paper, we tested the proposed method in documents clustering area with Particle Swarm Optimization (PSO) document clustering algorithm that performs a globalized search in the entire solution space. The proposed method can support the efficient clustering approach for document clustering of PSO algorithm using semantic relation in features extraction.

Keywords: Data Mining, Features Extraction, Semantic Relation, Ontology, PSO

1. Introduction

Data mining is the process of discovering interesting patterns and knowledge from huge amount of data. This process is actually performed by automatically or semi-automatically which uncovered the hidden patterns and relationships in data, and it can be used to make predictions that impact businesses.

Data mining, also popularly referred to as knowledge discovery from data (KDD), is the automated or convenient extraction of patterns representing knowledge implicitly stored or captured in large databases, data warehouses, the Web, other massive information repositories or data streams.

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.

Most traditional document clustering techniques were depended on frequency and co-occurrence of terms/words in a document. As a means that these techniques only consider the documents as bag of words, hence they don't considered the possible semantic relationships between the terms/words. The "bag of words"

[+] Corresponding author.
E-mail address: wai2020.smile@gmail.com

feature representation is not competent to imitate the semantic content of a document because of the synonym problem and polysemy problem [1].

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 [2].

Moreover sometimes utmost information can depreciate the potency of data mining. Some of the data attributes assembled for building and testing a model may not grant significant information to the model and take away from the quality and accuracy of model. To surmount these, many attributes in data sets are grouped as collection of attributes that are correlated to actually be measuring the same underlying feature. Additionally, we devote the feature extraction with semantic relation approach to minimize the effects of noise of improper attributes, correlation, and high dimensionality. Although irrelevant attributes add noise to the data and affect model accuracy, sometime some of them may contribute as meaningful information. For example, the names of players, managers and clubs in sports news significant affect to cluster the sports news group. As a result, these are applied likes the hyponyms of specific domain ontology of our proposed system. What's more, we apply the synonym, hyponym and Polysemy of semantic relationships between words, phrases and sentences.

In this paper, we present a semantic relation approach that using ontology and Particle Swarm Optimization (PSO) algorithm to cluster the news of News Websites. We developed domain ontology in order to represent generic knowledge about a target world which defines specific terms and semantic words like lexical database of related news documents for effective web documents clustering. As a consequence of the domain ontology for representing the terms with enriched semantic relation and synonyms, clustering of the documents using PSO clustering technique would achieve the preferred accomplishment, filter web documents precisely, and a better performance for inter and intra cluster similarity.

The rest of the paper is presented as follows. In section 2, it represents the related work of the system. Section 3 expresses the architecture of the system. Section 4 describes the methodology in the proposed system. Section 5 represents the discussion about the analysis results of the system. Section 6 describes the evaluation on the outcomes (clusters). Finally, section 7 is the conclusion and future work of the proposed system.

2. Related Works

In [3], they proposed ontological model to represent a list of clinical diagnoses from the data of signs, symptoms, risk factors and medical background. In this respect, they did not consider the semantic representation of medical knowledge to perform clinical diagnostic process.

In [4], the authors proposed a semantic similarity based model to pick up the semantic of the text. This proposed model in coincident with lexical ontology solves the synonyms and hypernyms problems. The proposed model use semantic weights added to the term frequency weight to calculate the semantic similarity between terms. Whereas, there are no support ontology that present the relation of semantic of terms such as names of people, countries, and items.

In [5, 6], authors introduced conceptual features in text representation. A concept feature is an aggregation of a few words that describe the same high level concept. They proposed three methods (i) adding concept features to the term space (i.e., term+concept); (ii) replacing the related terms with concept features and (iii) reducing the VSM to only concept features. For text clustering, experimental results [4] showed that only the term+concept representation improved clustering performance. But they all ignored the enhancing method of the features extraction.

In [7], authors presented the important of features selection and features extraction in data mining. They only described to select a particular feature from huge set of features that are residing within the dataset. Although there did not present how to improve the extraction of features in documents clustering.

3. Architecture of Proposed System

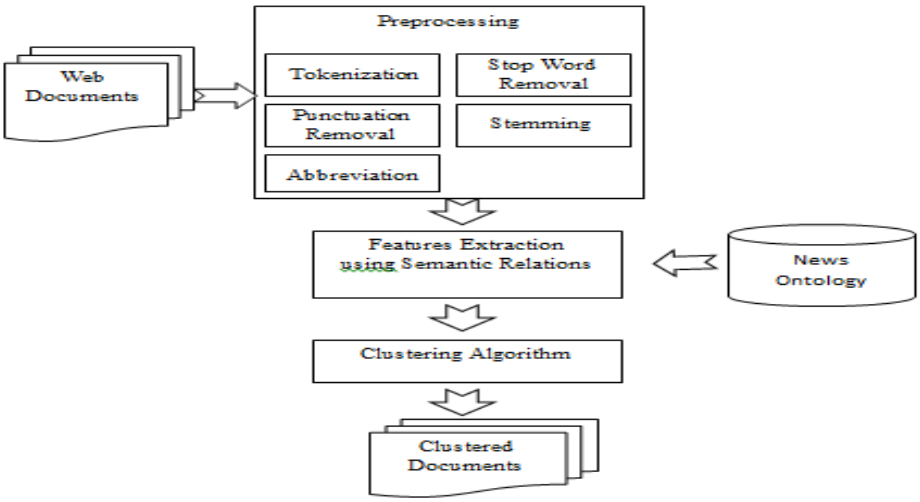


Fig 1: Architecture of Web Document Clustering using Domain Ontology

The first state of the proposed system is started with preprocessing. And then, the system performs generating features extraction using semantic relations in news ontology for document dataset to apply clustering algorithms on web document dataset. So proper features extraction of document based on Ontology is extremely important that they can be reduced the dimensionality of documents. The Fig: 1 describes the architecture of the system.

4. Methodology of the Proposed System

4.1 Feature Extraction

Data available for mining is raw data with many dimensions. Data comes from different sources with different formats and, it may consist of noisy data, irrelevant attributes, missing data etc. For data mining algorithms, wide, or unstacked data usually gives challenges in processing. Model attributes are the dimensions of the processing spaces. The higher dimensionality of the processing space makes the higher the computation cost in algorithmic processing. Dimensionality reduction is a very important step in the data mining process.

Irrelevant attributes cause noise to the data and directly influence the model accuracy. To minimize the effects of noise, correlation, and high dimensionality, dimension reduction is sometimes a desirable preprocessing step for data mining. Two approaches to dimension reduction are feature selection and extraction. **Feature selection** refers to the process of selecting the most relevant attributes and **Feature extraction** is combining attributes into a new reduced set of features

Feature Extraction is a process of an attribute reduction. Feature selection selects and returns the most significant attributes, whereas feature Extraction actually creates the new attributes, or features. The new attributes, or **features**, are linear combinations of the original attributes.

The Feature Extraction process causes in a much smaller and richer set of attributes. The maximum number of features can be user-specified or determined by the algorithm. By default, the algorithm determines it.

Since fewer and more meaningful attributes describe the data, models built on extracted features can be of higher quality. Feature Extraction intends a data set with higher dimensionality onto a smaller number of dimensions. It is useful for data visualization, because a complex data set can be effectively visualized when it is reduced to two or three dimensions.

Feature Extraction can be used to extract the features of a document collection, where documents are represented by a set of key words and their frequencies. Each feature is represented by a combination of keywords. The documents in the collection can then be expressed in terms of the discovered features.

4.2 Semantic Relationships

Semantic relationships are the associations that there exist between the meanings of words (semantic relationships at word level), between the meanings of phrases, or between the meanings of sentences (semantic relationships at phrase or sentence level). In this paper, we will discuss semantic relationships at word level as following: synonymy, hyponymy, and polysemy.

4.2.1 Synonymy

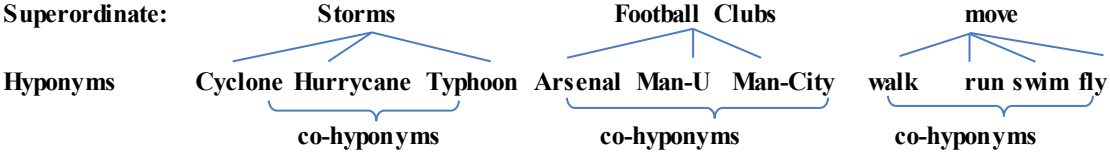
Synonymy is the semantic relationship that subsists within two (or more) words that have the same (or nearly the same) meaning and belong to the same part of speech, but are spelled differently. In other words, we can say that synonymy is the semantic exchangeability between lexical items. The (pairs of) words that have this kind of semantic relationship are called synonyms, or are said to be synonymous. [8]

E.g.	**football = soccer**	**match = game=competition**	**player = member**
	club = team	**to begin = to start**	**kind = courteous**
	beginning = start	**to cease = to stop**	**fast = quickly = rapidly**

4.2.2 Hyponymy

Hyponymy ([ÆhaI"pÓAn ́mi] or [hI"pÓAn ́mi]) or inclusion is the semantic relationship that subsists within two (or more) words in subway that the meaning of one word includes (or contains) the meaning of other words(s). We mean that the term whose meaning is included in the meaning of the other term(s) is the general term; linguists usually recognize it as a superordinate or hypernym. The term whose meaning includes the meaning of the other term is the specific term; linguists usually refer to it as a hyponym. If the meaning of a superordinate term is included in the meaning of several other more specific words, the set of specific terms which are hyponyms of the same superordinate term and are called co-hyponyms (cf. Crystal, 1991). [8]

E.g.



4.2.3 Polysemy

Polysemy ([p ‘"lIs ́mi]) is the semantic relationship that exists between a word and its multiple conceptually and historically related meanings (cf. Crystal, 1991; Fromkin & Rodman, 1998; Richards et al., 1992).

E.g.	**foot =**	**1. part of body;**	**2. lower part of something**	
	plain =	**1. clear;**	**2. unadorned;**	**3. obvious.**
	nice =	**1. pleasant;**	**2. kind;**	**3. friendly; etc.**

The different meanings of a word are not interchangeable; in fact, they are context-specific. [8]

4.3 Ontology

Ontology is more than a rule base of terminological problems and is worth to consider a promising methodology in the Semantic Web. In the semantic web, ontologies can be used to encode meaning into a web page, so that intelligent agents can understand what the web page is about. Ontologies increase the efficiency and consistency of describing resources, by enabling more sophisticated functionalities in development of knowledge management and information retrieval applications.[9]

Ontology is a formal description of knowledge as a set of concepts within a domain and the relationships that hold between them. To capacitate such a description, we need to formally specify components such as individuals (instances of objects), classes, attributes and relations as well as restrictions, rules and axioms. In this paper, we developed domain ontology in order to represent generic knowledge about a target world which defines specific terms and semantic words like lexical database of related news documents for effective documents clustering. Fig 2: presents the domain ontology of proposed system.

4.3.1 Domain Ontology of Proposed System

In this system, we generate domain ontology which actually defined attributes or features as combinations of the original attributes by using its semantic relations and properties. Some redundant or irrelevant attributes of documents which are valuable for specific domain are presented as instances. Therefore they transformed to relevant features by the accomplishing of semantic relation. In order to capability of ontology it can abstract the features more effectively and efficiently.

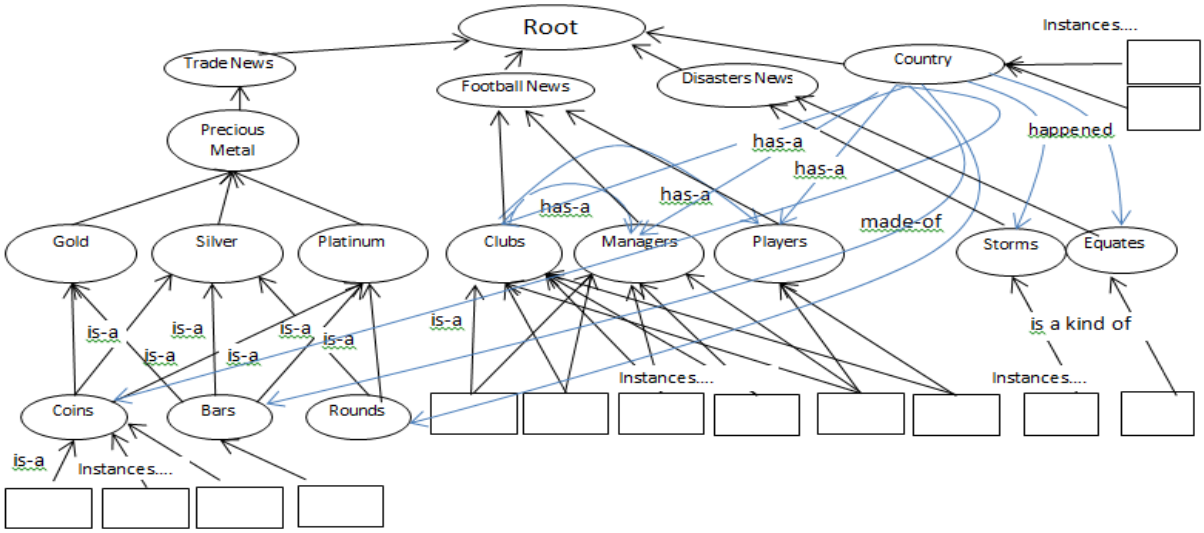


Fig 2: Domain Ontology of Proposed System

4.4. Clustering Algorithm

Document clustering is largely used in text mining and information retrieval and many fields. Clustering involves dividing a set of objects into a specified number of clusters, so that objects that are similar to other objects located in the same cluster. There are two major styles of clustering techniques: “Partitioning” and “Hierarchical” [8]. Particle Swarm Optimization (PSO) is another computational intelligence method.

PSO is a population based stochastic optimization technique that can be used to find an optimal, or near optimal, solution to a numerical and qualitative problem. A problem space in PSO has as many dimensions as needed to model the real problem space. A particle’s location in the multi-dimensional problem space represents one solution for the problem. When a particle moves to a new location, a different problem solution is generated. This solution is evaluated by a fitness function that provides a quantitative value of the solution’s utility. A particle will remember its current coordinates, its velocity that indicates the speed of its movement along the dimensions in a problem space, the best fitness value received so far, and the coordinates where these values were computed. It is this personal best value combined with its neighbor's best value that influences the movement of each particle through a problem space. [10]

5. Experimental Results

We generated dataset by collecting from Amazon.com, Alibaba.com, BBC.com, channelnewsasia.com, Goldprice.org, instaforex.com, skysports.com, premierleague.com and 24hgold.com. The system is used 100 test pages of the websites containing different data from different domains such as literature, media, business etc.

The results show that a domain Ontology using features extraction with semantic relation can fully support to decrease ambiguity to terms. And also the reasonable representation of documents leads PSO clustering algorithm to conduct an appropriate globalized searching for the optimal clustering.

In this ongoing study, the system will be obtained that Intra-cluster similarities (i.e. the distance between data vectors within a cluster) could be maximized and Inter-cluster similarity (i.e. the distance between the centroids of the clusters) could be minimized. The test result shows in the table 1.

Table 1: The result of using proposed system

Test Data Set	Test Method	Algorithm	Intra Cluster	Inter Cluster
100	Based on Term	Proposed System	83%	6.83%

6. Evaluation of cluster

There are numerous evaluation measures to validate the cluster quality. To evaluate the clustering results of proposed system, precision, recall, and F-measure has been used.

- **Recall:** ability of a classification model to identify all relevant instances
- **Precision:** ability of a classification model to return only relevant instances
- **F1 score:** single metric that combines recall and precision using the harmonic mean

For cluster *j* and cluster *i*:

$$\text{Recall}(i,j) = n_{ij}/n_i \quad (1)$$

$$\text{Precision}(i,j) = n_{ij}/n_j \quad (2)$$

where n_{ij} is the number of members of class *i* in cluster *j*, n_j is the number of members of cluster *j* and n_i is the number of members of class *i*.

Firstly, we defined the confusion matrix which is useful for quickly calculating precision and recall given the predicted labels from a model. A confusion matrix for binary classification shows the four different outcomes: true positive, false positive, true negative, and false negative. The actual values form the columns, and the predicted values (labels) form the rows. The intersection of the rows and columns show one of the four outcomes. For example, if we predict a data point is positive, but it actually is negative, this is a false positive.

		Predicted	
		Negative	**Positive**
Actual	**Negative**	TN: 3539	FP: 12
	Positive	FN: 264	TP: 1129

F-measure is computed using precision and recall as below:

$$F(i,j) = \frac{2 * recall(i,j) * precision(i,j)}{(precision(i,j) + recall(i,j)} \quad (3)$$

The result of the F-values for number of clusters 4 with proposed algorithm is 0.87.

7. Conclusion

In this paper, we show the experimental results of proposed system that uses domain Ontology which apply the semantic relation in feature extraction step, and Swarm Optimization (PSO) algorithm. According to the result, the performance of intra cluster similarity is high acceptability. Moreover, it is found that the inter cluster similarity accomplishes the proper results also. The proposed system gives the satisfactory result of F-measure value on News Group dataset (100 documents).

The future work of the system will upgrade the Ontology for specific domain and make force to grant the influential algorithm for clustering web document data.

8. Acknowledgements

Firstly, I would like to express my sincere gratitude to my rector Dr. Mie Mie Thet Thwin, at the University of Computer Studies, Yangon.

I would also like to acknowledge Dr. Nang Saing Moon Kham, the head of the Faculty of Information Science at University of Computer Studies, Yangon who shared me knowledge of data mining and I am gratefully indebted to her for her very valuable comments on this research. What's more, I would like to thank all "anonymous" reviewers for their so-called insights.

Finally, I must express my very profound gratitude to my colleagues for providing me with encouragement through the process of researching and writing this paper.

9. 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. X. Zhai, "*Mining Text Data*", Dordrecht Heidelberg London : Springer Publishing Company, 2012.

[3] Jos éA. Reyes-Ortiz, Ana L. Jiménez, Jaime Cater, Cesar A. Meléndez, et al, "*Ontology-based Knowledge Representation for Supporting Medical Decisions*", Research in Computing Science 68 (2013)

[4] W. K. Gad and M. S. Kamel, "Enhancing Text Clustering Performance Using Semantic Similarity", ICEIS, LNBIP 24, pp. 325–335, 2009

[5] A. Hotho, S. Staab, G. Stumme, "*Wordnet improves text document clustering*". In Proceedings of the semantic web workshop at 26th annual international ACM SIGIR conference, Toronto, Canada, 2003.

[6] A. Hotho, S. Bloehdorn "*Text classification by boosting weak learners based on terms and concepts"*. In Proceedings of the IEEE international conference on data mining, Brighton, UK, pp 7279,2004

[7] Aparna U.R., Shaiju Paul, Dept. Of Computer Science Engineering, Jyothi Engineering College, Kerala, India, *"Feature selection and extraction in data mining",* 2016 Online International Conference on Green Engineering and Technologies (IC-GET) , 19-19 Nov. 2016

[8] Prof. Argenis A. Zapata, Universidad de Los Andes, Facultad de Humanidades y Educación, Escuela de Idiomas Modernos, Inglés IV (B-2008)

[9] Gopinath Ganapathy and S. Sagayaraj , "Automatic Ontology Creation by Extracting Metadata from the Source code"

[10] Xiaohui Cui, Thomas E. Potok , "Document Clustering using Particle Swarm Optimization"