2011 International Conference on Information Communication and Management
IPCSIT vol.16 (2011) © (2011) IACSIT Press, Singapore

Ontology-Based Concept Weighting for Text Documents

Hmway Hmway Tar [1], Thi Thi Soe Nyunt [2]

[1, 2] University of Computer Studies, Yangon

hmwaytar34@gmail.com, ttsoenyunt@gmail.com

Abstract. Documents clustering become an essential technology with the popularity of the Internet. Clustering has been very popular for a long time because it provides unique ways of digesting and generalizing large amounts of information. As most of the recent text clustering research focuses on addressing specific issues (e.g., feature selection and dimensionality reduction), very few new approaches are being devised. The existing clustering technology mainly focuses on term weight calculation. To achieve more accurate document clustering, more informative features including concept weight are important based on the Semantic technologies. Feature Selection is important for clustering process because some of the irrelevant or redundant feature may misguide the clustering results. To counteract this issue, the proposed system presents the concept weight for text clustering system developed based on a k-means algorithm in accordance with the principles of ontology so that the important of words of a cluster can be identified by the weight values. To a certain extent, it has resolved the semantic problem in specific areas. The experimental results performed using dissertations papers from Google Search Engine and the proposed method demonstrated its effectiveness and practical value.

Keywords: Clustering, Concept Weight, Document clustering, Feature Selection, Ontology

1. Introduction

With the booming of the Internet, there are also a billion of textual documents. This factor put the World Wide Web to urgent need for clustering method based on ontology which are developed for sharing ,representing knowledge about specific domain.

Current text clustering approaches tend to neglect several major aspects that greatly limit their practical applicability. Text document clustering is mostly seen as an objective method, which delivers one clearly defined result, which needs to be "optimal" in some way. This, however, runs contrary to the fact that different people have quite different needs with regard to clustering of texts because they may view the same documents from completely different perspectives. Thus, what is needed are document clustering methods that provide multiple subjective perspectives.

Clustering text data faces a number of new challenges. Among others, the volume of text data, dimensionality, sparsity and complex semantics are the most important ones [12]. These characteristics of text data require clustering techniques to be scalable to large and high dimensional data, and able to handle sparsity and semantics. Most of the existing text clustering methods use clustering techniques depends only on term strength and document frequency where single terms are used as features for representing the documents and they are treated independently which can be easily applied to non-ontological clustering. For these problems, the proposed system has provided an efficient solution which is scalable clustering with onslaught the improper feature using Ontology-based computing.

This paper is organized as following. Section 2 describes some related work. Section 3 presents a summary of literature review relating to the research to be pursued. Section 4 will be discussing the proposed system and will propose the research approach and methodology in solving the problem. Section 5 presents the experimental work. Finally, concludes the paper in Section 6.

2. Related Work

Many clustering techniques have been proposed in the literature. Clustering algorithms are mainly categorized into hierarchical and partitioning methods [14, 15,16]. A hierarchical clustering method works by grouping data objects into a tree of clusters. These methods can further be classified into agglomerative and divisive hierarchical clustering depending on whether the hierarchical decomposition is formed in a bottom-up or top-down fashion. K- means and its variants [7, 8, 9] are the most well-known partitioning methods [10]. The proposed system derives the high level requirement for text clustering approaches that they either rely on concept weight.

Andreas Hotho proposed many methods that proved Ontology improve text document clustering. They stated that the ontology can improve document clustering performance with its concept hierarchy knowledge. This system integrates core ontologies as background knowledge into the process of clustering [7, 8].

Lei Zhang, Zhichao Wang [9] proposed ontology-based clustering algorithm with feature weights (OFW-Clustering). They have developed Ontology-based clustering method. Also feature graph is built to calculate feature weights in clustering. Feature weight in the ontology tree is calculated according to the feature's overall relevancy.

3. Ontology for Text Clustering

In the field of ontology, ontological framework is normally formed using manual or semi-automated methods requiring the expertise of developers and specialists. This is highly incompatible with the developments of World Wide Web as well as the new E-technology because it restricts the process of knowledge sharing. Search engines will use ontology to find pages with words that are syntactically different but semantically similar [3, 4, and 5]. Traditionally, ontology has been defined as the philosophical study of what exists: the study of kinds of entities in reality, and the relationships that these entities bear to one another [2]. In Computer Science Ontology is an engineering artefact describing what exists in a particular domain. An ontology belongs to a specific domain of knowledge. The scope of the ontology concentrates on definitions of a certain domain, although sometimes the domain can be very broad. The domain can be an industry domain, an enterprise, a research field, or any other restricted set of knowledge, whether abstract, concrete or even imagined. An ontology is usually constructed with a certain task in mind.

In recent years use of term ontology has become prominent in the area of computer science research and the application of computer science methods in management of scientific and other kinds of information. In this sense the term ontology has the meaning of a standardized terminological framework in terms of which the information is organized.

3.1. Overview of Ontology

Top level ontology or upper level ontologies are the most general ontologies describing the top-most level in ontologies to which other ontologies can be connected, directly or indirectly. Domain ontologies describe a given domain, eg medicine, agriculture, politics; etc. Task ontologies define the top level ontologies for generic tasks and activities. Domain task ontologies define domain-level ontologies on domain specific task and activities are primarily designed to fulfill the need for knowledge in a specific application. Application ontologies define knowledge on the application-level. Evaluating an ontology language is a matter of determining what relationships are supported by the language and required by the ontology or application domain [11].

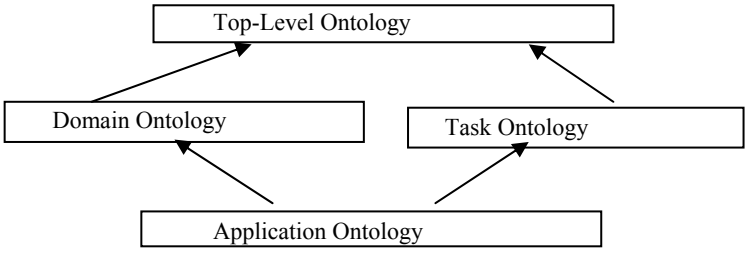

Fig. 1: Categorization of Ontology

4. Proposed System

This system is designed to perform clustering process based on the concept weight support by the ontology. With the help of a domain specific ontology, the proposed technique can transform a feature-represented document into a concept-represented one. Therefore, the target document corpus will be clustered in accordance with the concepts representing individual document, and thus, achieve the proceeding of document clustering at the conceptual level. The system uses the text documents for the clustering process .This system is divided into three major modules. They are document preprocessing, calculating concept weight based on the ontology and clustering documents with the concept weight. The concept weight is also called the Semantic weight. The following figure shows the overview of the proposed system architecture. This system is divided into three major modules. They are document preprocessing, calculating concept weight based on the ontology and clustering documents with the concept weight. The concept weight is also called the Semantic weight. The following figure shows the overview of the proposed system architecture. In the depicted Figure, the ultimate objective is to calculate the concept weight that will help when a subjective worthy of an in-depth analysis as the great advancement of the Semantic Web.

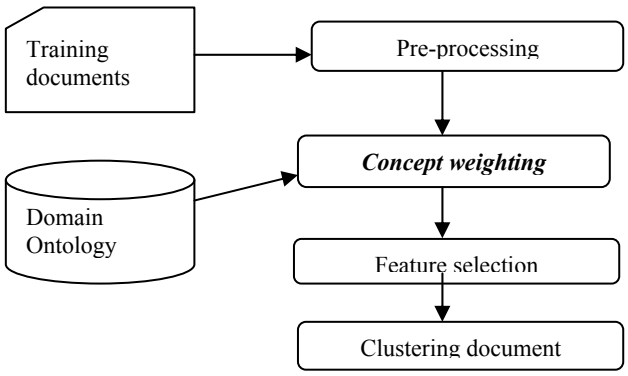

Fig. 2: Proposed System

4.1. Document Pre-processing module

In the preprocessing stage, the document is converted into text file format. The input documents are maintained in separate text files. Mainly, punctuation and special characters are removed on the documents. This is followed by applying some of the most popular choice: removing of common words (e.g., articles, pronouns, prepositions, etc). This is widely done by using a "stop word list collection". However, this approach suffers from being a language specific and domain specific choice.

4.2. Method of calculating the weight module

This system defines ontology as a set of concepts of interest domain organized as a hierarchical (or hierarchical) structure. When designing the method of calculating the weights, the proposed system makes the following assumptions:

1. More times the words appear in the document, more possibly it is the characteristic words;
2. The length of the words will also affect the importance of words. Apparently, one concept in the ontology is related to other concept in that domain ontology. That also means that the association between two concepts can be determined using the length of these two concept's connecting path (topological distance) in the concept lattice.
3. If the probabilities of one word is high, then the word will get additional weight;
4. One word may be the characteristic word even if it doesn't appear in the document.

Some researchers recently put their focus on calculating the words weight using TF-IDF formula in the document. But this method only considers the times which the words appear, while ignoring other factors which may impact the word weighs. A tighter combination of above depicted four assumptions leads to the proposed weighting structure with the ontological aspects. This paper takes into account frequency, length,

specific area and score of the concept when calculating the weighs, using the function with weight values as follows:

$$W = Len \times Frequency \times Correlation\ Coefficient + Probability\ of\ concept \quad (1)$$

where W is the weight of keywords, len is the length of keywords, Frequency is times which the words appear, and if the concept is in the ontology , then correlation coefficient =1 , else correlation coefficient=0. Probability is based on the probability of the concept in the document. The probability is estimated by following equation:

$$P(\text{concept}) = \frac{\text{Number of Occurences of the Concept}}{\text{Number of Occurences of All Concept in Document}} \quad (2)$$

Finally, the system ranks the weights and selects the keywords that have with bigger weight for pre clustering process. Ontology can be represented by standard ontology language. The motivation behind this step is that the OWL is one of the most used standards in describing the knowledge base and already use it in Semantic Web applications. Additional motivation for using OWL is the availability of the knowledge base development tools such as Protégé – OWL editor that supports OWL standard.

4.3. Clustering document module

This proposed system used K-means algorithm which is one of the oldest and most widely used clustering algorithm for clustering process as shown in below:

K-means algorithm is implemented in four steps:

1. Select K points as the initial centroids
2. Repeat
3. Form K clusters by assigning all points to the closest centroid
4. Recomputed the centroid of each cluster.
5. until the centroids don't change

The proposed system used original K-Means to cluster text documents efficiently. A document can be classified by comparing highlighted keywords after which can be got after weighting concept module.

5. Experiments and Results

The text documents are denoted as unstructured data. It is very complex to group text documents. The document clustering requires a preprocessing task to convert the unstructured data values into a structured one. The documents are large dimensional data elements. At first, the dimension is reduced using the stop word elimination and stemming process. The system is tested with 500 text documents collected from Goggle Search Engine relating with dissertation papers which were used in the evaluation. For each article (document) in the corpus, the system used only its abstract for the evaluation. After preprocessing the system can transform a feature represented document into concept represented one with the support of ontology. Therefore, the target document corpus will be clustered in accordance with the concept represented one and thus achieve the proceeding of document clustering at the conceptual level. Also an ontology tailored to the proposed system improves the clustering. Then the proposed technique anchors the analysis process. Finally, it is important to measure the efficiency of the proposed method. The proposed method of the research adopted the most commonly used measures in the data mining, namely, precision and recall for the general assessment (Han and Kamber, 2001). This is further illustrated in the following table:

Table 1: Accuracy of the proposed system

Method	Precision	Recall	F measure
K mean	0.7647	0.8125	0.7878
Ontological k means	0.7778	0.875	0.8235

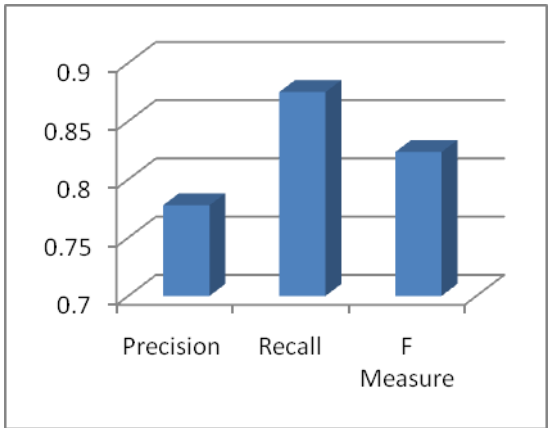

Fig. 3: Result of experiment based on Accuracy

6. Conclusion and Future Work

The World Wide Web grows and changes rapidly and many researchers are stepping into the era of ontology. There is a highly diverse group of text documents. For this reason, document clustering is an important area in data mining. The paper articulates the unique requirements of text document clustering with the support of specific domain ontology. With the use of domain-specific ontology, the proposed system is able to categorize documents on the basis of the concept level. This method present a concept weighting that tries to capture some aspect of the Semantic Web. When weighed by the concept, the clustering system can improve the accuracy and performance of text documents. Finally, the proposed method provides a basis for continued ontology-based document management research. The development and evaluation of advanced ontology-based techniques for text clustering represent interesting and essential future research directions. Another direction is to link this work to web document clustering

7. References

[1] Smith, B.: Ontology. In: Blackwell Guide to the Philosophy of Computing and Information, pp. 155–166. Oxford Blackwell, Malden (2003).

[2] Berners-Lee, T., Weaving the Web, Harper, San Francisco, 1999.

[3] Decker, S., Melnik, S., Van Harmelen, F., Fensel, D., Klein, M., Broekstra, J., Erdmann, M. and Horrocks, I. (2000) *The semantic web: the roles of XML and RDF*, IEEE Internet Computing, Vol.4, No. 5, pp.63–74.

[4] Ding, Y., and Foo, S., (2002). *Ontology Research and Development: Part 1 – A Review of Ontology Generation*. Journal of Information Science 28 (2).

[5] A. Hotho and S. Staab. *Ontology based Text clustering*.

[6] Andreas Hotho, *Ontologies improve Text Document Clustering*.

[7] Lei Zhang, Zhichao Wang. *Ontology-based clustering algorithm with feature weights*, 2010 Journal of Computational Information Systems 6:9 (2010) 2959-2966.

[8] A. Maedche and V. Zacharias, *Clustering Ontology-based Metadata in the Semantic Web*. In Proceedings of the 6th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD'02), Helsinki, Finland, pp. 342-360, 2002.

[9] L. Jing, M. K. Ng, J. Xu and Z. Huang, *Subspace clustering of text documents with feature weighting k- means algorithm*, Proc. of PAKDD, pp. 802-812, 2005.

[10] W. Fan, L. Wallace, S. Rich, and Z. Zhang, *Tapping into the power of text mining*, the Communications of ACM, 2005.

[11] M. Steinbach, G. Karypis, and V. Kumar. 2000. *A comparison of document clustering techniques*. KDD Workshop on Text Mining'00.

[12] P. Berkhin. 2004. *Survey of clustering data mining techniques* [Online]. Available: http://www.accrue.com/products/rp_cluster_review.pdf.