International Journal of Information, Control and Computer Sciences
ISSN: 2517-9942
Vol:5, No:9, 2011

Ontology-based Concept Weighting for Text Documents

Hmway Hmway Tar , Thi Thi Soe Nyaunt

Abstract**—**Documents clustering become an essential technology with the popularity of the Internet. That also means that fast and high-quality document clustering technique play core topics. Text clustering or shortly clustering is about discovering semantically related groups in an unstructured collection of documents. Clustering has been very popular for a long time because it provides unique ways of digesting and generalizing large amounts of information. One of the issues of clustering is to extract proper feature (concept) of a problem domain. The existing clustering technology mainly focuses on term weight calculation. To achieve more accurate document clustering, more informative features including concept weight are important. 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.

Keywords**—**Clustering, Concept Weight, Document clustering, Feature Selection, Ontology

I. 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. At the same time, as a result of the vigorous developments and accessibility of the World Wide Web (WWW), a great quantity of information was suddenly made available to people. However, due to the enormousness of the data, users waste a lot of time browsing the Internet and searching for the information they need; it makes the tasks of searching, accessing, displaying, integrating and maintaining data more laborious. With the aim of solving this difficulty, Berners-Lee and Fischetti (1999) conceived the concept of the Semantic Web. Based on this concept, ontology constitutes the foundations of the Semantic Web. Semantic Web is anything but a new kind of network; it is built within the existing network environment and provides a highly readable data without modifying or altering any of the contents. It also enables computers and people to work in cooperation. It is the idea of having data on the Web defined and linked in a way that it can be used for more effective discovery, automation, integration and reuse across various applications [1].

Hmway Hmway Tar, University of Computer Studies , Yangon, Myanmar(e-mail: hmwaytar34@ gmail.com).
Thi Thi Soe Nyunt, University of Computer Studies , Yangon, Myanmar(e-mail:ttsoenyunt@ gmail.com).

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. Text clustering is one of the fundamental functions in text mining [13]. Clustering is to divide a collection of text documents into different category groups so that documents in the same category group describe the same topic. There are many uses of clustering in real applications, for example, grouping the Web search results and categorizing digital documents. Unlike clustering structured data, 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. 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. This proposed system also considers concept weight for selecting the trait of the documents with the support of ontology.Clustering text documents into category groups is an important step in indexing, retrieval, management and mining of abundant text data on the Web or in corporate information systems. Among others, the challenging problems of text clustering are big volume, high dimensionality and complex semantics [12]. For these problems, the proposed system has provided an efficient solution which is scalable clustering with onslaught the improper feature using Ontology-based computing.Recently, there has been a flurry of activity in the clustering area. Every clustering algorithm learns in a slightly different way and introduces biases. Algorithm will often behave better in a given domain. Furthermore, interpretation of the resulting clusters may be difficult or even entirely meaningless. This makes for an interesting and active field of research. Many of the drivers for clustering analysis stem from computer and natural sciences.Clustering has been applied to field such as information retrieval, data mining, image processing, segmentation, Gene expression clustering and pattern classification .The problem of document clustering is generally defined as follows: Given a set of documents,

would like to partition them into a predetermined or an automatically derived number of clusters, such that the documents assigned to each cluster are more similar to each other than the documents assigned to different clusters. In other words, the documents in one cluster share the same topic, and the documents in different clusters represent different topics. 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.

II. 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]. All clustering approaches based on frequencies of terms and similarities of data points suffer from the same mathematical properties of the underlying spaces (Beyer et al., 1999; Hinneburg et al., 2000). Therefore, the proposed system derives the high level requirement for text clustering approaches that they either rely on concept weight.In the proposal for feature selection made in (Devaney & Ram, 1998) describe feature selection for an unsupervised learning task, namely conceptual clustering. They discuss a sequential feature selection strategy based on an existing COBWEB conceptual clustering system. In their evaluation they show that feature selection significantly improves the results of COBWEB. The drawback that Devaney and Ram face, however, is that COBWEB is not scalable like K-Means. In [6] Prof K. Raja identifies the semantic relations using the ontology. The ontology is used represent the term and concept relationship. The synonym, meronym and hypernym relationships are represented in the ontology. The concept weights are estimated with reference to the ontology. The concept weight is used for the clustering process. In [6], the concept weight is highlight for complex semantic problem area. But in this proposed system the concept weight is calculating for curse of dimensions problem.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 integrate 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.

Maedche and Zacharias examine hierarchical clustering of ontology-based metadata for the Semantic Web [11]. That system measures compute relatedness scores based on the relational similarity of two concepts. Unlike their approach, the proposed system clusters text documents using weighting sets derived from the ontologies. As a result, the proposed approach avoids the complexity of comparing conceptual graphs. In addition, this system is able to demonstrate the using of ontological expansion over traditional clustering that does not use non-ontological aspects.

III. Ontology for Text Clustering

As a result of the extensive developments in the Internet, sharing knowledge with each other has finally become a reality. Unfortunately, it is for the same reason that we are facing an overflow of data and information. Nevertheless, the Semantic Web concept proposed by Berners-Lee and Fischetti (1999) paved the way to the formulation of possible and effective solutions. The most vital tools in searching for information and related resources in a Semantic Web are the ontology and intelligent agent. 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, this task focus restricts the content and structure of the ontology. An ontology can, for example, be used with a reasoning engine to classify instances, check consistency of facts, or answer queries. On the other hand there can be different kinds of tasks, where complex reasoning is not the main focus, such as annotating information, or acting as a user interface for structured document browsing. The nature or the task imposes requirements on the content and structure of the ontology.In contrast to purely statistical correlations, ontologies encode semantic relationships between terms. Present-day ontologies can be grouped into two general categories: those that form meta-language dictionaries and those that are derived from knowledge bases built for inference engines and expert systems.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.Text clustering and classification are two promising approaches to help users organize and contextualize textual information. Existing text mining systems typically based on term weighting. Recent work has shown improvements in text clustering by means of conceptual features extracted from ontologies (Bloehdorn and Hotho (2004), Hotho et al.(2003).So far, however the ontological structures employed for this task are created manually by knowledge engineers and domain experts which requires a high initial modeling.

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. In theory, these ontologies are shareable as they express very basic knowledge, but this is not the case in practice, because it would require agreement on the conceptualization.Domain ontologies describe a given domain, eg medicine, agriculture, politics, etc. They are normally attached to top level ontologies, if needed, and thus do not include common knowledge. Different domains can be overlapping, but they are generally only reusable in a given domain. The overlap in different domains can sometimes be handled by a so called middle layer ontology, which is used to tie one or more domain ontologies to the top level ontology.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.Method ontologies give definition of the relevant concepts and relations applied to specify a reasoning process so as to achieve a particular task.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].

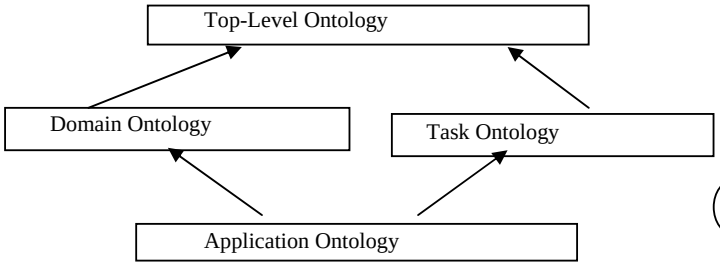

Fig. 1 Categorization of Ontology

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; this task focus restricts the content and structure of the ontology.

IV. 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.

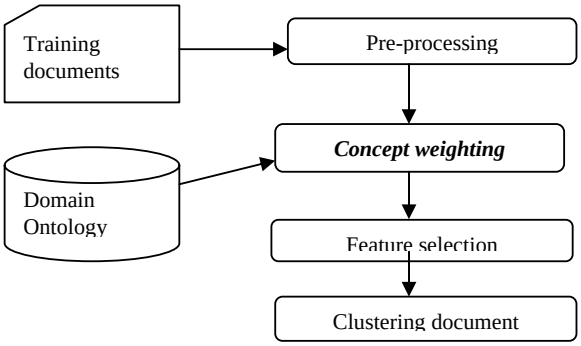

Fig. 2 Proposed System Architecture

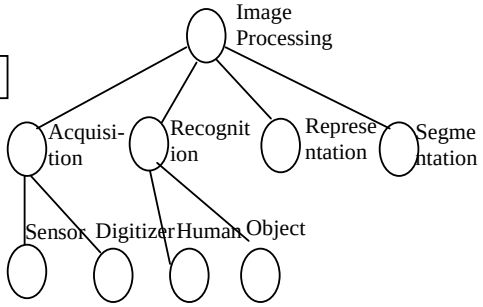

Fig.3 Hierarchical taxonomy of Image Processing Ontology

4.1Document 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.

4.2Method 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 of the concept is estimated by following equation:

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

Finally, the system ranks the weights and selects the keywords that have with bigger weight for preclustering process. After selecting the concepts the proposed system represents each document as a concept vector; i.e., the concept-based document representation.

Ontology can be represented by standard ontology language. The motivation behind this step is that the OWL is one of the most used standard 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. The proposed ontology snippet of proposed domain ontology is shown as in Figure 4.

```
<owl:Class rdf:ID="ImageAquisition">
<rdfs:subClassOf rdf:resource="#Image" />
/owl:Class>
<owl:Class rdf:ID="ImageRecognition">
<rdfs:subClassOf rdf:resource="#Image" />
</owl:Class>
<owl:Class rdf:ID="ImageRepresentation">
<rdfs:subClassOf rdf:resource="#Image" />
</owl:Class>
<owl:Class rdf:ID="ImageSegmentation">
<rdfs:subClassOf rdf:resource="#Image" />
</owl:Class>
```

Fig 4 A part of the OWL source

4.3Clustering 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. Recompute the centroid of each cluster.
5. until the centroids don't change.

The similarity between two documents needs to be measured in a clustering analysis. Over the years, many prominent ways have been used to compute the similarity between documents mp and mj. The most commonly used distance functions for clustering algorithm are the Euclidean distance, Manhattan (city block) distance and Cosine correlation measure. The commonly used similarity measure in document clustering is the cosine correlation measure, given by

$$Cos(m_p, m_j) = \frac{m_p, m_j}{|m_p||m_j|} \quad (4)$$

where $m_p\ m_j$ denotes the dot-product of the two document vectors. |.| indicates the length of the vector.

V. 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 I
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

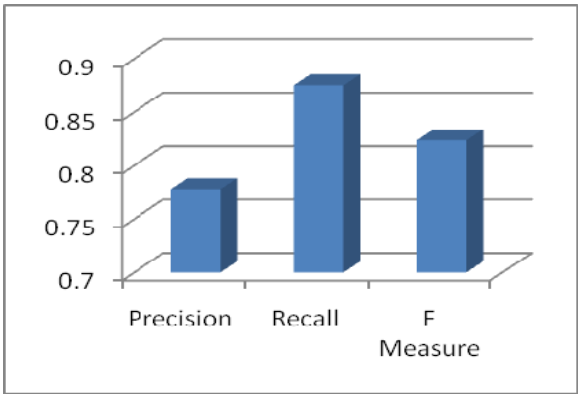

Fig. 5 Result of experiment based on Accuracy

VI. Conclusion And Futrure 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.

References

[1] W3C Semantic Web Activity Statement: W3C's Technology and Society domain(2001). www.w3.org/2001/sw/Activity
[2] Smith, B.: Ontology. In: Blackwell Guide to the Philosophy of Computing and Information, pp. 155–166. Oxford Blackwell, Malden (2003).
[3] Berners-Lee, T., Weaving the Web, Harper, San Francisco, 1999
[4] 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.
[5] Ding, Y., and Foo, S., (2002). Ontology Research and Development: Part 1 – A Review of Ontology Generation. Journal of Information Science 28 (2).
[6] Prof.K.Raja(2010) Clustering Technique with Feature Selection for Text Documents.
[7] A. Hotho and S. Staab "Ontology based Text clustering.
[8] Andreas Hotho,"Ontologies improve Text Document Clustering".
[9] Lei Zhang , Zhichao Wang "Ontology-based clustering algorithm with feature weights",2010Journal of Computational Information Systems 6:9 (2010) 2959-2966.
[10] 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
[11] Travis D. Breaux "Using Ontology in Hierarchical Information Clustering", Proceedings of the 38th Hawaii International Conference on System Sciences – 2005
[12] 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.
[13] W. Fan, L. Wallace, S. Rich, and Z. Zhang, "Tapping into the power of text mining," the Communications of ACM, 2005.
[14] Jain, A.K, Murty, M.N., and Flynn P.J. 1999. Data clustering: a review. ACM Computing Surveys, pp. 31, 3, 264-323.
[15] M. Steinbach, G. Karypis, and V. Kumar. 2000. A comparison of document clustering techniques. KDD Workshop on Text Mining'00
[16] P. Berkhin. 2004. Survey of clustering data mining techniques [Online]. Available: http://www.accrue.com/products/rp_cluster_review.pdf.