Enhancing Traditional Text Documents Clustering based on Ontology

Hmway Hmway Tar
University of Computer Studies, Yangon
Myanmar

Thi Thi Soe Nyunt
University of Computer Studies, Yangon
Myanmar

ABSTRACT
Ontologies currently are a hot topic in the areas of Semantic Web. The current clustering research emphasizes the development of a more efficient clustering method and mainly focuses on term weight calculation without considering the domain knowledge. This paper investigates how ontologies can also be applied to the clustering process. To complement the traditional clustering method, more informative features including concept weight are important based on recent developments in the area of the Semantic technologies. 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 weighted values. To a certain extent, it has resolved the semantic progeny in specific areas. The experimental results performed using dissertations papers from Google Search Engine and the proposed method demonstrated its effectiveness and practical value.

General Terms
Text mining, Machine Learning, Semantic Web

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

1. INTRODUCTION
Today is the knowledge age, and the Internet contains many non-trivial of documents .All year people read in a utilitarian fashion .This factor put the World Wide Web to urgent need for clustering method based on knowledge technology – ontology which are developed for sharing ,representing knowledge about specific domain.

The management of non-numerical data traditionally is a task typically associated to Artificial Intelligence methods. Data-mining techniques and in particular clustering algorithms were conceived for managing non-numerical data. From the different methods included in the field of Data Mining, we have focused on knowledge discovery from data using clustering (Han and Kamber 2000). Clustering is a masterpiece in many data mining methodologies, because it builds a classification or partition into coherent clusters from unstructured data sets.

Clustering algorithms have focused on the management of categorical data. However, in the last years, textual information has grown in importance. Proper processing of this kind of information within data mining methods requires an interpretation of their meaning at a semantic level. In this work, we can get document clustering at conceptual level for clustering text documents.

Text document clustering is mostly seen as an objective method, which delivers one clearly defined result, which needs to be "optimal" in some way. With the rapid development and widely use of the Internet, we have to clustered document based on theme becomes a heated topic and will be more significant than before. Text clustering is one of the fundamental functions in text mining [1]. 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. 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 so that the utility of ontology can be applied in clustering process.

Traditional knowledge-representation systems typically have been centralized, requiring everyone to share exactly the same definition of common concepts such as "parent" or "vehicle." But central control is stifling, and increasing the size and scope of such a system rapidly becomes unmanageable [2].

To counteract this issue, this paper investigates which beneficial effects can be achieved for text document clustering by using ontological computing. Ontologies can enhance the functioning of the clustering in many ways. A major reason is that calculating concept weight is that the feature space that possesses none of the conceptual irregularities that underlay the domain (the distance from a purple grape to red apple is not the same as from a green orange to a red apple). The main goal of this research is to achieve an ontology-based clustering for the exploitation of domain ontologies to support semantic capabilities.

This paper is organized as following. Section 2, 3 presents a summary of literature review relating to the research to be pursued. Section 4 motivates for this research. Section 5 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. 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 [6]. In Computer Science, ontology is an engineering artifact describing what exists in a particular domain. 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. 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. OVERVIEW OF ONTOLOGY

Ontologies are designed for being used in applications that need to process the content of information, as well as, to reason about it, instead of just presenting information to humans. They permit greater machine interpretability of content than that supported by XML, RDF and RDF Schema (RDF-S), by providing additional vocabulary along with a formal semantics.

From a structural point of view (Stumme, Ehrig et al. 2003; Cimiano 2006), an ontology is composed by disjoint sets of concepts, relations, attributes and data types. Ontologies can be classified in several forms. An interesting classification was proposed by Guarino (Guarino 1998), who classified types of ontologies according to their level of dependence on a particular task or point of view: 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 .Domain ontologies, on one hand, are general enough to be required for achieving consensus between a wide community of users or domain experts and, on the other hand, they are concrete enough to present an enormous diversity with many different and dynamic domains of knowledge and millions of possible concepts to model. Being machine readable, they represent a very reliable and structured knowledge source.

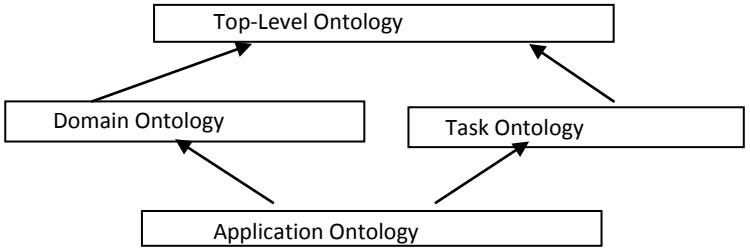

Fig. 1: Categorization of Ontology

4. MOTIVATION FOR TEXT CLUSTERING

In the last years, with the enormous growth of the Information Society, the Web has become a valuable source of information for almost every possible domain of knowledge. This has motivated many researches to start considering the Web as a valid repository for Information Retrieval and Knowledge Acquisition tasks. So, the Web, thanks the huge amount of information available for every possible domain and its high redundancy, can be a valid knowledge source for similarity computation. In this sense, the amount and heterogeneity of information is so high that it can be assumed that the Web approximates the real distribution of information (Cilibrasi and Vitányi 2004), representing the hugest repository of information available (Brill 2003). In many knowledge related tasks the use of statistical measures (e.g. co-occurrence measures) for inferring the degree of relationship between concepts is a very common technique when processing unstructured text (Lin 1998a). However, these techniques typically suffer from the sparse data problem (i.e. the fact that data available on words may not be indicative of their meaning). So, they perform poorly when the words are relatively rare (Sánchez, Batet et al. 2010b). In that sense, the size and the redundancy of the Web has motivated some researches to consider it as a corpus from which extract evidences of word relationships. Some authors (Turney 2001; Brill 2003) have demonstrated the convenience of use a wide corpus as the Web to address the data sparse problem. However, the analysis of such an enormous repository is, in most cases, impracticable. Here is where the use of web search engines (e.g. Google, Bing, Yahoo) can properly scale this high amount of information, obtaining good quality and relevant statistics. So, robust web-scale statistics about information distribution in the whole Web can be obtained in a scalable and efficient way from queries performed into a web search engine (Sánchez 2008) (Sánchez 2009). Therefore, this system applied the ontology-based concept weighting to improve the clustering process.

5. PROPOSED ONTOLOGICAL SYSTEM FOR TEXT CLUSTERING

The primary contribution of this research is the calculation of concept weight for dimensions reduction for concept vector. This research work deals with three step phases: document preprocessing, concept weighting, and the clustering document collection. Domain ontology has been created using the software Protégé software. And, finally, apply k-means as a baseline algorithm.

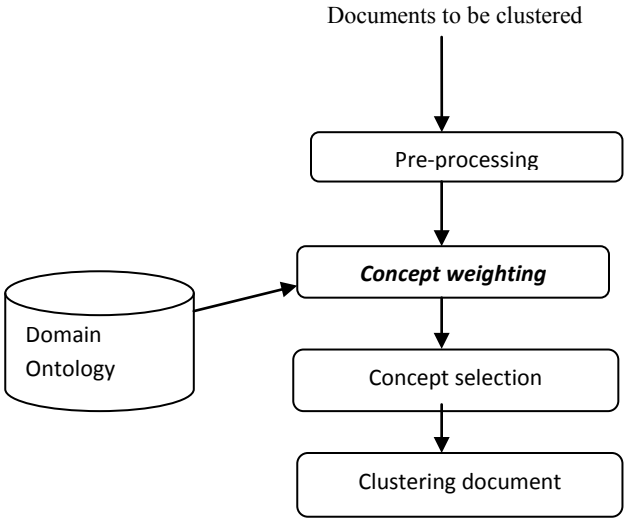

Fig 2: Overall process of proposed technique

5.1 Pre-processing Phase

The text document collection is the initial stage for this phase. The textual information is stored in many kinds of machine readable form, such as PDF, DOC, PostScript, HTML, and XML and so on. After the text document are collected from Google search engine , the abstract page is elective from those pdf file and transformed into TXT format and maintained in the text files. After that phase, the system removes the stop words and stemming on the extracted text document. The system transform the term related document to concept represented one. 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". The stop words lists is download from the Wikipedia stop word lists [7].

5.2 Concept Weighting Phase

The advantages of text mining based on domain ontology are one of the effective mining methods. After the preprocessing step changes the text objects, the system converts the attributes to numeric one and uses the weighted vector to represent the text objects. So that each data point has a specified measured value. The use of ontology in text mining leads to more meaningful and interesting results and can reveal a more general concept. Our goal is slightly different from previous approaches. We also examine how new concept weighting processes can aid in extracting precise and useful information from the ontology data, thus reducing the curse of dimension problem in feature weighting. One thing's behind this is that the system accuracy also depends on accuracy domain ontology [8].This means that the construction of ontology needs to be good enough for supporting. Also to address the issue of text clustering, a suitable method for calculating and selecting the feature vector is proposed. Different terms have different importance in a text, thus an important indicator for concept weight contributes to the semantics of document is calculated by the equation (1). 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 [9];(this means that if the number of occurrence of word is high then the frequency of that word will be high)
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. And also this method is only a binary weighting method. A tighter combination of above depicted four assumptions leads to the proposed weighting structure with the ontological aspects. This ontological computing can give more accurate result course of its concept hierarchy. For example, if we have to survey the pet owner and non-pet owner to draw conclusion. But the cat owners among all pet owners are not responding to the survey. In this situation we can use ontological aspects to be adjusted with the data weigh. That does can drive following weighting scheme. 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 = Length × Frequency× Correlation Coefficient + Probability of concept (1)

where W is the weight of keywords, Length is the depth of concept in the ontology 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.

5.3 Document Clustering Phase

Clustering is generally seen as the task of finding groups of similar individuals based on information found in data, which means that the data individuals in the same group are more similar to each other than to individuals in other groups. So, clustering algorithms partition data into a certain number of clusters (groups, subsets, or categories) (Xu and Wunsch 2005).

The k-Means algorithm (Duda and Hart 1973) implemented as a simple procedure that initially selects k random centroids, assigns each example to the cluster whose centroid is closest, and then calculates a new centroid for each cluster. Examples are reassigned to clusters and new centroids are re-calculated repeatedly until there is no change in clusters.

Cluster analysis has been of long-standing interest in statistics, numerical analysis, machine learning (where it is commonly called unsupervised learning) and other fields. Some trace it back to the work of Adanson as early as 1757 for classifying botanic species [Adanson 1757]. Various clustering methods are used in various fields of applications.

All the general purpose clustering algorithms can be applied to document/text clustering. Some algorithms have been developed solely for document/text clustering. All these algorithms can be classified into partitional, hierarchical, and others such as probabilistic, graph-based, and frequent term-based, etc.

The choice of which of the clustering algorithms is the best candidate for clustering process cannot be supported by theory. Each of these algorithms has pros and cons. The proposed system, the most widely used the k-means algorithm [Lloyd, 1957] in the literature is applied for the clustering results because of its simplicity. In this system, centroids selection update the cluster centers after the iteration, instead of after all documents are classified for evaluation.

6. EXPERIMENTS AND RESULTS

The text documents are denoted as unstructured data. It is very complex to group text documents. The document clustering requires a pre-processing 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 pre-processing 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:

Paper id	Paper Title	Significance
D 110	Design and Implementation of Android Phone Surveillance System	Disturbed Cluster
D 112	Android English Dictionary Development Based On Intelligence Mobile Phone Platform	Disturbed Cluster
D 221	Moment in Online Handwritten Character Recognition	Image Cluster
D 229	An Image Recognition and Interpretation System for the Dutch Postbank	Image Cluster
D 995	Massive Centralized Cloud Computing (MCCC) Exploration in Higher Education	Disturbed Cluster
D 994	Cloud Computing Based on Service Oriented Platform	Disturbed Cluster

Table 1: Resulted output from the training documents

Method	Precision	Recall	F-measure
k-mean	0.7466	0.7501	0.7541
Ontological k-means	0.7778	0.875	0.8235

Table 2: Accuracy of the proposed system

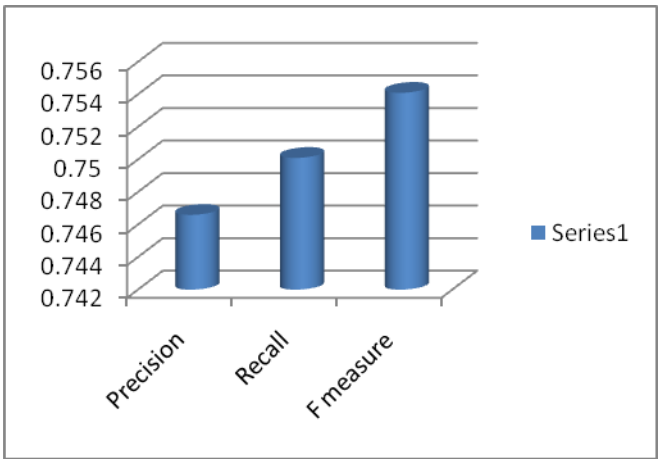

Table 3 : Accuracy of the proposed method

7. 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 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 used of domain ontology in the proposed system will extend perfect ontology in the future work. 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.

8. REFERENCES

[1] A new form of Web content that is meaningful to computers will unleash a revolution of new possibilities by TIM BERNERS-LEE, JAMES HENDLER and ORA LASSILA

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

[7] http://www.textfixer.com/resources/common-english-words.txt

[8] Stefan Brueggemann,Using Domain Knowledge Provided by Ontologies for Improving Data Quality Management.