Ontology-Based Criterion in Categorical Clustering

Hmway Hmway Tar,Thi Thi Soe Nyunt
University of Computer Studies, Yangon
hmwaytar34@gmail.com, ttsoenyunt@gmail.com

Abstract

Currently the Internet is the first solution for scholars in finding information. But, analyzing and interpreting this volume of information will lead to researchers overload in pursuing their research. Document clustering has become an increasingly important task in analyzing huge documents. The challenging aspect to analyze the enormous documents is to organize them in such a way that facilitates better search and knowledge extraction without introducing extra cost and complexity. In this paper, ontology-based document clustering method has been proposed. Ontology based document categorization helps scholars to find relevant categorization of special areas by applying ontology based measure. However, there are few researches on ontology based measure with weighting methods. Previous works lack of an appropriate representation scheme for research topics. According to this issue this papers constructs ontology and apply this ontology in the clustering process and after that the topic discovery phase is followed. This paper describes a method that combines Semantic Web and ontology to support clustering unstructured text documents .

1. Introduction

Clustering is useful in several exploratory pattern-analysis, grouping, decision making, and machine-learning situations. This includes data mining, document retrieval, image segmentation, and pattern classification.

Data clustering has been considered as a primary data mining method for knowledge discovery. There have been many clustering algorithms in the literature. In general, major clustering methods can be classified into the hierarchical or the partition category. A hierarchical method creates a hierarchical decomposition of the given set of data patterns. A partition approach produces k partitions of the patterns, where each partition represents a cluster.

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, the proposed system has 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.

There are a number of approaches to solving document clustering problems. Numerous document clustering algorithms appear in the literature [1]. 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, the proposed method 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 [2]. Traditional knowledge-representation systems typically have been centralized. But central control is stifling, and increasing the size and scope of such a system rapidly becomes unmanageable [3].

Dealing with text documents is a hard task. This is due to some intrinsic characteristics of human languages. For example, the same word can have different meanings according with the context in which it is referred. That is one of the major reasons why these systems use the support of the ontology. The semantic web also motivates the effort of researchers to apply the ontological computing. The major contribution of the proposed system is an innovative semantic framework for document clustering. Inwardly, the main goal of the proposed system is to propose Ontology-based Clustering method for text documents which are able to overcome the difficulties of dealing with sparse data.

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 the proposed system 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 and will be discussing the proposed system and will propose the research approach and methodology in solving the problem. Section 5 presents the experimental work.

2. Ontology for Text Clustering

The core of Semantic Web is ontology, which is used to explicitly represent our conceptualizations. Ontology engineering in the Semantic Web is primarily supported by languages such as RDF, RDFS, and OWL [7]. 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 [4, 5, and 6]. 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. In the context of computer and information sciences, ontology defines a set of representational primitives with which to model a domain of knowledge or discourse. The representational primitives are typically classes (or sets), attributes (or properties), and relationships (or relations among class members). In computer and information science, ontology is a technical term denoting an artifact that is designed for a purpose, which is to enable the modeling of knowledge about some domain, real or imagined [8].

3. Types of Ontology

There are two main classes of Ontologies: the first would be the one that is employed to explicitly capture “static knowledge” about a domain, in contrast to Ontologies (the second) that provide a reasoning point of view about the domain knowledge(problem solving knowledge).

In the first class a distinction between types is made on the basis of the level of generality, as summarized in the table below:

1.Domain Ontologies	Designed to represent knowledge relevant to a certain domain type, e.g. medical, mechanical, etc.
2.Generic Ontologies	Can be applied to a variety of domain types. Mereology(Part-Whole theory) Ontologies are applicable to many technical domains. Also, called “super theory”.

3.Representational Ontologies	These formulate general representation entities without defining what should be represented. The Frame Ontology is well known example.

For the problem solving knowledge class, two types may be found:

1.Task Ontologies	Provide terms specific for particular Tasks
2.Method Ontologies	Provide terms specific to particular Problem Solving Methods

Another type of Ontology is the Application Ontology. The Application Ontology is a combination of the Domain and Method Ontologies that includes all the knowledge static and problem solving needed for the modeling of a particular domain.

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. Moreover ontology is one of the ways to put things in order.

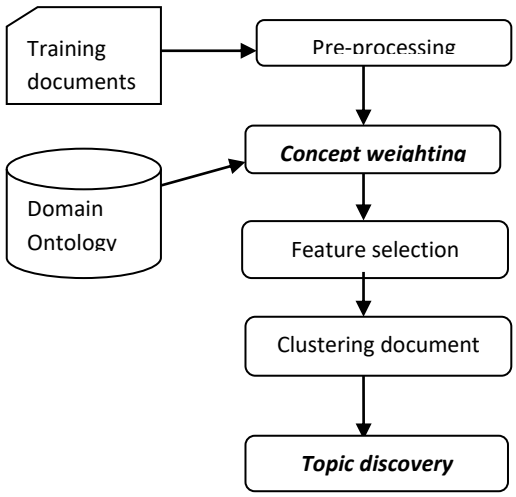

Figure 1: The Proposed System Architecture

4.1 Pre-Processing Phase

The text document collection is the initial stage for this phase. The input for the proposed system is set of text documents. 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. Pre-processing is a process in which the system extract the meaningful terms and count their frequencies while reading the input text file. Here several formatting procedures are performed on the document text, in order to make sure that the computed statistics are meaningful. 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].

4.2 Concept Weighting Phase

The advantages of text mining based on domain ontology are one of the effective mining methods. Compared to traditional databases model, the ontological model can be mines concepts at a higher level. The high-level concepts can provide more accurate and clearer summary of text documents. The semantic document clustering approach is unique in that it provides users with document cluster models from an ontology-enriched scale-free representation of a set of documents, which are the summaries for each document cluster. The use of ontology in text mining leads to more meaningful and interesting results and can reveal a more general concept. The proposed method goal is slightly different from previous approaches. The System 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. The system takes into account frequency, length, specific area and score of the concept when calculating the weighs, using the function with weight values as follows:

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

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. Use these weights to do document clustering on the training data.

4.3 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. This system use k-Means algorithm (Duda and Hart 1973) groups a collection - of examples into k clusters so as to minimize the sum of squared distances to the cluster centers. It can be 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.

4.4 Topic Discovery Phase with rule based

In the last phase of the system, apply simply rule for resulting output. In general, the rules are in the following form:

IF <antecedent conditions>

THEN <consequent conditions>

This system try to test that if document ‘i’ is in k=1 cluster then it will be determined and output as this document belong to Distributed Cluster and else the document is in Image Processing Cluster.

5. Experimental 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 1000 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).

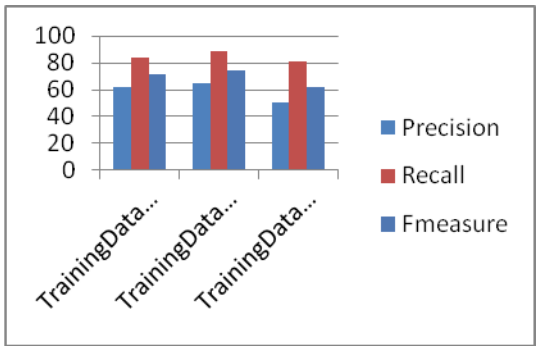

Figure2. Result of the proposed method

References

[1] EL-HAMDOUCHI, A., and P. WILLETT. 1987. "Techniques for the Measurement of Clustering Tendency in Document Retrieval Systems." J. Information Science, 13, 361-65.

[2] 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

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

[7] Li Ding , Pranam Kolari, Zhongli Ding, Sasikanth Avancha,Tim Finin,University of Maryland Baltimore Country "Using Ontologies in the Semantic Web: A Survey"

[8] Ling Liu and M. Tamer Özsu (Eds.),"Ontology to appear in the Encyclopedia of Database Systems", Springer-Verlag, 2008.