Clustering Technique based on Concept Weight to Text Documents

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

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 issue of clustering is to extract proper feature (terms) 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 uses the concept weight for clustering in accordance with the principles of ontology. To a certain extent, it has resolved the semantic problem in specific areas.

1. Introduction

With the explosion of the amount of textual data in the information age, there is an urgent need for clustering method based on ontology which are developed for sharing ,representing knowledge about specific domain. Following the rapid widespread of computer networks, data handling had slowly grown dependent on computer networks and servers particularly in carrying out automated exchange tasks. In other words, data storage, data management, data transmission and even analysis rely on computer and network technology. 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. Ontology is made capable to 'describe metadata' in order to build one complete glossary that will clearly define the data found in the World Wide Web (WWW). 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. In fact, the Semantic Web is an extension of the current Web in which information is given well-defined meaning. 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].

With the booming of the Internet, the World Wide Web currently contains billions of documents. To explore and utilize the huge amount of text documents, many methods are developed to help users effectively navigate, summarize, and organize text documents that is why clustering become an important factor. There are several approaches for Web document clustering which can organize large amount of documents into a number of meaningful clusters. In other words, the document in one cluster share

the same topic, and the document in different clusters represent different topics. However, 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.

Traditional clustering techniques depend only on term strength and document frequency which can be easily applied to non-ontological clustering. This proposed system also considers concept weight with the support of ontology.

Clustering is a form of data analysis that finds patterns in data. Clusters are learnt in an unsupervised manner where no a priori labeling of patterns has occurred. Supervised learning differs because labels or categorization. When a collection is clustered all items are represented using the same set of features. 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 this type of 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. Finally, concludes the paper in Section 5.

2. Related Work

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.

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]. In clustering metadata, they introduce a number of semantic measures required to compute the similarity matrix prior to constructing the clusters. These measures compute relatedness scores based on the relational similarity of two concepts, such as comparing their locations in a classification hierarchy or evaluating the intersection of their attributes. Unlike their approach, the proposed system clusters text documents using expanded term 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 ontologies.

3. Ontology for Text Clustering Techniques

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. Ontology Construction

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].

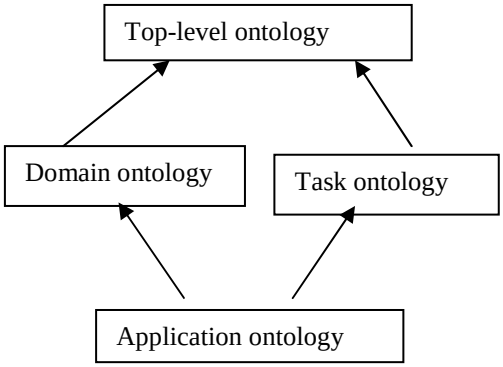

Figure 1. Categorization of Ontology

4. Proposed System

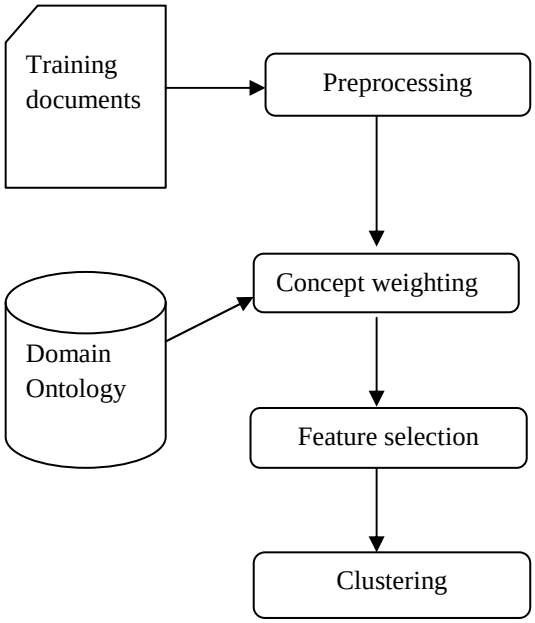

Figure 2. Proposed system

4.1. Document Pre-processing module

This system is designed to perform clustering process based on the concept weight support by the ontology. 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.

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

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 will also affect the importance of words. Generally, one concept in the ontology is related to other concept in that domain ontology.

(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. In this paper, we take into account frequency, length, specific area and related information and so on when calculating the weighs, using the function with weight values as follows:

W = len× Frequency× Correlation coefficient + probability score of concept (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 score 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, we rank the weights and select the keywords that has with bigger weight for clustering process.

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:

Given k, the k-means algorithm is implemented in four steps:

1. Partition objects into k nonempty subsets
2. Compute seed points as the centroids of the cluster of the current partition (the centroid is the center, i.e mean point, of the cluster)
3. Assign each object to the cluster with the nearest seed point
4. Go back to Step 2, stop when no more new assignment.

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

$$\cos(m_p, m_j) = \frac{m_p m_j}{|m_p||m_j|} \quad (3)$$

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

5. 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 web documents. For this reason, document clustering is an important area in web mining. The paper articulates the unique requirements of text document clustering with the support of specific domain Ontology. This proposed system applied semantic weight calculation in specific domain Ontology. When weighed by the concept, the clustering system may improve the accuracy and performance of text documents.

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.