International Journal of Applied Information Systems (IJAIS) – ISSN : 2249-0868
Foundation of Computer Science FCS, New York, USA
Volume 6 – No. 3, September 2013 – www.ijais.org

Ontology Employment in Text Document Clustering combined with Grouping Algorithm

Hmway Hmway Tar
University of Computer Studies, Loikaw
15/66 AyeTharYar, Taunggyi

Pye Phyo Oo
University of Medicine (1), Yangon
584 Rugby Rd. Apt 23S,
Brooklyn, NY11230

ABSTRACT

Incorporating semantic knowledge from ontology into text document clustering is an important but challenging problem. Moreover, there are many of computer science and medical based subject related papers and journals cited on the Internet. The purpose of this system is to cluster the documents based upon the statistical method and from the semantic web point of view, the system advances in the field of scientific endeavor. Moreover this system is the advanced and extended version of the paper we have been published before. After time passed the testing data amount becomes lager and lager and we have been found that our previous methods should have to improve in more mathematically. Finally, it also reports on the experiments that performed to test the system utilization weighting scheme which is used to encode the importance of concepts inside documents. For the experiments the system has to use ontology that enables us to describe and organize this from heterogeneous sources, and to cluster about it. The experiments reveal that even the testing documents increased; the system may actually be able to produce useful results.

General Terms

Semantic Web, Ontology, Text Document Clustering

Keywords

Semantic Web, Clustering, Text Clustering Algorithm

1. INTRODUCTION

While the capabilities of today's Web directed towards the Semantic area, many research for the field of ontology become more interested area. With the booming of the Internet, the World Wide Web contains a billion of textual documents. This factor put the World Wide Web to urgent need for clustering method based on ontology which mare developed for sharing, representing knowledge about specific domain. 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. However, as more text documents are populated, many systems urgently need to rely on well model technologies such as Semantic Web.

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

Ontologies currently are hot topics in the area of Semantic Web. To effectively use that data and information this system applies ontology concepts to develop well defined model for data with well structure. This research is mainly concerned with the concept weighting and grouping algorithm by taking the advantages of the concepts of domain ontologies. Moreover, it is vital to have a reliable way to cluster massive amounts of text data. This method present a new way to mine documents and cluster them using ontology.

One of the goals of the system is to cluster text documents based on their concept weight similarity rather than keywords. This phase focuses on the introduction of the concept of semantic features weight similarity into the clustering methods. Text clustering algorithms have focused on the management of numerical and categorical data. However, in the last years, textual information has grown in importance. Proper processing of that kind of information within data mining methods requires at a semantic level. In the system's work, the concept of concept weighting is introduced to provide a formal framework for clustering documents. Available knowledge is formalized by means of ontology. Clustering cover approaches completely or partially relying on ontology. In the system values represent concepts weight rather than simple term weight. As a consequence, applying the results obtained in the first part of this research to the clustering processes should also have benefits on having a better identification of the clusters than non-semantic clustering. On the other hand, it has been proposed a method to include semantic features weight into an unsupervised clustering.

2. Background Theory

Text mining is a technique developed from data mining to analyze textual data especially unstructured (free text, abstract, etc). A text document is unclear, and according to [1, 2, 3]. 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 [4,5,6,7].

3. Problem Statements

Recently, researchers in digital communities have witnessed the tremendous growth of publications. Even though search engines on the Internet provide the efficient way for researchers to search publications of interests, the overwhelming amount of information still makes it a time-consuming task. Clustering, a technique used in any areas, is one way to facilitate this. Ontologies can also help in addressing the problem of searching related entities, including research publications.

Most of the existing text clustering methods use clustering techniques depends only on term strength and document frequency 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. 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. Moreover, this system wishes to utilize the ontology hierarchy structure it added the categorical information table before weighting phases. Also, the previous system meets some obstacle when the applied document sets become lagers [8]. To overcome this issue this system incorporated with the grouping algorithm [9].

4. Overview of the System

Searching the World Wild Web can be frustrating. Past studied have indicated that applying concept weight becomes biased towards some of the text documents when applying the tremendous testing data [8]. To counter act this problem we use grouping algorithm. Figure 1 provides main development of the system and also describes a detailed description of all the process that was taken out in this system.

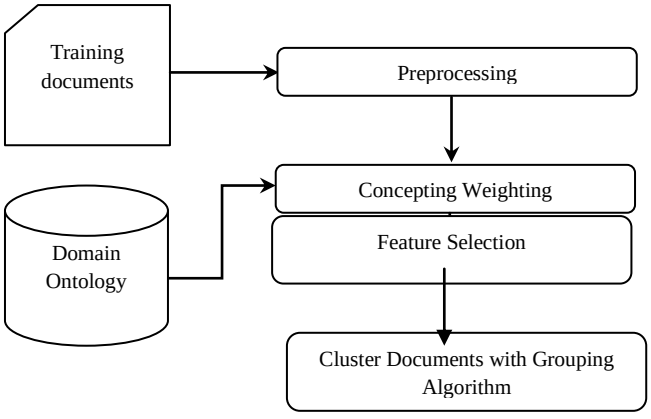

Fig 1: Overview of the System

The implementation of the system consists of five parts. The first part is ontology creation; the second part is weighting calculation. The rest of the path is for clustering. The goal of this research is the development of a domain-specific ontology, which will be used for technology clustering. This section presents a detailed explanation of the system work, which can be used in combination with ontology concepts. The work has been based on information extracted and inferred from Google Search Engine relating with the dissertation papers about image processing domain, distributing system and medical domain. With the growing demands in the research and development community of image research, distributed system and medical field, it is necessary to capture concept hierarchical data in order to provide an efficient means and efficient model of these areas of research. Therefore, the system creates an ontology which can be queried to gain knowledge for this research area and discovery has been conceived. The basic steps in building ontology are straightforward. The system has explored the ontology construction using text documents as shown in Figure 1.4. This ontology is captured in the OWL DL (Ontology Web Language Description Logics) language and supported by the current ontology editors, valuators, and reasoners.

4.1 Preprocessing Phase

The text document collection is the initial stage for this phase. In the preprocessing stage, the document is transferred to a format suitable to the representation process. The textual information is stored in many kinds of machine readable form, such as PDF, DOC, PostScript, HTML, and XML and so on. However, there are still a lot of papers stored in the plain pdf format. After the text document are collected from Google search engine, the abstract of the paper 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 stop-words are high frequent words that carry no information (i.e. pronouns, prepositions, conjunctions etc).

4.2 Concept Weighting Phase

In this phase the system calculate the weight of the concept as [8] as shown in below.

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

where W is the weight of keywords, Length is the depth of concept in the ontology. Frequency is the times which count the words appear in the document, and if the concept is in the ontology Correlation Coefficient is taken as 1 and otherwise 0. Probability is based on the probability of the concept in the document.

4.3 Clustering with Grouping Algorithm

The basic idea is that each text could gather its most related texts to form an initial group, Yllias Chali decide which groups have more strength over other groups, make the stronger groups as final clusters, and use them to bring any possible texts to their clusters. First, Yllias Chali's system uses each text as a leading text (Tl) to form a cluster. To do this, they put all the texts which have a score greater than the high-threshold with Tl into one group and add each score to the group's total score. By doing this for all texts, they will

have N possible different groups with different entries and group scores, where N is the number of the total texts in the set. Next, they select the final clusters from those N groups. They arrange all the groups by their scores in a non-increasing order. They choose the group with the highest score and check if any text in this group has been clustered to the existing final clusters or not. If not more than 2 texts are overlapping with the final clusters, then their algorithm take this group as a final cluster, and remove the overlapping texts from other final clusters. Yllias Chali 's stated that the process the group with the next highest score in the same way until the groups' entries are less than 4. For those groups, they would first try to insert their texts into the existing final clusters if they can fit in one of them. Otherwise, they will let them go to the leftover cluster which holds all the texts that do not belong to any final clusters. After the concept weighting phase we apply the grouping algorithm. The following is the pseudocode for the grouping algorithm , Yllias Chali applied in their system:

// Get the Initial Clusters

For each text t i

Construct a text cluster including all the texts (t j) which score (ti,tj)>=high threshold;

Compute the total score of the text cluster;

Find out its neighbor with maximum relation score;

End For

//Build the final clusters

Sort the clusters by their total score in non-increasing order;

For each cluster g i in the sorted clusters

If member gi >3 and overlap-mean gi <=2

Take g i as a final cluster c i;

Mark all the texts in ci as clustered;

Else

Skip to process next cluster;

End If

End For

//Process the leftover texts and insert them into one of the final clusters

For each text t j

If t j has not been clustered

Find cluster c i with the highest score (ci, tj);

If the average-score (ci, tj) >= low- threshold

Put tj into cluster ci;

Else If the max score neighbor tm of tj is in ck

Put tj into cluster ck;

Else

Put tj into the final leftover cluster;

End if

End if

End For

Output the final clusters and the final leftover cluster;

5. Experimental Results

The proposed system has been tested with three test cases. The experiments in this section are conducted on the papers that are downloaded from the Google. The system downloaded 2000 papers from the Google Search track of recent World Wide Web conference websites. These 2000 test documents came from three subcategories (types) of Image documents, Distributed System documents and Medical related documents respectively.

Table 1 Testing Results

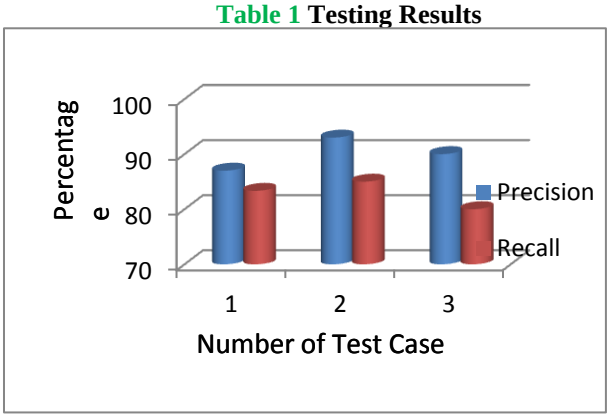

Table 1 shows the results of the three dataset's statistical relatedness analysis measures using precision and recall rate. As expected, the highest scoring produce the highest precision, which shows that the system score is a good measure of the degree of relevancy between concepts and the documents. The performance of the method is influenced by a number of factors. The CPU requirements for the experiments described above are of the order of 2-3 hours. The memory requirements are quite excessive, and there is a trade-off between the number of abstracts (instances) and the number of concepts (features).

6. Performance Analysis

The performance of the method is influenced by a number of factors. The system processing time may be slow degradation caused by the continuous growth of the ontology size and sudden improvement gains due to more successful arrangements of concepts and clustering methods has be seen. The CPU requirements for the experiments described above are of the order of 2-3 hours. The memory requirements are quite excessive, and there is need to be trade-off between the number of documents and the number of concepts if the processing time needs to be decreased. The concept/documents ratio may have to be reduced for very large-scale experiments that influence the ontology. The system performance also degrades if the number of documents is increased and if the ontology size is growing.

From the results of this research it has been shown that the idea of using ontology to represent the clustering document in place of its concept weighting, as conventionally is used, is a sound principle under certain conditions. The dominating condition with regard to the test dataset and the testing methods were the length of the article i.e. the number of words occurring in the document. The achieved results indicate that as the average number of words in a document corpus increases the less of an impact that the ontology methods. In conclusion, the ontology-based approach performed as good even it have some minor differences than traditional clustering method and statistically method.

The experimental results show that the concept weighting based approach is indeed helpful in clustering with the support of the ontology. But the improvement is too small to be cost effective for such an ontology based approach.

7. Conclusion

The main aim of this work has been the development of a methodology able to exploit ontological computing when used in clustering methods, called ontology-based clustering. This work is a contribution in the field of ontology, in which the system has studied how domain knowledge can been exploited before the clustering process. Moreover, these approaches do not attempt to interpret the conceptual meaning of textual terms, which is crucial in many applications related to textual data. In this work, the system has focused on applying semantic issue. Moreover, we applied medical domain related papers in this system to test whether the system can give accurate cluster and also required medical knowledge is acquired from the medical knowledge obtained from our second author. As expected as earlier, the experimental results illustrating the effectiveness of our technique. Therefore we should extend this technique more statistically for further experiments to conduct more accurate text document clustering because the author interests area is the clustering system which can catch up with google clustering method which was hit 90% in this year for all domain and they used in the Web documents clustering.

A key aspect in the clustering process is the way in which concept are evaluated and compared. For doing so, ontology-based clustering have been studied. Also, the use of ontologies has been exploited as more suitable for clustering purposes. The system observed that the clustering results are affected by the degree of completeness of the ontology. For the case of evaluating the behavior of the system, three test cases were analyzed. The system also observed that the more accurate the ontology the more accurate the clustering results. Successful results were obtained with the domain ontology. In that sense, the proposed system clustering approach can be used in several tasks and domains such as electronic commerce (e.g. grouping similar products or to obtain a characterization of users), medicine (e.g. clustering of electronic health records), tourism (recommending tourist destinations to users), even in privacy preserving.

Future work includes the idea to acquire knowledge from the domain expert. The domain expert will validate the extracted concept Also, the methodology for building ontology from unstructured data such web pages and documents. Another direction is to link the system to the web document clustering and to construct the domain ontology automatically.

8. REFERENCES

[1] M. Jursic , N. Lavrac, “Fuzzy Clustering Of Documents”, Department of Knowledge Discovery.

[2] M. SteinBach, G. Karypis and V. Kumar, “A Comparison of Document Clustering Techniques” in KDD Workshop on Text Mining, 2000.

[3] M-L. Reinberger, P. Spyns, “Discovering Knowledge in Texts for the Learning of DOGMA-inspired ontologies”. In Proceedings of ECAI 2004 Workshop on Ontology Learning and Population, 2004.

[4] N. F. Noy and D.L. McGuiness. “Ontology Development 101: A Guide to Creating Your First Ontology”, Stanford Knowledge Systems Laboratory Technical Report KSL-01-05 and Stanford Medical Informatics Technical Report SMI-2001-0880, March 2001.

[5] S. Bechhofer, I. Horrocks, C. Goble, R. Stevens, “ OILEd: A Reasonable Ontology Editor for the Semantic Web” , In KI2001, Joint German/Austrian conference on Artificial Intelligence, volume LNAI Vol. 2174, pages 396-408, Vienna ,2001.

[6] S. Karapiperis and D. Apostolou, “Consensus Building in Collaborative Ontology Engineering Processes”, Journal of Universal Knowledge Management, 1(3), 199-216, 2006

[7] T. Berners-Lee, “Weaving the Web”, Harper, San Francisco, 1999, HarperCollins Publishers, New York, NY, 1999.

[8] H. H. Tar, T. T. S. Nyunt, “Ontology-based Concept Weighting for Text Documents”, World Academy of Science, Engineering and Technology.

[9] Yllias Chali and Soufiane Noureddine, “Document Clustering with Grouping and Chaining Algorithms”, University of Lethbridge.