Journal of Engineering and Applied Sciences 7 (4): 342-347, 2012
ISSN: 1816-949X

Study of Ontology or Thesaurus Based Document Clustering and Information Retrieval

G. Bharathi and D. Venkatesan
School of Computing, SASTRA University, Tamil Nadu, India

Abstract: Document clustering generate clusters from the whole document collection automatically and is used in many fields including data mining and information retrieval. 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. In the traditional vector space model, the unique words occurring in the document set are used as the features. But because of the synonym problem and the polysemous problem such a bag of original words cannot represent the content of a document precisely. Most of the existing text clustering methods use clustering techniques which depend 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. To overcome these issues, this study makes a survey of recent research done on ontology or thesaurus based document clustering.

Key words: Ontology, thesaurus, document clustering, intelligent information retrieval, semantics, Wikipedia, Wordnet

INTRODUCTION

Data clustering partitions a set of unlabeled objects into disjoint/joint groups of clusters. In a good cluster, all the objects within a cluster are very similar while the objects in other clusters are very different. When the data processed is a set of documents, it is called document clustering. Document clustering is very important and useful in the information retrieval area. Document clustering can be applied to a document database so that similar documents are related in the same cluster. During the retrieval process, documents belonging to the same cluster as the retrieved documents can also be returned to the user. This could improve the recall of an information retrieval system.

Document clustering can also be applied to the retrieved documents to facilitate finding the useful documents for the user. Generally, the feedback of an information retrieval system is a ranked list ordered by their estimated relevance to the query. When the volume of an information database is small and the query formulated by the user is well defined, this ranked list approach is efficient. But for a tremendous information source, such as the world wide web and poor query conditions (just one or two key words), it is difficult for the retrieval system to identify the interesting items for the user. Sometimes most of the retrieved documents are of no interest to the users. Applying documenting clustering to the retrieved documents could make it easier for the users to browse their results and locate what they want quickly. A successful example of this application is Vivisimo (http://vivisimo.com/) which is a web search engine that organizes search results with document clustering.

Generally, clustering is used in statistics to discover the structure of large multivariate data sets. It can often reveal latent relationships hidden in complex data. Within information retrieval, clustering (of documents) has several promising applications, all concerned with improving efficiency and effectiveness of the retrieval process. Some of the more interesting include:

Finding similar documents to a given document: This feature is often used when the user has spotted one good document in a search result and wants more-like-this. The interesting property here is that clustering is able to discover documents that are conceptually alike in contrast to search-based approaches that are only able to discover whether the documents share many of the same words.

Search result clustering: This allows the user to get a better overview of the documents returned as results in the search and to navigate towards clusters that are relevant to the user's information need.

Guided/interactive search: Here clustering is used to help the user drill down and find the desired information step-by-step by gradually refining the search.

Corresponding Author: G. Bharathi, School of Computing, SASTRA University, Tamil Nadu, India

Organising site content into categories: Allows browsing of the site in a Yahoo-like fashion.

Recommender system: This based on the documents, the user has already visited, recommends other documents. A typical use of this is in an e-Commerce setting where products that might interest the customer are suggested based on products the user has already examined/bought.

Faster/better search: Utilizes the clustering to optimise the search. A user query could for instance be compared to clusters instead of the individual documents, effectively limiting the search space.

Many clustering techniques have been proposed in the literature. Clustering algorithms are mainly categorized into hierarchical and partitioning methods. 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 are the most well-known partitioning methods. Semantic document clustering provides a means of clustering documents on the basis of the actual content inside the document. One important prerequisite for semantic analysis is the construction of a large-scale thesaurus and dictionary.

Currently, one of the most widely used thesaurus for English is WordNet. The Vector Space Model is a widely used method for document representation in information retrieval. In this model, each document is represented by a feature vector. The unique terms occurring in the whole document collection are identified as the attributes (or features) of the feature vector. Different term weighting methods may be used in the Vector Space Model such as the Binary Method, tf (term frequency) Method (Salton and Buckley, 1988) and tf-idf (inverse document frequency) (Salton, 1971) Method. Traditionally, the single words or compound words occurring in the document set are used as the features. Because of the synonym and polysemous problem, generally such a bag of words cannot reflect the semantic content of a document (Han and Kamber, 2006). One way to resolve this problem is to enrich document representation with the background knowledge represented by ontology.

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 (Decker *et al*., 2000; Ding and Foo, 2002; Hotho *et al*., 2001). 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 (Berners-Lee, 1999). 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.

Overview of ontology: Top level ontology or upper level ontology 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, e.g., 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 (Fig. 1) (Steinbach *et al*., 2000).

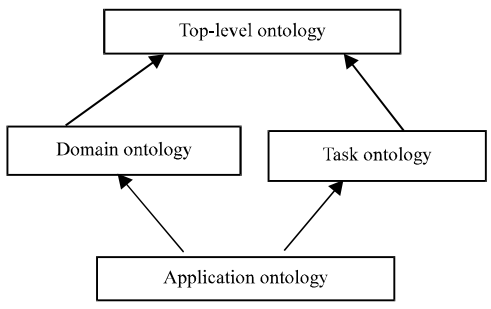

Fig. 1: Categorization of ontology

Ontology usually includes at least three components: concepts, attributes and the relationships among concepts. All of them can be used for document representation and clustering. The most common way of applying ontologies for clustering is to match ontology concepts to the topical terms appearing in the documents. Then the matched ontology concepts are either used as replacement or introduced as additional features to the original text. Further, the attributes of and relationships among the ontology terms can be exploited for clustering.

PROBLEMS IN ONTOLOGICAL APPROACH

Hu *et al.* (2009) say that a major problem of this ontological approach for document clustering is that it is usually difficult to find a comprehensive ontology which can cover all the concepts mentioned in a collection, especially when the documents to be clustered are from general domain. Previous research has adopted WordNet (Hotho *et al.*, 2001, 2003) and Mesh (Yoo *et al.*, 2006; Zhang *et al.*, 2007) as the external ontology for text enrichment. However, they all have limited coverage. Another problem is that using ontology terms either as replacement or additional features has its disadvantages. While replacing original content with ontology terms may cause information loss, especially when the coverage of the ontology is limited, adding ontology terms to the original document vector can bring data noise into the dataset. Therefore in order to enhance text clustering by leveraging ontology semantics, two issues need to be addressed: an ontology which can cover the topical domain of individual document collections as completely as possible and a proper matching method which can enrich the document representation by fully leveraging ontology terms and relations but without introducing more noise.

Hu *et al.* (2009) address both issues. In terms of ontology, they rely on Wikipedia concepts and categories for document enrichment. Wikipedia has become the largest electronic knowledge repository on the web with millions of articles contributed collaboratively by volunteers. Unlike other standard ontologies such as WordNet and Mesh, Wikipedia itself is not a structured thesaurus. However, it is much more comprehensive and up to date.

WIKIPEDIA AS AN ONTOLOGY

In Wikipedia, each study only describes a single topic. The title of each study is a succinct phrase that resembles an ontology term. Equivalent concepts are grouped together by redirected links. Meanwhile, it contains a hierarchical categorization system, in which each study belongs to at least one category. All these features make Wikipedia a potential ontology which can be exploited for enriching text representation and enhancing text clustering. As for how to integrate ontology concepts into the document representation and clustering process (Hu *et al.*, 2009), they propose two approaches for mapping ontology concepts to the documents. The first approach, called exact-match is a dictionary-based approach. It maps the topical terms present in the documents directly to Wikipedia concepts. It is especially useful when Wikipedia concepts can cover most of the topic terms in a collection.

The second mapping approach is called relatedness match. Instead of mapping Wikipedia concepts to each document directly; this approach builds the connection between Wikipedia concepts and each document based on the contents of Wikipedia articles. This approach is more useful when Wikipedia concepts cannot fully cover the topical domain of a collection. After the mapping process, each document is associated with a set of concepts. Then based on the hierarchical structure of Wikipedia, each document is further mapped to a set of Wikipedia categories. Finally, the text documents are clustered based on a similarity metric which combines document content information, concept information as well as category information.

CONCEPT OR FEATURE WEIGHTING

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 integrates core ontologies as background knowledge into the process of clustering (Zhang and Wang, 2010; Maedche and Zacharias, 2002).

Jing *et al.* (2005) 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.

Hmway Hmway Tar and Thi Thi Soe Nyunt proposed a system which has been 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

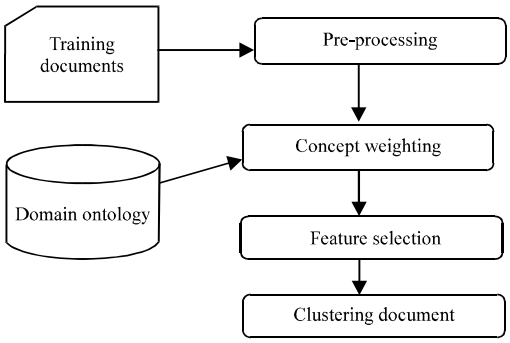

Fig. 2: Concept weighting

clustering at the conceptual level. The system uses the text documents for the clustering process. Here the 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. Figure 2 shows the overview of the proposed system architecture.

SEARCH RESULTS CLUSTERING

Web search results clustering is an increasingly popular technique for providing useful grouping of web search results or snippets into clusters. The Lingo algorithm, proposed by Stanislaw Osinski and Dawid Weiss, uses frequent phrases to identify candidate cluster labels and then assigns snippets to these labels. Sameh and Kadray (2010) extends on the Lingo algorithm by adding semantic recognition to the frequent phrase extraction phase. This is achieved by finding the synonyms of frequent words in the WordNet database and adding the synonyms to the pool of frequent terms that comprise the cluster label candidates. The detection of synonyms helps in grouping together snippets that contain different but synonymous words that would otherwise have not been grouped together using the original Lingo algorithm.

The study's contribution is adding semantic recognition to enable the recognition of synonyms in snippets, thus improving the quality of the clusters generated. The semantic recognition is achieved using the WordNet database which is a lexical database for the English language, in which nouns, verbs, adjectives and adverbs are grouped into sets of cognitive synonyms (synsets), each expressing a distinct concept (Hua-Jun *et al*., 2004).

In WordNet, each word can be a part of various synsets, each synset relating the word to other words with the same conceptual meaning. For example, the word doctor is part of a synset that relates it to the word physician for the meaning of a licensed medical practitioner for this meaning, the words doctor and physician are synonymous. Similarly, doctor is also related to Dr. in the synset that has the meaning of a person who holds PhD degree (or the equivalent) from an academic institution.

FREQUENT CONCEPTS BASED DOCUMENT CLUSTERING (FCDC)

In English dictionary, most of the words have multiple synonyms, therefore it is possible that two different documents which have no common words may represent the same topic. As Baghel and Dhir (2010), concepts are the set of synonym words which have the same meaning. The proposed FCDC first searches the concepts in documents and then finds the frequent concepts by apriori paradigm (Agrawal and Srikant, 1994; Agrawal *et al*., 1993). Finally, it forms the initial clusters of documents with each cluster representing a single frequent concept. Then, the proposed algorithm processes these initial clusters to create disjoint clusters. The final results are represented using the hierarchical tree like structure.

The proposed document clustering algorithm consists of the following phases: finding frequent concepts using apriori algorithm, creating initial clusters for each frequent concept, making clusters disjoint using score function, building cluster tree and tree pruning.

INTELLIGENT INFORMATION RETRIEVAL USING DOMAIN ONTOLOGY

A digital library is a type of Information Retrieval (IR) system. Swe (2011) proposed a model which uses concept-based approach (ontology) and metadata case base. This model consists of identifying domain concepts in user's query and applying expansion to them. The system aims at contributing to an improved relevance of results retrieved from digital libraries by proposing a conceptual query expansion for intelligent concept-based retrieval.

There is a need to import the concept of ontology, making use of its advantage of abundant semantics and standard concept. Domain specific ontology can be used to improve information retrieval from traditional level based on keyword to the lay based on knowledge (or concept) and change the process of retrieval from traditional keyword matching to semantics matching. One approach is query expansion techniques using domain ontology and the other would

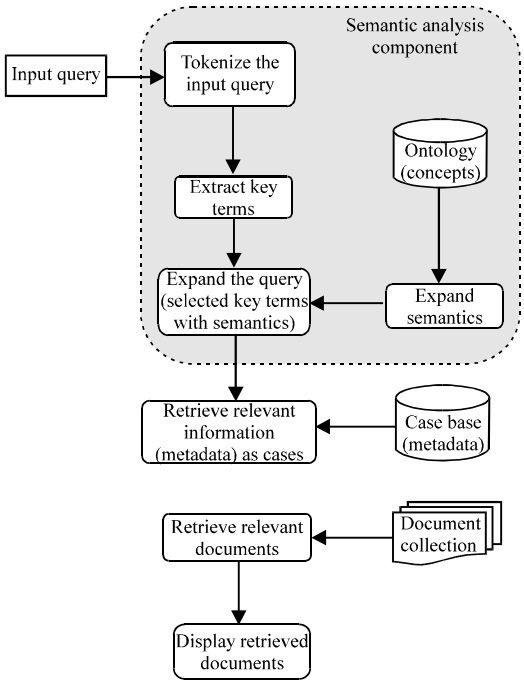

Fig. 3: Intelligent information retrieval

be introducing a case based similarity measure for metadata information retrieval using Case Based Reasoning (CBR) approach (Fig. 3).

CONCLUSION

Thus in this study, researchers make a survey of recent methodologies and approaches that are followed, used and developed to improve document clustering and information retrieval processes by using ontology or thesaurus as background knowledge.

REFERENCES

Agrawal, R. and R. Srikant, 1994. Fast algorithms for mining association rules. Proceedings of the 20th International Conference on Very Large Data Bases, September 12-15, 1994, San Francisco, USA., pp: 487-499.

Agrawal, R., T. Imielinski and A. Swami, 1993. Mining association rules between sets of items in large databases. Proceedings of the ACM SIGMOD International Conference on Management of Data, May 25-28, 1993 Washington, DC., USA., pp: 207-216.

Baghel, R. and R. Dhir, 2010. Text document clustering based on frequent concepts. Proceedings of the1st International Conference on Parallel, Distributed and Grid Computing, October 28-30, 2010, Solan, pp: 366-371.

Berners-Lee, T., 1999. Weaving the Web. Harper, San Francisco.

Decker, S., S. Melnik, F. Van Harmelen, D. Fensel and M. Klein *et al*., 2000. The semantic web: The roles of XML and RDF. IEEE Int. Comp., 4: 63-74.

Ding, Y. and S. Foo, 2002. Ontology research and development: Part 1-A review of ontology generation. J. Inf. Sci., 28.

Han, J. and M. Kamber, 2006. Data Mining: Concepts and Techniques. 2nd Edn., Morgan Kaufmann Publisher, San Fransisco, USA., ISBN: 1-55860-901-6.

Hotho, A., A. Maedche and S. Staab, 2001. Text clustering based on good aggregations. Proceedings of the 2001 IEEE International Conference on Data Mining, November 29-December 2, 2001, San Jose, pp: 607-608.

Hotho, A., S. Staab and G. Stumme, 2003. Wordnet improves text document clustering. Proceedings of the SIGIR 2003 Semantic Web Workshop, (SWW'03), Toronto, Canada, pp: 541-544.

Hu, X., X. Zhang, C. Lu and X. Zhou, 2009. Exploiting wikipedia as external knowledge for document clustering. Proceedings of the KDD'09, June 28-July 1, 2009, Paris, France.

Hua-Jun, Z., H. Qi-Cai, Z. Chen, M. Wei-Ying and J. Ma, 2004. Learning to cluster web search results. SIGIR'04, Sheffield, South Yorkshire, UK.

Jing, L., M.K. Ng, J. Xu and Z. Huang, 2005. Subspace clustering of text documents with feature weighting k-means algorithm. Proc. PAKDD, 3518: 802-812.

Maedche, A. and V. Zacharias, 2002. Clustering ontology-based metadata in the semantic web. Proceedings of the 6th European Conference on Principles and Practice of Knowledge Discovery, August 19-23, 2002, Helsinki, Finland, pp: 348-360.

Salton, G. and C. Buckley, 1988. Term-weighting approach in automatic text retrieval. Inf. Proc. Manag., 24: 513-523.

Salton, G., 1971. The Smart Retrieval System Experiments in Automatic Document Retrieval. Prentice Hall Inc., New Jersey, Pages: 556.

Sameh, A. and A. Kadray, 2010. Semantic web search results clustering using lingo and word net. Int. J. Res. Rev. Comp. Sci., 1.

Steinbach, M., G. Karypis and V. Kumar, 2000. A comparison of document clustering techniques. Proceedings of the 6th ACM SIGKDD World Text Mining Conference, August 20-23, 2000, Boston, pp: 1-2.

Swe, T.M.M., 2011. Intelligent information retrieval within digital library using domain ontology. Comp. Sci. Inf. Technol.

Yoo, I., X. Hu and I.Y. Song, 2006. Integration of semantic-based bipartite graph representation and mutual refinement strategy for biomedical literature clustering. Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, August 20-23, 2006, Philadelphia, USA, pp: 791-796.

Zhang, L. and Z. Wang, 2010. Ontology-based clustering algorithm with feature weights. J. Comp. Inf. Syst., 6: 2959-2966.

Zhang, X., L. Jing and X. Hu, 2007. A comparative study of ontology based term similarity measures on documentclustering. Proceedings of 12th International Conference on DatabaseSystems for Advanced Applications, April 9-12, 2007, Bangkok, Thailand, pp: 115-126.