Impact of Ontology based Approach on Document Clustering

S.C. Punitha
HOD, Department of Computer Science, P.S.G.R. Krishnammal College for Women, Coimbatore, India.

K. Mugunthadevi
Mphil Scholar, P.S.G.R. Krishnammal College for Women, Coimbatore, India.

M. Punithavalli
Director, Department of Computer Science, Sri Ramakrishna College of Arts and Science for Women, Coimbatore, India

ABSTRACT
Document clustering is considered as an important tool in the fast developing information explosion era. It is the process of grouping text documents into category groups and has found applications in various domains like information retrieval, web or corporate information systems. Ontology-based computing is emerging as a natural evolution of existing technologies to cope with the information onslaught. This paper discusses the concepts behind ontology-based document clustering and compares the performance with existing traditional system. The results prove that introducing ontology concepts with document clustering is promising and improves clustering process.

Keywords : Clustering, Document Clustering, Ontology, Similarity Measure, Text Mining.

1. INTRODUCTION
In the fast developing information explosion era, much of the knowledge available is stored as text. It is not surprising, therefore, that data mining (DM) and information retrieval (IR) from text collections (text mining) has become an active and exciting research area. Clustering or segmentation of data is a fundamental data analysis step that has been widely studied across multiple disciplines for over 40 years. Clustering text documents into different category groups is an important step in indexing, retrieval, management and mining of abundant text data on the Web or in corporate information systems.

Current clustering methods can be divided into generative (model-based) approaches Cadez *et al.*, 2000 [1] and discriminative (similarity-based) approaches Karypis *et al.*,1999 [8]. Parametric, model-based approaches attempt to learn generative models from the data, with each model corresponding to one particular cluster. In similarity-based approaches, one determines a distance or similarity function between pairs of data samples, and then group similar samples together into clusters. While considering solutions to document clustering problem, there are many algorithms for automatic clustering like the K Means algorithm, Expectation be applied to a set of vectors to form the clusters. Traditionally the document is represented by the frequency of the words that make up the document (the Vector space model and the Self-organizing semantic map). Different words are then given importance according to different criteria like Inverse Document frequency and Information Gain. A comparative evaluation of feature selection methods for text documents can be found in Yang and Pedersen,1997 [13]. These methods consider the document as a bag of words, and do not exploit the relations that may exist between the words.

The rapidly growing availability of large tracts of textual data such as online news feeds, blog postings, emails, and discussion board messages, has made the need for improved text clustering an important current research area. However, despite the extensive research, clustering unstructured, textual information remains a challenging problem. For example, the nature of the unstructured textual information makes it hard for current clustering algorithms to capture the intrinsic structure that is desired Geo *et al.*, 2006 [5]. Individual data sets also have unique characteristics, which add more complexity to mapping or deciding upon the clustering methodology that works best for a particular data set. Moreover, the lack of labeled examples in unsupervised clustering make the partitioning task an ill-posed problem since there is no adopted methodology well known to produce the ideal clustering. To overcome these challenges, researchers have begun to investigate alternative clustering approaches that incorporate background knowledge to guide each partitioning task and thus alleviate the difficulty of finding a single, best approach Hotho *et al.*, 2003; Sedding and Kazakov, 2004 [7],[10]. Thus, the most challenging problems of text clustering are big volume, high dimensionality and complex semantics. Moreover, traditional clustering algorithms have the disadvantage that they do not understand the text. For example, consider two sentences “Mr. A and Mr. B are standing near Neem tree” and “The Neem tree is near to the place where Mr. A and Mr. B is standing”. Both the sentences mean the same. Similarly, the two sentences “Mr. A is intelligent” AND “Mr. A is brilliant” mean the same but are constructed using different synonymous words. Latent Semantic Indexing Deerwester *et al*., 1990 [2] uses a word category map to solve such problems in text clustering. But the drawback here is that due to polysemy or homography, where a word with different meanings or meaning shades in different contexts (Example: “Lots of money from bank” and “Boat beside the river bank”). Recent works has shown that ontology is useful to improve the performance of text clustering in these situations.

The primary objective of this paper is to understand the basic concepts behind ontology with particular emphasis on its application to document clustering problem. For this purpose, the paper explains the general concepts behind ontology in Section II, followed by a general description of document clustering in Section III. Section IV explains the working of ontology-based clustering. Section V concludes the study.

2. ONTOLOGY

The term "ontology" has been used for a number of years by the artificial intelligence and knowledge representation community but is now becoming part of the standard terminology of a much wider community including information systems modelling. The term is borrowed from philosophy, where ontology means 'a systematic account of existence'.

Ontology is "the specification of conceptualisations, used to help programs and humans share knowledge". Ontology is a set of concepts - such as things, events, and relations that are specified in some way in order to create an agreed-upon vocabulary for exchanging information. Ontology's establish a joint terminology between members of a community of interest. These members can be human or automated agents.

In information management and knowledge sharing arena, ontology can be defined as follows:

- Ontology is a vocabulary of concepts and relations rich enough to enable us to express knowledge and intention without semantic ambiguity.
- Ontology describes domain knowledge and provides an agreed-upon understanding of a domain.
- Ontology: are collections of statements written in a language such as RDF that define the relations between concepts and specify logical rules for reasoning about them.

Mathematically it can be defined Yang *et al*., 2008 [12] as follows:

"An ontology can be defined as an Vector O: = (C, V, P, H, ROOT), where C is the set of concepts, V ($v_i \subset C$) contains a set of terms and is called the vocabulary, P is the set of properties fore each concept, H is the hierarchy and ROOT is the topmost concept. Concepts are taxonomically related by the directed, acyclic, transitive, reflexive relation $H \subset C * C$. $H(c_1, c_2)$ shows that c_1 is a subclass of c_2 and for all $c \subset C$ it holds that H(c, ROOT)."

Ontology is an explicit and formal specification of a conceptualization Gruber,1993 [6] . Ontology defines as a common vocabulary for researchers who need to share information in a domain. It includes machine interpretable definitions of basic concepts in the domain and relations and has become common on the World-Wide Web. An example of a basic ontology is shown in Figure 1.

Ontology describes the relationships between entities on a conceptual level. It shows the hierarchy of classes and subclasses for an object-entity, for example (computer). It describes subclass relationships disjointness, constraints, and information between objects. It provides vital information to search agents, intelligent agents and databases.

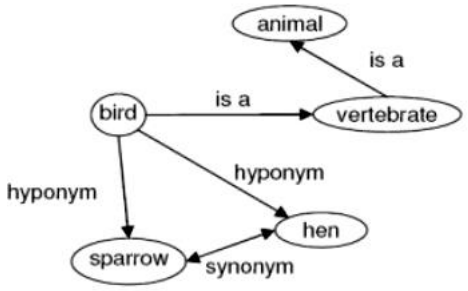

Figure 1 : Ontology – An Example

2.1. Terms and Definition

This section describes some of the commonly used terms along with their meaning with respect to ontology.

- Concept : An idea or thought that corresponds to some distinct entity or class of entities, or to its essential features, or determines the application of a term, and thus plays a part in the use of reason or language
- Holonym : A concept of which this concept forms a part
- Hypernym : Word with a broad meaning which more specific words fall under: a super ordinate
- Hyponym : Word of more specific meaning
- Meronym: A term that denotes part of something: a member of an information set
- Ontology: The branch of metaphysics dealing with the nature of being
- Semantic: Relating to meaning in language or logic
- Synonym: A word or phrase that means exactly or nearly the same as another word or phrase in the same language
- Whole:A term used to identify a concept that consists of multiple parts

The relationship between the component parts of the semantic model is shown in Figure 2.

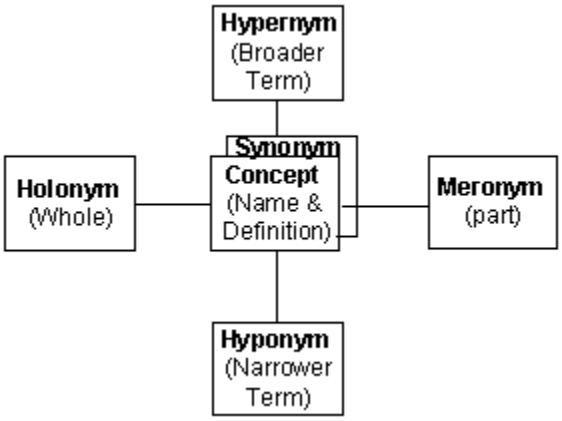

Figure 2 : Relationship between Ontology Components

2.2. Benefits of Ontology

Ontology provides many benefits as listed below.

- To facilitate communications among people and organisations
    - Aid to human communication and shared understanding by specifying meaning
- To facilitate communications among systems with out semantic ambiguity. i.e. to achieve inter-operability
- To provide foundations to build other ontology (reuse)
- To save time and effort in building similar knowledge systems (sharing)
- To make domain assumptions explicit
    - Ontological analysis
        - Clarifies the structure of knowledge and allow domain knowledge to be explicitly defined and described.

2.3. Application Areas of Ontologies

Usage of Ontology's has been prominent in various fields and some of them are listed below.

- Information Retrieval - As a tool for intelligent search through inference mechanism instead of keyword matching, Easy retrievability of information without using complicated Boolean logic, Cross Language Information Retrieval, Improve recall by query expansion through the synonymy relations, Improve precision through Word Sense Disambiguation (identification of the relevant meaning of a word in a given context among all its possible meanings)
- Digital Libraries - Building dynamical catalogues from machine readable meta data, Automatic indexing and annotation of web pages or documents with meaning, To give context based organisation (semantic clustering) of information resources, Site organization and navigational support
- Information Integration - Seamless integration of information from different websites and databases
- Knowledge Engineering and Management -As a knowledge management tools for selective semantic access (meaning oriented access), Guided discovery of knowledge
- Natural Language Processing - Better machine translation, Queries using natural language

3. GENERAL DOCUMENT CLUSTERING FRAMEWORK

The major concern in information retrieval and text mining area is the question of finding the best method to explore and utilize the huge amount of text documents. Document clustering helps users to effectively navigate, summarize, and organize text documents. By organizing a large amount of documents into a number of meaningful clusters, document clustering can be used to browse a collection of documents or organize the results returned by a search engine in response to a user's query. Using clustering techniques to group documents can significantly improve the precision and recall in information retrieval systems and it is an efficient way to find the nearest neighbors of a document. A general definition of clustering as stated by Everitt *et al*. (2001) [4] is given below.

"Given a number of objects or individuals, each of which is described by a set of numerical measures, devise a classification scheme for grouping the objects into a number of classes such that objects within classes are similar in some respect and unlike those from other classes. The number of classes and the characteristics of each class are to be determined". A document clustering techniques performs the desired clustering activity in three stages (Figure 3).

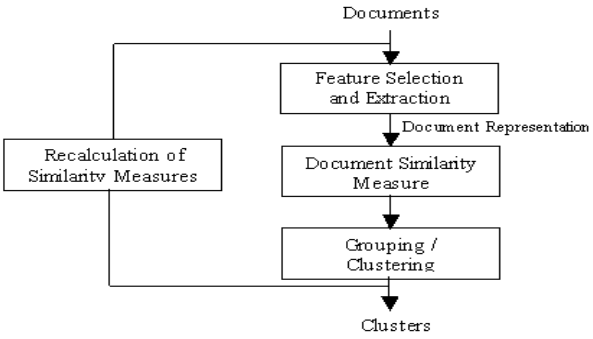

Figure 3 : Stages in Document Clustering

Document representation refers to the number of clusters, the number of documents, and the number, type and scale of the features available to the clustering algorithm. Feature selection is the process of identifying the most effective subset of the original features to use in clustering. Feature extraction is the use of one or more transformations of the input features to produce new salient features. Either or both of these techniques can be used to obtain an appropriate set of features to use in clustering. Document similarity is usually measured by a pair-wise similarity function. A simple similarity measure, like cosine function, is often used to reflect the similarity between two documents. The grouping step of text clustering can be performed in a number of ways. Three methods namely, traditional K-Means, Ontology-based and Hybrid technique that combines pattern recognition and clustering are studied in this research. The performance of text clustering algorithm could be evaluated by the cluster validity analysis, which is the assessment of a clustering procedure's output. There are three types of validation studies. An external assessment of validity compares the recovered structure to a-priori structure. An internal examination of validity tries to determine if the structure is intrinsically appropriate for the data. A relative test compares two structures and measures their relative merit.

4. ONTOLOGY-BASED DOCUMENT CLUSTERING

The main motivation behind ontology is that different people have different needs with regard to the clustering of texts. Empirical and mathematical analysis has shown that clustering in a high-dimensional space is very difficult and explanation why particular texts were categorized into one cluster is required. The goal of cluster analysis is the division of a set of objects into homogeneous clusters. The general steps followed by ontology-based clustering algorithms are given below.

1) Calculate distance matrix (or similarity matrix) between every pair of objects using ontology-specific methods. Here, every object constitutes a separate cluster (obtaining similarity matrix).
2) Using distance matrix, merge the two closest clusters (clustering process)
3) Modify or rebuilt distance matrix, by treating merged clusters as one object. Methods that calculate similarity between an object and a cluster and methods that estimate similarity between clusters and ontology objects are used for this purpose (evaluation process).
4) If the desired number of clusters have been reached, then stop else go to Step 2.

The similarity between the objects is normally calculated using Equation (1).

$$Sim(I_i, I_j) = f_{agr}(TS(I_i, I_j), RS(I_i, I_j), AS(I_i, I_j)) \quad (1)$$

where TS is the taxonomy similarity, RS is the relationship similarity and AS is the attribute similarity. TS is the similarity or dissimilarity between classes on the scheme and can be calculated in many ways. Some examples are Wu-Palmer measure Wu and Palmer,1994 [11]. The idea of the relationship similarity is very simple. Similar objects should have relationships with objects that are similar to each other. When two objects O_1 and O_2 are compared, it should indicate all objects that have relationships with object O1 and all objects that have relationships with O_2, calculate taxonomy similarity and/or attribute similarity between these two sets of objects and finally aggregate calculated similarities. The estimation of attribute similarity depends on the data types of the objects. As text documents have only strings, a lexical similarity measure is often used Euzenat and Shvaiko,2007 [3] . Another method is to use some distance measure like Euclidean distance or as one proposed by Manning and Schutze,1999 [9].

For clustering process, the traditional K-means algorithm is often used. After implementing the first prototype following main benefits have been achieved:

- The process of aggregating is automated and has reduced the manual operation and therefore reduced the costs.
- Retrieving of data and creating different analysis provides a higher precision.
- Different documents from various content sources are connected based on their content, meaning on their semantics that has been automatically extracted.

5. EXPERIMENTAL RESULTS

This section reports experimental results when applying the basic ontology algorithm to cluster documents. During experimentation, Reuters-21578 dataset was used. More information about Reuters-21578 can be found at http://www.daviddlewis.com/resources/testcollections/reuters21578/ readme.txt. To ascertain the performance of the models, several experiments were conducted. All the experiments were conducted using a Pentium IV machine with 2GB RAM. Three performance metrics, namely, purity of a cluster, F-measure and CPU execution time were used. The results were compared with the traditional K-means clustering algorithm. The overall purity obtained for the three algorithms for different number of clusters is shown in Table I.

Table 1: Purity of a Cluster

No. of Clusters	K-Means	Ontology
20	0.66	0.75
40	0.68	0.81
60	0.69	0.83
80	0.70	0.85
100	0.72	0.88

The F-measure calculated from the precision and recall is shown in Table II.

Table 2: Accuracy of the Algorithm

Algorithm	Precision	Recall	F Measure
K-Means	0.515	0.832	0.64
Ontology	0.698	0.902	0.79

While considering the time taken or speed of clustering, it was found that the ontology-based algorithm is fast and takes only 79.66 minutes on average while tested with the Reuters dataset. The K-means algorithm took 98.77 minutes, which is slow when compared with ontology-based algorithm.

All these results from the various experiments show that the clustering algorithm that uses semantics of the documents, that is, ontology-based clustering produces significant improvement in clustering results when compared with traditional existing algorithm and therefore proves to be a promising field of research in terms of text mining.

6. CONCLUSION

As the volume of information continues to increase, there is growing interest in helping people better find, filter and manage these resources. Text clustering, which is the process of grouping documents having similar properties based on semantic and statistical content, is an important component in many information organization and management tasks. Ontology-based computing is emerging as a natural evolution of existing technologies to cope with the information onslaught. Future work is planned in comparing the performance of document clustering when various similarity measures and clustering algorithms are combined with ontology features of documents.

7. REFERENCES

[1] Cadez, I.V., Gaffney, S. and Smyth, P. (2000) A general probabilistic framework for clustering individuals and objects, Proc. 6th ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining, Pp.140–149.

[2] Deerwester, S., Dumais, S.T., Furnas, G.W., Landauer, T.K. and Harshman, R. (1990) Indexing by Latent Semantic Analysis, Journal of the American Society of Information Science.

[3] Euzenat, J. and Shvaiko, P. (2007) Ontology Matching, Springer-Verlag. Berlin Heidelberg.

[4] Everitt, B.S., Landau, S. and Leese, M. (2001) Cluster Analysis, Oxford University Press, Fourth Edition.

[5] Goe, J., Tan P.N. and Cheng, H. (2006) Semi-supervised Clustering with Partial Background Information. In Proc. of SIAM International Conference on Data Mining, Bethesda, MD.

[6] Gruber, T.R. (1993) A translation approach to portable ontology specifications, Technical Report, KSL, Knowledge System Laboratory, Pp.92-71.

[7] Hotho A., Staab S. and Stumme G, (2003) WordNet improves text document clustering, Proc. of the SIGIR 2003 Semantic Web Workshop, Pp. 541-544.

[8] Karypis, G., Han, E.H. and Kumar, V. (1999) Chameleon: Hierarchical clustering using dynamic modeling. Computer, Vol. 32, No. 8, Pp. 68–75.

[9] Manning, C. and Schütze, H. (1999) Foundations of Statistical Natural Language Processing, MIT Press, Cambridge, MA.

[10] Sedding J. and Kazakov, D. (2004) WordNet-based text document clustering, Proc. of the 3rd Workshop on Robust Methods in Analysis of Natural Language Processing Data, Pp.104-113.

[11] Wu, Z. and Palmer, M. (1994) Verb Semantics and Lexical Selection, Proc. of the 32nd Annual Meeting of the Assoc. for Computational Linguistics, Pp. 133-138.

[12] Yang, X., Guo, D., Cao, X. and Zhou, J. (2008) Research on Ontology-Based Text Clustering, Proceedings of the 2008 Third International Workshop on Semantic Media Adaptation and Personalization, EEE Computer Society Washington, DC, USA.

[13] Yang, Y. and Pedersen, J.O. (1997) A Comparative Study on Feature Selection in Text Categorization, Proc. of the 14th International Conference on Machine Learning ICML.