STATUS QUO OF SEMANTIC-BASED TEXT DOCUMENT CLUSTERING: A REVIEW

S SUNEETHA[1], M. USHA RANI[2]
[1]Research Scholar, Professor & Head
[1, 2]Department of Computer Science, Sri Padmavati Mahila Visvavidyalayam (Women'sUniversity), Tirupati, Andhra Pradesh, INDIA
[1]suneethanaresh@yahoo.com, [2]musha_rohan@yahoo.com

ABSTRACT

Text Document Clustering has been a focused theme in Data Mining research for over a decade. This task pioneered copious publications in data mining conferences and journals. In spite of abundant literature and tremendous progress made on document clustering, still there exist several challenges for increasing clustering quality. Traditional document clustering algorithms use bag-of-words approach and thus concentrate on the syntax in a document, producing unsatisfactory clustering results. Semantic-based text document clustering concentrates on the computational semantics of the grammar in the document, rather.

With about 3-4 years of substantial and fruitful research in applying semantics to document clustering, the current paper provides an overview of this flourshing field with an eye towards future avenues i.e., what more to be done in order to turn this technology a cornerstone approach in data mining text document applications.

Keywords: *Document Clustering, Bag-of-words Approach, Semantic Document Clustering, Special Requirements, Semantic Similarity Measures.*

1. INTRODUCTION AND MOTIVATION

Technology Melioration provided a tremendous growth in the volume of the text documents available on the internet, digital libraries & repositories, news sources, company-wide intranets, and digitized personal information such as blog articles and emails. As a result, computer understanding of text has acquired great interest in the research community in order to enable a proper exploitation, management, classification or retrieval of textual data. A critical issue is to design and organize the vast amounts of on-line documents on the World Wide Web (WWW) according to their topic and this brought challenges for the effective and efficient organization of text documents automatically. Even for the search engines, when a query is submitted to the system it is very important to group similar documents in order to improve their performance. Text document clustering, an unsupervised machine learning approach organizes large amounts of information into a small number of meaningful clusters, where the documents in each cluster share some common properties according to defined similarity measure. Clustering is useful for taxonomy design and similarity search of documents on such a domain.

'Text Document Clustering' is the organization of an unstructured text corpus into subsets or clusters so as to maximize intra-cluster distances between documents, while minimizing inter-cluster distances between documents. Various applications of document clustering include finding similar documents, document classification, organization and browsing, corpus summarization, duplicate content detection, news aggregation, and search optimization. There are various subcategories of document clustering like soft, hard, partitioning, hierarchical etc. which are again classified further.

The standard document representation technique is the vector space model (VSM).Vector Space Model represents each document as a feature vector of the terms in the document. The existing vector space model is well suited for the search engines and websites based on keywords. Keyword based search engines such as Google, Yahoo, Msn, Ask and Bing are the main tools to use the web. The VSM representation creates problems during retrieval as semantically related words are not taken into account. For example, the sentences "John is an intelligent boy" and "John is a brilliant kid" mean the same thing; "He went to the bank to withdraw money" and "he went to the bank to repair his boat" mean different. Also, traditional document clustering methods that use bags-of-words approach are not suitable for the semantic web searching process due to synonymy, polysemy or homography. Thus, there is a need for semantic driven text document clustering methods to improve the quality of the clusters further.

The rest of this paper is organized as follows: Section 2 highlights requirements for text clustering results improvisation. A brief overview of traditional document algorithms is furnished in Section 3. Section 4 gives an idea about semantic similarity measures. State-of-the-art of semantic driven text document clustering methods is presented in Section 5. The paper is concluded in Section 6 followed by future avenues in Section 7.

Thus, this paper presents a brief review on semantic based document clustering with a focus to enhance the quality and performance of text document clustering further through status in quo of it

2. REQUIREMENTS FOR TEXT DOCUMENT CLUSTERING.

Special Requirements for improving text document clustering result follows:

♣ *To find a suitable model to represent the document.* Accurate meaning of a sentence has close relationship with the sequential occurrences of words in it. So, the document model should preserve the sequential relationship among words in the document for context sensitive representation.

♣ *To reduce high dimensionality of text documents.* In order to efficiently process a huge text database, then text clustering algorithm should have a way to reduce high dimension.

♣ *To allow overlapping between document clusters.* A document can cover several topics. So, overlapping among document clusters should be allowed. For instance, a document discussing “Computational Intelligence and Computer Engineering” should be assigned to both of the clusters “Computational Intelligence” and “Computer Engineering”.

♣ *To associate a meaningful label to each final cluster.* The problem of clustering is to group a collection of unlabeled data into meaningful clusters and the label provides an adequate description of the cluster that will guide users in the process of information retrieval. Hence, final clusters should be meaningfully labelled.

♣ *To estimate the number of clusters.* Clustering, the most uncommon form of unsupervised learning needs little prior information about data and the number of clusters is unknown prior to the clustering. It is difficult to specify a reasonable number of clusters for a data set when little information about it is available and so the text document clustering should provide a rough estimate of the number of clusters.

♣ *To improve scalability.* Many document clustering algorithms work fine on small document sets, but fail to deal with large document sets efficiently and therefore scalability is a big requirement.

♣ *To extract semantics from text.* The bag-of-words representation used for clustering is often unsatisfactory as it ignores the conceptual similarity of terms that do not co-occur actually. So, semantic understanding of text is necessary to improve the efficiency and accuracy of clustering.

3. TEXT DOCUMENT CLUSTERING ALGORITHMS: AN OVERVIEW

The data mining techniques are essentially designed to operate on structured databases. When the data is structured it is easy to define the set of items and hence, it becomes easy to employ the traditional mining techniques. Specific text mining techniques have to be developed to process the unstructured textual data to aid in knowledge discovery. For an unstructured document, features are extracted to convert it to a structured form. Some of the important features are document processing like stop words elimination, stemming, POS tagging. Other higher order features include Semantic grammar, semantic relation between words and similarity measure. Once the features are extracted the text is represented as structured data, and traditional data mining techniques like clustering can be used.

Document clustering methods accept input data in either numerical or categorical form. Documents are represented at the lexical and semantic levels by the words they contain. This creates an independent representation called bag-of-words. To create this representation documents are segmented into tokens based on white space, paragraph separators and punctuation marks. Then all words are extracted and stemmed, stop words are removed and the number of occurrences of each word is counted.

The syntax analysis is used to find the syntactic structure of the sentences. It is the process of analyzing a text made of sequence of tokens(words) to determine its grammatical structure with respect to a given document sentence. Syntax analysis refers to the way that human beings rather than computers analyze a sentence or phrase in terms of grammatical constituents, identifying the parts of speech, syntactic relations. Semantics is the study of meaning and focuses on the relation between words and their literal meaning. The greatest source of difficulty is identifying semantics and traditional document clustering algorithms concentrated on the syntax of the sentence in a document rather than semantics.

Traditional document clustering methods start with partitional and hierarchical methods. In this Unweighted Pair Group Method with Arithmetic Mean (UPGMA) of agglomerative hierarchical clustering is reported to be the most accurate one. Bisecting k-means algorithm, a combination of the strengths of partitioning and hierarchical clustering methods, is reported to outperform the basic k-means as well as the agglomerative approach in terms of accuracy and efficiency.

To resolve the problem of high dimensionality and to produce understandable cluster descriptions with scalability and increased efficiency, several frequent item set based methods were proposed and frequent item sets are good candidates for clusters. Beil et al. developed the first frequent item set based algorithm, namely Hierarchical Frequent Term-based Clustering (HFTC). Only low-dimensional frequent item sets are considered as clusters. HFTC also discovers overlapping clusters, which is useful for a search engine. However, the experiments of Fung et al. showed that HFTC is not scalable.

Fung et al. proposed Frequent Itemset-based Hierarchical Clustering (FIHC) algorithm by using the notion of frequent itemsets derived from association rule mining. Frequent itemsets were also used to construct a hierarchical topic tree for clusters. The experimental results have shown that this method outperforms its competitors in terms of accuracy, efficiency and scalability, ease of browsing with meaningful cluster labels.

Yu et al. presented another frequent itemset-based algorithm, called TDC, to improve the clustering quality and scalability. This algorithm dynamically generates a topic directory from a document set using only closed frequent itemsets and thus reduces the dimensionality further. But, the clusters generated by FIHC and TDC are non-overlapping.

A problem with these algorithms is that they strongly depend on the frequent word sets, which are unordered and cannot represent text documents well in many cases. Although high accuracy is achieved, it affects the overall clustering quality when terms in the document set are highly correlated because of too much node duplication. Moreover, HFTC, FIHC, and TDC only account for term frequency in the documents and ignore the important semantic relationships between terms.

A word sequence is frequent if it occurs in more than certain percentage of the documents in the text database. Frequent word sequences can represent a document well. So, a new text clustering algorithm, named Clustering based on Frequent Word Sequences (CFWS) was proposed. It was shown that CFWS has much better performance.

Then, Suffix Tree Clustering (STC) based on the overlap of their document sets was proposed to get a better clustering result and a comparative small number of clusters. STC just performs word form matching that too with high complexity as it totally ignores the semantic relationships between words and similarity between the non-overlap parts of different clusters.

Recently, WordNet, the most widely used thesauruses for English, has been used to group documents with its semantic relations of terms. However, Synonym sets (synsets) would decrease the clustering performance in all experiments without considering word sense disambiguation.

Documents can be parsed, by using any standard parsers, for generating the syntactic structure also called parts of speech (POS) tagging for the sentences in the document. POS tagging is the process of assigning a parts of speech such as a noun, verb, pronoun, preposition, adverb and adjective to each word in a sentence. Partial disambiguation of words by their POS is beneficial in text clustering. But, taking into account synonyms and hypernyms disambiguated only by POS tags, is not successful in improving clustering effectiveness because of the noise produced by all the incorrect senses. A possible solution that uses a word-by-word disambiguation was proposed by Rekha Baghel in order to choose the correct sense of a word in FCDC (Frequent Concepts based Document Clustering); based on frequent concepts. The proposed FCDC algorithm utilizes semantic relationship between words to create concepts. In turn, it exploits WordNet Ontology to create low dimensional feature vector that allows developing an efficient clustering algorithm. FCDC that uses a hierarchical approach to cluster text documents having common concepts was found to be more accurate, scalable and effective when compared to existing clustering algorithms.

In Clustering based on Frequent Word Meaning Sequences, various document representation methods to exploit noun phrases and semantic relationships for clustering were proposed.

‘Hypernymy’ is defined as the semantic relation of being super-ordinate or belonging to a higher rank or class; ‘Hyponymy’ as the semantic relation of being subordinate or belonging to a lower rank or class; ‘Holonymy’ as the semantic relation that holds between a whole and its parts; and ‘Meronymy’ as the semantic relation that holds between a part and the whole. Using WordNet, hypernymy, hyponymy, holonymy, and meronymy have been utilized for clustering and they found that hypernymy is most effective for clustering, through a series of experiments.

Only these algorithmic and conceptual changes are not enough in current scenario of large data sets. Most of the traditional clustering algorithms suffer from the curse of dimensionality. In sparse and high dimensional space, any distance measure that assumes all features to have equally important is likely to be not effective. This is due to the reason that semantically related words are not taken into account.

4. SEMANTIC SIMILARITY MEASURES

Semantic similarity measurement plays an important role in information retrieval on the semantic Web. They also play an important role for ontology engineering, alignment and matching. The main challenge for semantic similarity measurement is the comparison of meanings as opposed to a purely structural (syntactical) comparison. Depending on the representation language, concepts are specified as collections of features, regions in a multidimensional space, or formal restrictions specified on sets using description logics. Besides representation, context is a major challenge for similarity. In most cases, meaningful measures cannot be defined without specifying a context in which similarity is measured.

‘A semantic relatedness measure’ also called as ‘semantic distance’ or ‘semantic similarity’ is a criterion to find the relatedness of two senses in a semantic network. WSD is finding the correct sense of a word in given context. In word sense disambiguation (WSD) algorithms, a semantic relatedness measure is a very important factor for the performance. Lesk, Fagos et. al., Gomes et. al., Sussna, Li et. al., and Ramakrishnan and

Bhattacharyya have suggested different methods for WSD. Some other methods include Banerjee and Pedersen's method, and Patwardhan's method. WordNet also has similarity measures implemented in WordNet::Similarity, a Perl software package which consists of several sub-modules to implement different semantic relatedness measures.

Jiang and Conrath classified semantic similarity measurement methods into two categories: edge-based methods and node-based methods. Edge-based methods measures the distance between two senses according to the length of the path between them in the semantic networks. The simplest method is to count the number of edges or nodes between them. Some other edge-based (information content-based) methods include Hso method, Lch method, Sussna's method, and Wup method. Node-based methods measure the distance between two senses according to the statistical information contained in the nodes within the semantic network. Some node-based methods include Res method and Lin method. Jcn method is a combined method which considers both edge and node information.

Thus, several measures for semantic-based similarity exist but, a single similarity value is difficult to interpret. There is no global and application independent law on how similarity is measured. Strictly speaking, there is even no single definition of what similarity measures. This makes the selection of an appropriate measure for a particular application area and also the comparison of existing similarity measures difficult.

5. STATUS IN QUO OF SEMANTIC DRIVEN DOCUMENT CLUSTERING

The problem of document clustering has two main components: to represent inherent semantics of the document, and to define a similarity measure based on the semantic representation such that it assigns higher numerical values to document pairs which have higher semantic relationship. In data mining literature, a handful of researches for clustering text data exist and among them the concentration of recent researchers is focused on semantic based text document clustering to further improvise the clustering quality. They differ in document representation, semantic measure, usage of background semantic information etc. A review of few researches related to semantic based text document clustering is furnished below:

In traditional document clustering methods, a document is considered a bag of words. The fact that the words may be semantically related – a crucial information for clustering- is not taken into account. In the paper entitled 'Text Clustering Using Semantics' a new method is proposed for generating feature vectors, using the semantic relations between the words in a sentence. The semantic relations are captured by the Universal Networking Language (UNL), a semantic representation for sentences. The clustering method applied to the feature vectors is the Kohonen Self Organizing Maps (SOM), a neural network based technique, which takes the vectors as inputs and forms a document map in which similar documents are mapped to the same or nearby neurons. This approach performed better than the methods based on only frequency.

The bag of original words cannot represent the content of a document precisely, because of the synonym problem and the polysemous problem. The problem of finding the correct sense of a word in a context is called word sense disambiguation (WSD). In Document Clustering with Semantic Analysis, the sense disambiguation method, the sense of words is investigated to construct the feature vector for document representation. Experimental results demonstrate that using sense can improve the performance of document clustering system, in most conditions. But the comprehensive statistical analysis performed indicates that the differences between using original single words and using senses of words are not statistically significant. The future work focuses on two aspects: Performing syntactic analysis to find the important word in a context and Combining the semantic analysis and syntactic analysis to realize a further improvement.

Latent semantic indexing (LSI) is an indexing and retrieval method that uses a mathematical technique called singular value decomposition (SVD) to identify patterns in the relationships between the terms and concepts contained in an unstructured collection of text. LSI is based on the principle that words that are used in the same contexts tend to have similar meanings. A key feature of LSI is its ability to extract the conceptual content of a body of text by establishing associations between those terms that occur in similar contexts. From Lingo's viewpoint, these concepts are perfect cluster label candidates. The Lingo algorithm combines common phrase discovery and LSI technique to separate search results into meaningful groups. It looks for meaningful phrases to use as cluster labels and then assigns documents to the labels to form groups.

The problem of LSI is that it seeks to uncover the most representative features rather the most discriminative features for document representation. So, LSI might not be optimal in discriminating documents with different semantics. Moreover, the document space is generally of high dimensionality, and clustering in such a high dimensional space is often infeasible due to the curse of dimensionality. Locality Preserving Indexing (LPI) is used to tackle high dimension issue of document clustering. By using Locality Preserving Indexing (LPI), the documents can be projected into a lower dimensional semantic space in which the documents related to the same semantics are close to each other. Different from previous document clustering methods based on Latent

Semantic Indexing (LSI) or Non-negative Matrix Factorization (NMF), LPI tries to discover both the geometric and discriminating structures of the document space.

Text clustering typically involves clustering in a high dimensional space that appears difficult with regard to virtually all practical settings. In addition, given a particular clustering result it is typically very hard to come up with a good explanation of why the text clusters have been constructed the way they are. A new approach for applying background knowledge during preprocessing in order to improve clustering results and allow for selection between results is proposed. The preprocessing method, COSA, that proposed is a very general one. In addition, various views can be built based on the selection of text features on a heterarchy of concepts. Thereby, the user can rely on a heterarchy to control and possibly interpret clustering results onto the same input. The results may be distinguished and explained by the corresponding selection of concepts in the ontology.

Feature space of the documents can be very challenging for document clustering. A document may contain multiple topics, it may contain a large set of class independent general words, and a handful class-specific core words. With these features in mind, traditional agglomerative clustering algorithms, which are based on either Document Vector model (DVM) or Suffix Tree model (STC), are less efficient in producing results with high cluster quality. So, a new approach for document clustering based on the Topic Map representation of the documents is proposed. The document is transformed into a compact form and a similarity measure is proposed based upon the inferred information through topic maps data and structures. This method was implemented using agglomerative hierarchal clustering and tested on standard Information retrieval (IR) datasets. The comparative experiment revealed that the proposed approach was effective in improving the cluster quality.
In order to improve the quality of document clustering results, an effective Fuzzy-based Multi-label Document Clustering (FMDC) approach that integrates fuzzy association rule mining with an existing ontology WordNet was proposed. In this approach, the key terms will be extracted from the document set, and the initial representation of all documents is further enriched by using hypernyms of WordNet in order to exploit the semantic relations between terms. Then, a fuzzy association rule mining algorithm for texts is employed to discover a set of highly-related fuzzy frequent itemsets, which contain key terms to be regarded as the labels of the candidate clusters. Finally, each document is dispatched into more than one target cluster by referring to these candidate clusters, and then the highly similar target clusters are merged. Experiments were conducted to evaluate the performance based on Classic, Re0, R8, and WebKB datasets and the l results proved that the suggested approach outperforms the influential document clustering methods with higher accuracy. This approach not only provides more general and meaningful labels for documents, but also effectively generates overlapping clusters.

Traditionally, clustering techniques do not consider the semantic relationships between words, such as synonymy and hypernymy. To exploit semantic relationships, ontologies such as WordNet were used to improve clustering results. However, WordNet-based clustering methods mostly rely on single-term analysis of text; they do not perform any phrase-based analysis. In addition, these methods utilize synonymy to identify concepts and only explore hypernymy to calculate concept frequencies, without considering other semantic relationships such as hyponymy. To address these issues, detection of noun phrases was combined with the use of WordNet as background knowledge to explore better ways of representing documents semantically for clustering. First, based on noun phrases as well as single-term analysis, different document representation methods were exploited to analyze the effectiveness of hypernymy, hyponymy, holonymy, and meronymy. Second, the most effective method was chosen and it was compared with the existing WordNet-based clustering methods and the experimental results shown that the effectiveness of semantic relationships for clustering is (from highest to lowest): hypernymy, hyponymy, meronymy, and holonymy. Moreover, noun phrase analysis improved the WordNet-based clustering method.

Most of the common techniques in text mining are based on the statistical analysis of a term, either word or phrase. Statistical analysis of a term frequency captures the importance of the term within a document only. However, two terms can have the same frequency in their documents, but one term contributes more to the meaning of its sentences than the other term. Thus, the underlying text mining model should indicate terms that capture the semantics of text. In this case, the mining model can capture terms that present the concepts of the sentence, which leads to discovery of the topic of the document. A new concept-based mining model that analyzes terms on the sentence, document, and corpus levels is introduced. The concept-based mining model can effectively discriminate between non important terms with respect to sentence semantics and terms which hold the concepts that represent the sentence meaning. The proposed mining model consists of sentence-based concept analysis, document-based concept analysis, corpus-based concept-analysis, and concept-based similarity measure. The term which contributes to the sentence semantics is analyzed on the sentence, document, and corpus levels rather than the traditional analysis of the document only. The proposed model found significant matching concepts between documents efficiently, according to the semantics of their sentences. The similarity between documents was calculated based on a new concept-based similarity measure that takes full advantage of using the concept analysis measures on the sentence, document, and corpus levels in calculating the similarity between documents. The experiments on large data sets demonstrated substantial enhancement of the clustering quality.

The clustering of documents presents difficult challenges due to the sparsity and the high dimensionality of text data, and due to the complex semantics of the natural language. Subspace clustering is an extension of traditional clustering that is designed to capture local feature relevance, and to group documents with respect to the features (or words) that matter the most. A subspace clustering technique based on a Locally Adaptive Clustering (LAC) algorithm was proposed to improve the subspace clustering of documents and the identification of keywords achieved by LAC. For this, kernel methods and semantic distances were deployed. The basic idea was to define a local kernel for each cluster by which semantic distances between pairs of words can be computed to derive the clustering and the local term weightings. The proposed approach, called Semantic LAC, was evaluated using benchmark datasets and the results shown that Semantic LAC improved the clustering quality.

Polysemy, phrases and term dependency are the limitations of search technology. A single term cannot identify a latent concept in a document, for instance, the term Network associated with the term Computer, Traffic, or Neural denotes different concepts. To discriminate term associations is a concrete way to distinguish one category from the others. A group of solid term associations can clearly identify a concept. Most methods, such as k-means, HCA, AutoClass or PDDP classify or cluster documents from the represented matrix of a set of documents. It is inefficient and complicated to discover all term associations from such a high-dimensional and sparse matrix. The term - associations (frequently co-occurring terms) of a given collection of documents, form a simplicial complex. The complex can be decomposed into connected components at various levels and each such connected component properly identifies a concept in a collection of documents. So, a novel view based on finding maximal connected components for document clustering was proposed i.e., an agglomerative method for finding geometric maximal connected components without the use of distance function is proposed. Maximal r-simplexes of highest dimensions can represent a maximal primitive concept in a collection of documents. Such maximal simplexes of highest dimension can be effectively discovered and used to cluster the collection of documents. Compared with some traditional methods, such as k-means, AutoClass and Hierarchical Clustering (HAC), and the partition based hypergraph algorithm, PDDP demonstrated its superior performance on three datasets and illustrated that geometric complexes are effective models for automatic document clustering.

The Semantic Model (SM) method only concentrates on the compositional semantics of the grammar. But it is insufficient to get the original semantic meaning of the documents. The document sentences contain idiom phrases does not have compositional semantics, the words collectively do not give the original meaning. Compositional semantics are useful for common sentences or phrases. For example a phrase like “kick the bucket” (meaning is die) does not have compositional semantics as the meaning of the whole is unrelated to the meanings of the component words. Considering compositional semantics, disambiguity and idioms, Idiom Semantic Based Mining Model was proposed in which the documents are clustered based on their meaning using the techniques of idiom processing, semantic weights using Chameleon clustering algorithm. The enhanced quality in creating meaningful clusters has been demonstrated with the use of performance indices, entropy and purity and the results exhibited improved performance.

Metaphor is a pervasive feature of human language that enables us to conceptualize and communicate abstract concepts using more concrete terminology. Unfortunately, it is also a feature that serves to confound a computer’s ability to comprehend natural human language. A method to detect linguistic metaphors by inducing a domain aware semantic signature for a given text was proposed and compared against a large index of known metaphors. By training a suite of binary classifiers using the results of several semantic signature-based rankings of the index, linguistic metaphors in unstructured text were detected at a significantly higher precision as compared to several baseline approaches. Although this technique is necessarily limited by the coverage of the metaphors in the index, it is a viable technique for metaphor detection as more and more examples become available. In future work, existing features supplemented with information such as term imageability, transmission of affect, and selectional preference violation will result in a robust system for linguistic metaphor detection to further aid in the computer understanding of natural language.

The large volume of unstructured text data available at various sources such as digital libraries, news, internet, has given arise a need to organize the information as per the user's requirement. Search for relevant information is efficient when context of the selected word in the document is considered. Document Clustering aims to discover natural groupings, and present an overview of classes (topics) in a document collection. Thus, documents with similar contents are related to the same query. A novel method of word fusion based on word context using hypernym was proposed in which the term frequency of the document collection is computed and contexts based terms are fused. Agglomerative clustering and Bisecting K-Means are used to cluster the extracted features. The transcribed Reuters dataset was used to evaluate the hypothesis and the results obtained are promising with classification accuracy of up to 82.65%.

CONCLUSION

Text Document clustering is a fundamental unsupervised operation used in document organization, automatic topic extraction, and information retrieval. K-means, hierarchical agglomerative clustering and various variations of these algorithms were proposed for document clustering over years. Concepts of frequent item set, fuzzy theory, neural network, genetic algorithm, self-organizing map, non-negative matrix factorization etc. were also studied for increasing efficiency of document clustering. Then, to increase the quality and efficiency semantics are also applied to document clustering; this includes latent semantic indexing, frequent word phrases, WordNet, ontology, part-of-speech, tagging, sense disambiguation, machine learning, and many more. Only these algorithmic and conceptual changes are not enough in current scenario of large data sets. Still there is research going to find out more semantically oriented approaches to further increase the accuracy and quality of clusters.

The Semantic Model (SM) method concentrates on the compositional semantics of the grammar. Compositional semantics signifies a system of constructing logical forms for sentences or parts of sentences in such a way that the meanings of the components of the sentences (phrase) are used to construct the meanings of the whole sentence. This paper elucidates special requirements for improving text clustering results, semantic similarity measures and a brief overview of research done in the area of semantic driven text document clustering. Though abundant literature on semantic based text document clustering abide, there are still some challenging research issues to further increase the accuracy and quality of clusters.

FUTURE AVENUES

To increase the quality and efficiency of document clustering, semantics are applied to document clustering; this includes latent semantic indexing, frequent word phrases, WordNet, ontology, part-of-speech tagging, sense disambiguation, machine learning, and many more. Still there is research going on to find out more semantically oriented approaches to further increase the accuracy and quality of clusters.

Future Avenues in enhancing document clustering using semantic analysis include reducing the computational complexity of semantic mapping, dimension reduction and making use of semantics like ontology and natural language processing.

In Semantic Based Model for Text Document Clustering with Idioms, the documents are clustered based on their meaning using the techniques of idiom processing, semantic weights using Chameleon clustering algorithm. Further work can be concentrated on data documents consisting of metaphors and ellipses. Adopting a multilevel or hybrid clustering may improve cluster quality and justification of time complexity need to be made. A robust system for linguistic metaphor detection to further aid in the computer understanding of natural language may be developed in near future.

With more and more development of information technology, data set in many domains is reaching beyond petascale; making it difficult to work with the document clustering algorithms in central site and leading to the need of increasing the computational requirements. Thus, the concept of distributed computing is explored for document clustering giving rise to the concept of distributed document clustering, providing a great potential for future research.

This survey paper may help in the thriving research area of text document clustering.

REFERENCES

[1] Alexander Budanitsky (1999), “Lexical Semantic Relatedness and its Applications in Natural Language Processing”, Technical Report, CSRG-390.

[2] Andreas Hotho, Alexander Maedche, Steffen Staab (2002), “Ontology-based Text Document Clustering,” *K¨unstliche Intelligenz,* **16**, *4*, pp. 48–54.

[3] B. Choudhary, P. Bhattacharyya (2002), “Text clustering using semantics,” *Proc. 11th International World Wide Web Conference.*

[4] B. Drakshayani and E. V. Prasad, (2013), “Semantic Based Model for Text Document Clustering with Idioms”, Intl. *J. Date Engg.*, **4**, *1*, pp. 1□13.

[5] Chun-Ling Chen, Frank S.C. Tseng, Tyne Liang (2010), “An Integration of WordNet and fuzzy association rule mining for multi-label document clustering”, *J. of Data and Knowledge Engineering,* **69**, *11*, pp. 1208-1226.

[6] Deng Cai, Xiaofei He, and Jiawei Han (2005), "Document Clustering Using Locality Preserving Indexing", *IEEE Transactions on Knowledge and Data Engineering,* **17**, *12*.

[7] Kevin Lind (2006), "Concept Based Document Clustering using a Simplicial Complex, a Hypergraph", Master's Thesis.

[8] Loulwah AlSumait, Carlotta Domeniconi (2007), "Local Semantic Kernels for Text Document Clustering," *Intl. Conf. on Data Mining.*

[9] Michael Mohler, David Bracewell, David Hinote and Marc Tolinson (2013), "Semantic Signatures for Example-Based Linguistic Metaphor Detection", *Proc. First Workshop on Metaphor in NLP,. Control*, Atlanta, Georgia, pp. 27□35.

[10] Muhammad Rafi, M. Shahid Shaikh, Amir Farooq (2010), "Document Clustering based on Topic Maps", *Intl J. of Comp. Applications,* **12.**

[11] Neepa Shah and Sunita Mahajan (2012), "Semantic based Document Clustering: A Detailed Review", *Intl. J. Comp. Appls*, **5**, *52*, pp. 42□51.

[12] Neepa Shah and Sunita Mahajan (2013), "Semantic based Distributed Document Clustering: Proposal", *Intl. J. Comp. Sci. & Engg.*, **3**, *2*, pp. 379□388.

[13] Shehata, S. Karray, F. Kamel, M.S. (2010), "An Efficient Concept-Based Mining Model for Enhancing Text Clustering", *IEEE Transactions on Knowledge and Data Engineering*, **22**, *10*, pp. 1360 – 1371.

[14] Stanislaw Osinski, Dawid Weiss (2005), "A Concept-Driven Algorithm for Clustering Search Results", *J. of IEEE Intelligent Systems,* **20**, *3*, pp. 48-54.

[15] Tao Zheng, Bo-Yeong Kang, Hong-Gee Kim (2009), "Exploiting noun phrases and semantic relationships for text document clustering", *J. of Information Sciences*, **179**, *13*, pp. 2249-2262 .

[16] Venkatesh Kumar P, A Subramani (2013), "Using Data Fusion for a Context Aware Document Clustering", Intl. J. of Comp. Applications, **72,** *6,* pp. 17-20.,

[17] Wei Song, Cheng Hua Li, and Soon Cheol Park (2009), Genetic Algorithm for Text Clustering using Ontology and Evaluating the Validity of Various Semantic Similarity, *J. of Expert systems with Applications*, **5**, *36*, pp. 9095□9104.

[18] Yong Wang and Julia Hodges (2006), Document Clustering with Semantic Analysis, *Proc. Intl. Conf. Sys. Sciences*, Hawaii, pp. 54.3.

AUTHORS BIOGRAPHIES

S.SUNEETHA received Bachelor's Degree in Science and in Education, Master's Degree in Computer Applications (MCA) from SVU, Tirupati and M.Phil. in Computer Science from SPMVV, Tirupati. Currently, she is pursuing her Ph.D. under the able guidance of Dr. M. Usha Rani, Preofessor & Head, Department of Computer Sceience, SPMVV, Tirupati. She is a life time member of ISTE. Her areas of interest are Data Mining & Software Engineering. She has 9 papers in National/International Conferences/ Journal to her credit. She has 9 papers in National/International Conferences/ Journal to her credit. She also attended several workshops in different fields. She has 9 papers in National/International Conferences/ Journal to her credit. She also attended several workshops in different fields. She served Narayana Enginering College, Nellore, Andhra Pradesh as Sr. Asst. Professor heading the departments of IT and MCA.

Dr. M.Usha Rani is Professor & Head in the Department of Computer Science in Sri Padmavati Mahila Visvavidyalayam (Women's Univeristy), Tirupati. She did Ph.D. in Computer Science in the area of Artificial Intelligence & Expert Systems. She is in teaching since 1992. She presented many papers at National & Inter-National Conferences and published articles in National and International Journals. She has also written 4 books like Superficial Overview of Data Mining Tools, Data Mining Applications, Opportunities and Challenges. She is guiding M. Phil. & Ph.D. Students in the areas of Artificial Intelligence, Data Warehousing & Data Mining, Computer Networks and Network Security etc.,