Semantic based Document Clustering: A Detailed Review

Neepa Shah
Assistant Professor, IT Department
D.J. Sanghvi College of Engineering,
Vile Parle(West), Mumbai-56

Sunita Mahajan
Phd, Principal, Institute of Computer Science,
M.E.T., Bandra (west),
Mumbai – 50.

ABSTRACT
Document clustering, one of the traditional data mining techniques, is an unsupervised learning paradigm where clustering methods try to identify inherent groupings of the text documents, so that a set of clusters is produced in which clusters exhibit high intra-cluster similarity and low inter-cluster similarity. The importance of document clustering emerges from the massive volumes of textual documents created. Although numerous document clustering methods have been extensively studied in these years, there still exist several challenges for increasing the clustering quality. Particularly, most of the current document clustering algorithms does not consider the semantic relationships which produce unsatisfactory clustering results. Since last three-four years efforts have been seen in applying semantics to document clustering. Here, an exhaustive and detailed review of more than thirty semantic driven document clustering methods is presented. After an introduction to the document clustering and its basic requirements for improvement, traditional algorithms are overviewed. Also, semantic similarity measures are explained. The article then discusses algorithms that make semantic interpretation of documents for clustering. The semantic approach applied, datasets used, evaluation parameters applied, limitations and future work of all these approaches is presented in tabular format for easy and quick interpretation.

General Terms
Data Mining, Semantic Document Clustering

Keywords
Document clustering, semantic based document clustering, requirements of document clustering, semantic similarity for document clustering

1. INTRODUCTION
With the enormous success of the Information Society and the World Wide Web, the amount of textual electronic information available has significantly increased. 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 [1]. Text document clustering plays an important role in providing intuitive navigation and browsing mechanisms by organizing such large amounts of information into a small number of meaningful clusters [2]. In traditional document clustering methods, Vector Space Model (VSM) is used which uses linear-algebra operations to compare textual data (bag-of-word approach). VSM associates a single multidimensional vector with each document in the collection, and each component of this vector reflects a particular keyword or term related to the document. This represents a set of documents by arranging their vectors in a term-document matrix. The value of a single component of the term-document matrix depends on the strength of the relationship between its associated term and the respective document. Importance of different words is then calculated using different criteria like Inverse Document frequency and Information Gain [3]. Inherent shortages of VSM include breaking multi-word expressions, like Machine Learning, into independent features, mapping synonymous words into different components, and treating polysemous as one single component. Moreover, the VSM representation of text data can easily result in tens or hundreds or thousands of features. As a consequence, any clustering algorithm would suffer from the curse of dimensionality. In such sparse and high dimensional space, any distance measure that assumes all features to have equally important is likely to be not effective [4]. This is due to the semantically related words are not taken into account; which can cause problems. For example, if we consider the two sentences "John eats the apple standing beside the tree" and "The apple tree stands beside John's house". These are two different sentences formed from same words. On the other hand there may be some sentences, which have the same meaning but have been constructed from different sets of words. For example in the sentences, "John is an intelligent boy" and "John is a brilliant lad"; mean more or less the same thing [5]. There are some methods like Latent Semantic Indexing, which try to solve this problem. The word category map method can also be used for the same purpose. Shortcoming of these methods is due to polysemy or homography, where a word has different meanings or meaning shades in different contexts (for example, the word bank in "He went to the bank to withdraw some money" and "The boat was beside the bank").

Here, we present detailed survey of various semantic driven techniques for document clustering which enhances the quality and performance of the clusters formed. Section 2 highlights special requirements for improving the clustering results. The brief overview of traditional clustering algorithms is given in section 3. Section 4 gives idea and importance of semantic similarity measures. Detailed overview of more than thirty semantic driven document clustering methods along with tabular representation of the semantic approach applied, datasets used, evaluation parameters applied, limitations and future work of all these approaches is presented for easy and quick interpretation in section 5. The paper is concluded in section 6.

2. SPECIAL REQUIREMENTS FOR IMPROVING THE CLUSTERING RESULT
Here, we highlight special requirements of document clustering to improve the quality of clustering result.

- Finding 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 between words in the document for context-sensitive representation [6].

- To reduce the high dimension of text documents: Usually there are about 200–1000 unique words in a document. In order to efficiently process a huge text database the text clustering algorithm should have a way to reduce the high dimension [6].
- To allow overlapping between document clusters: A document can cover several topics [6]. For instance, a document discussing "Natural language and Information Retrieval" should be assigned to both of the clusters "Natural language" and "Information Retrieval".
- Associating a meaningful label to each final cluster: As the label can provide an adequate description of the cluster [6] and will guide users in the process of browsing the retrieved results [7].
- To estimate the number of clusters: As the number of clusters is unknown prior to the clustering. Also it is difficult to specify a reasonable number of clusters for a data set when little information about it is available [6].
- To improve the scalability: Many document clustering algorithms work fine on small document sets, but fail to deal with large document sets efficiently [7]. So scalability is also a big requirement.
- To extract semantics from text: the bag-of-words representation used for clustering algorithms is often unsatisfactory as it ignores the conceptual similarity of terms that do not co-occur actually [7]. So semantic understanding of text is necessary to improve the efficiency and accuracy of clustering.

3. OVERVIEW OF CLUSTERING ALGORITHMS

Traditional clustering algorithms have not focused on semantic extraction from text but have tried to address other requirements mentioned section 2.

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, combining 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 [6].

To resolve the problems of high dimensionality, large size, and understandable cluster description, number of frequent itemsets-based methods have been seen. Beil et al. developed the first frequent itemsets-based algorithm, namely Hierarchical Frequent Term-based Clustering (HFTC). Only low-dimensional frequent itemsets 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. For a scalable algorithm, Fung et al. proposed the Frequent Itemset-based Hierarchical Clustering (FIHC) algorithm by using frequent itemsets derived from association rule mining to construct a hierarchical topic tree for clusters. 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 further reduces the dimensionality. But, the clusters generated by FIHC and TDC are non-overlapping [7].

An advantage of these frequent-itemsets based algorithms is that a label is provided for each cluster. The label is the frequent word sets shared by the documents in each cluster. A problem of these algorithms is that they strongly depend on the frequent word sets, which are unordered and cannot represent text documents well in many cases [6]. Also, though high accuracy is achieved, it affects the overall clustering quality because of too much node duplication when terms in the document set are highly correlated. Moreover, HFTC, FIHC, and TDC only account for term frequency in the documents and all ignore the important semantic relationships between terms [8].

Frequent word sequences can represent the document well. So, clustering text documents based on frequent word sequences is meaningful. Ahonen-Myka et al. also pointed out that the sequential aspect of word occurrences in documents should not be ignored to improve the information retrieval performance. The idea of using word sequences (phrases) for text clustering was proposed in [9]; and then the Suffix Tree Clustering (STC) based on this idea was proposed in [10]. However, STC does not reduce the high dimension of the text documents; hence its complexity is quite high for large text databases. And STC just performs the word form matching, which ignores the semantic and lexical relationships between words [6].

Recently, WordNet, which is one of 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 [8].

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 extracted from WordNet. A possible solution is proposed which uses a word-by-word disambiguation in order to choose the correct sense of a word in [11]. In [6] Clustering based on Frequent Word Sequences (CFWS) has been proposed. In [12] the authors have proposed various document representation methods to exploit noun phrases and semantic relationships for clustering. Using WordNet, hypernymy, hyponymy, holonymy, and meronymy have been utilized for clustering. Through a series of experiments, they found that hypernymy is most effective for clustering.

WordNet [13] has defined hypernymy 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.

All these methods, with WordNet, hypernymy, word sense disambiguation, and other approaches are explained evolutionary in section 5.

4. SEMANTIC SIMILARITY MEASURES

A semantic relatedness measure is a criterion to find the relatedness of two senses in a semantic network. It is also called semantic distance or semantic similarity. In word sense disambiguation (WSD) algorithms, a semantic relatedness measure is a very important factor for the performance. WSD is finding the correct sense of a word in given context. Lesk,

Fagos et.al., Gomes et.al., Sussna, Li et.al., and Ramakrishnanan and Bhattacharyya have suggested different methods for WSD. Budanitsky presented a comprehensive overview of various semantic relatedness measures. Jiang and Conrath classify those methods into two categories, edge-based methods and node-based (information content-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 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. 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 [14].

The semantic similarity in WordNet: WordNet is an online semantic dictionary which is developed at Princeton by a group led by Miller. Synonym sets are named as synsets. WordNet organizes the lexicon by nouns, verbs, adjectives, and adverbs; represented by synsets. The synset reflects a concept in which all words have similar meaning. The functions of synset include the concept definition for each word and the semantic relationship pointed to other related synsets. WordNet provides a number of 18 kinds of relations to represent nouns concepts. The "ISA" hierarchical structure of the knowledge base is important in determining the semantic distance between words. Fig. 1 shows a part of such a hierarchical semantic knowledge base [15].
One direct approach for calculating semantic similarity between two concepts is to find the minimum length of the path connecting these two concepts. For example, the shortest path between "teacher" and "student" is "teacher–educator–professional–adult–person–intellectual–student". The minimal length of the path is 6. While the minimal path length between "teacher" and "parent" is 9. Thus we could conclude that "student" is more similar to "teacher" than "parent" to "teacher". If a word has multiple meaning, various paths may exist [15].

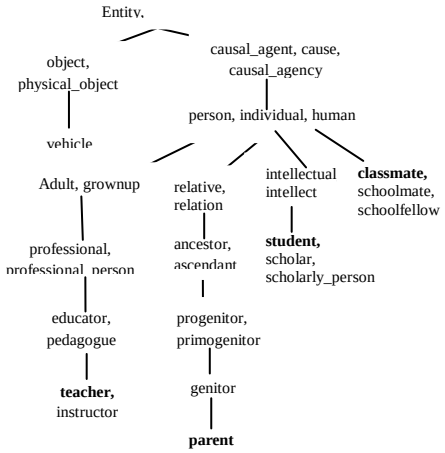

Fig. 1: Part of hierarchical semantic knowledge base

5. REVIEW OF SEMANTIC DRIVEN DOCUMENT CLUSTERING METHODS

The problem of document clustering has two main components: (1) to represent inherent semantics of the document, and (2) 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 [16]. Various approaches have been proposed by many researchers to take care of semantic relation in document clustering. They differ in document representation, semantic measure, usage of background semantic information etc. The brief description of these methods is given below.

In [5] a new method for generating feature vectors is given. The semantic relations between the words in a sentence is described t generate the feature vectors. The semantic relations are captured by the Universal Networking Language (UNL), the semantic representation for sentences. The UNL presents the document in the form of a semantic graph with universal words as nodes and the semantic relation between them as links. The clustering method applied to the feature vectors is the Kohonen Self Organizing Maps (SOM). Experiments show that UNL method for feature vector generation tends to perform better than the term frequency based method.

In [17], a new approach to improve the clustering result is suggested by applying background knowledge during preprocessing. During preprocessing an ontology-based heuristics for feature selection and feature aggregation is applied to construct a number of alternative text representations. This approach is referred as COSA (Concept Selection and Aggregation). It consists of two stages. In first stage, COSA maps terms onto concepts using a shallow and efficient natural language processing system. Then, COSA uses the concept heterarchy to propose good aggregations for subsequent clustering. The results are found to be comparable with a sophisticated baseline preprocessing strategy on tourism domain dataset.
Wordnet is incorporated as background knowledge into a representation for text document clustering in [2]. Here, word sense disambiguation and feature weighting is used to achieve improvements in clustering. For Clustering standard partitional Bisecting K-means algorithm is used and improvement is observed in clustering with background knowledge compared to clustering without background knowledge.

A distributional clustering algorithm, Contextual Document Clustering (CDC), for document clustering is proposed in [18]. Basic principle of CDC is to split document corpus into relatively large groups of documents that are covered by relatively small number of concepts. For this, subject related words, which have a narrow context, are identified to form meta-tags for that subject. These contextual words form the basis for creating thematic clusters of documents. The method outperforms the K-means method and information theoretic method of sequential information bottleneck.

The hybrid algorithm called gradient descent with constrained least squares (GD-CLS) (a nonnegative matrix factorization (NMF) method) for text mining is proposed in [19]. This method introduces a partitional clustering that identifies semantic features in a document collection and groups the documents into clusters on the basis of shared semantic features. Textual data is encoded using a low rank NMF

algorithm to retain natural data non-negativity. NMF outperforms traditional vector space approaches to information retrieval (such as latent semantic indexing) for document clustering on few topic detection benchmark collections.

A guided neural network based on topical information and SOM to exploit the domain knowledge by integrating topical and semantic information from WordNet is proposed in [20]. SOM combines nonlinear projection, vector quantization, and data-clustering functions. The approach is tested in three static competitive learning models: competitive learning, SOM, neural gas, and in three dynamic competitive learning models: growing grid, growing cell structure, and growing neural gas models. The outcome is compared for these six models with WordNet, without WordNet, with one-level, two-level and three-level hypernyms to prove that the model can potentially handle real world tasks.

Latent Semantic Indexing (LSI) aims to represent the input collection using concepts found in the documents. To do this, LSI approximates the original term-document matrix using a limited number of orthogonal factors. These factors represent a set of abstract concepts, each conveying some idea common to a subset of the input collection. From Lingo's [3] 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. Lingo is available in the online demo of the Carrot system at http://carrot.cs.put.poznan.pl.

CDC presented in [18] is enhanced for the scalability and quality for large data-sets using the whole Reuters Corpus Volume 1 (RCV1) collection in [22]. The contexts identified by distributional approach of CDC acts as attractors for clustering documents that are semantically related to each other. Once clustered, the documents are organized into a minimum spanning tree so that the topical similarity of adjacent documents within this structure can be assessed. It is demonstrated that CDC is a powerful and scalable technique with the ability to create stable clusters of high quality. Also, the time complexity is found to be less than the time complexity of partitional clustering algorithm such as K-means. This is the first time that a collection as large as RCV1 has been analyzed in its entirety using a static clustering approach. The RCV1 collection is approximately 35 times more documents than the popular Reuters-21278 collection and contains approximately 10 times the number of distinct words after stemming. The experiments are conducted for quality, homogeneity, and stability of clustering.

In [23], Locality Preserving Indexing (LPI) is used to tackle high dimension issue of document clustering. In dimension reduction, local geometric structure is more important than global structure. In this paper, LPI is proved to be more useful compared to LSI for dimension reduction purpose. 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. Whereas, LPI aims to discover the local geometrical structure. LPI can have more discriminating power. Thus, the documents related to the same semantics are close to each other in the low-dimensional representation space. Also, LPI is good linear approximation to spectral clustering. Thus, linearity of LPI makes it more applicable compared to spectral clustering when dataset is large.

A new algorithm for clustering search results for search engines is proposed in [24]. Many other clustering systems analyze snippets, a short document abstract returned by search engines, post-processing step. As snippet might not always represent whole document content, the quality of cluster can be worse. So, here comparison between snippet analysis and whole document content analysis is produced. A dynamic Singular Value Decomposition (SVD) clustering approach based on LSI is presented to discover the optimal number of singular values to be used for clustering purposes. Also, the algorithm is incremental; thus, eliminating the computation of whole SVD matrix. As whole document content is needed for analysis, this algorithm has been integrated into the Noodles search engine, a tool for searching and clustering Web and desktop documents.

The thesis [25] evaluates the effectiveness of using a combinatorial topology structure (a simplicial complex) for document clustering. It is a geometric structure formed by terms and the associations between them, as only terms fail to identify the concept. A simplicial complex identifies the latent concept space defined by a collection of documents better than the use of hypergraphs or human categorization. Hypergraphs are used to represent term associations (co-occurring terms). The method is compared with human classifiers and proved to work better.

As bag-of-words approach fail to capture semantics in document, sense disambiguation method is used to construct feature vector for document representation in [14]. In this system, words are first mapped to word senses using a semantic relatedness based word sense disambiguation algorithm. Then these senses are used to construct the feature vector to represent the documents. Two different sense representation methods, namely, senseno and offset, are used. Different semantic relatedness measures are also evaluated in the experiments. As large-scale thesaurus and dictionary WordNet is used. K-means, Buckshot, HAC, and bisecting k-means clustering algorithms are used for comparison. For small dataset, HAC outperforms other algorithms and for large datasets bisecting k-means has shown better performance.

As seen in section 2, the high dimensionality of text data, and the complex semantics of the natural language are major requirements of document clustering. To deal with these issues, a subspace clustering technique based on a Semantic Locally Adaptive Clustering (LAC) algorithm is presented in [4]. 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.
In LAC, a weighted cluster is defined as a subset of data points with a weight vector. In this cluster, the points are clustered as per their weighted Euclidean distance. Thus, the objective of LAC is to find cluster centroids, and weight vectors. These local weights are used to find keywords for each cluster. In semantic LAC a semantic distance between pairs of words is used by defining a local kernel for each cluster. Experiments show that Semantic LAC is capable of improving the clustering quality and enhance the subspace clustering of documents.

Since single word-based representation of feature vector doesn't convey semantics, so work is done in the direction of phrases and sequence of words for feature vector generation. In [6], two text clustering algorithms, Clustering based on Frequent Word Sequences (CFWS) and Clustering based on Frequent Word Meaning Sequences (CFWMS) are proposed. If documents share frequent word (meaning) sequences then they are more close to each other. To find this semantic measure, each document is first converted to a compact form by keeping only the frequent words (meanings). Then Generalized Suffix Tree for all such compact documents is built. The frequent word (meaning) sequences and the documents sharing them are found. Then, frequent word (meaning) sequences are used to create clusters and describe their content for the user.
In CFWMS, words are converted into word meanings considering the synonymy, polysemy, and hyponymy/hypernymy relationships between words making use of WordNet. CFWMS is found to be more accurate than CFWS, and both of them are more accurate than bisecting k-means, and FIHC algorithms.

Genetic Algorithm (GA) belongs to search techniques that can efficiently evolve the optimal solution in the reduced space. But GA cannot be applied to high dimensional VSM representational model of documents due to scalability issue and high computational cost. So, in [26] GA method based on a latent semantic model for text clustering is presented. LSI uses SVD technique to decompose the large term-by-document matrix into a set of k orthogonal factors. So, the approach first applies SVD technique to raw data. This outcome is then given to a variable string length GA; which finds proper number of clusters automatically as well as provides near optimal dataset clustering. Experiments are shown to prove improvement of GA in conjunction with the reduced latent semantic structure compared to GA in conjunction with VSM in terms of dimensionality reduction, computational cost, and clustering efficiency and accuracy.

[15] Proposes a self-organized genetic algorithm for text clustering based on ontology (thesaurus-based and corpus-based ontology). A transformed LSI model which can appropriately capture the associated semantic similarity is proposed and demonstrated as corpus-based ontology. Two hybrid strategies are put forward as the semantic similarity measures (Ontology based semantic similarity measure, LSI for semantic similarity calculation) to overcome the limitation of each sole similarity measure. Experiments show that this method of GA in conjunction with the ontology strategy, the combination of the transformed LSI-based measure with the thesaurus-based measure, apparently outperforms that with traditional similarity measures.

PubMed is the most comprehensive database of biomedical literature. Many document clustering methods have been suggested to better understand this literature; but they lack in semantic meaning. An ontological clustering method called GOClonto for conceptualizing these abstracts is proposed in [27]. There is Gene Ontology (GO) to provide controlled vocabulary for describing gene and gene product attributes. GOClonto makes use of this ontology with latent semantic analysis (LSA) to identify key gene-related concepts and their relationships. Then the abstracts are allocated based on these key gene-related concepts, helping users to browse and conceptualize this collection. It also generates corpus-related ontology automatically. Experiments are performed to show that this generated ontology is more informative compared to other algorithms developed for this domain.

To consider semantics of the document many methods use WordNet as thesaurus. But WordNet-based clustering methods rely only on single-term analysis of text; they do not perform phrase-based analysis. Also, these methods use synonymy to identify concepts and explore only hypernymy to calculate concept frequencies; i.e. they don't consider other semantic relationships. To address these issues, in [12] authors have combined detection of noun phrases and WordNet. This integration helps in exploring documents more semantically for clustering purpose. Also other semantic relationships such as hypernymy, hyponymy, holonymy, and meronymy are exploited. The experimental results show the hypernymy is most effective and useful for clustering. Other relationships are useful; their effectiveness for clustering is in order from highest relevant to lowest relevant: hyponymy, meronymy, and holonymy. It is proved through experiments that noun phrase analysis improves the WordNet-based clustering method.

An effective Fuzzy-based Multi-label Document Clustering (FMDC) approach is proposed in [7] to improve the quality of clustering results. In this approach, fuzzy association rule mining is integrated with WordNet ontology. The key terms are extracted from the document set, and the generated feature vector is then enriched using the hypernyms of WordNet to exploit the semantic relations between terms. Now, to discover fuzzy frequent itemsets a fuzzy association rule mining algorithm for texts is applied. These frequent itemsets become 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. Thus, this approach provides meaningful labels to clusters and also generates overlapping clusters effectively. Experiments proved quality enhancement over FIHC, K-means, UPGMA, and bisecting K-means.

In [28], a new concept-based mining model that analyzes terms on the sentence, document, and corpus levels is introduced. So, semantic structure of each term is captured within a sentence, document, and corpus instead of checking only in document. The model consists of sentence-based concept analysis, document-based concept analysis, corpus-based concept-analysis for this requirement. Also, concept-based similarity measure is proposed. The similarity between documents is calculated based on this new concept-based similarity measure. Comparison is provided for single-term against concept-based for term frequency (TF), conceptual; term frequency (CTF), document frequency (DF), and combined frequency; and improvement is observed. Thus, this model brings NLP to the document clustering process.

A new approach based on the Topic Map representation of the documents is introduced in [16]. This approach is more applicable to multi-topic documents. Topic maps help in finding relations among knowledge content. Similarity measure for this new representation is also proposed. Here, the document is first converted to a compact form. The topic map information is generated and information is then extracted from these data structure. A similarity measure is applied to this inferred information through topic maps data and structures. AHC algorithm is implemented and tested on standard information retrieval datasets. The comparative

experiment reveals that the proposed approach is effective in terms of quality compared to CFWS, FIHC, and BKM.

In [29], four different clustering approaches for text collection using probabilistic models are proposed. Dimension reduction is performed using latent variables which compose a concept space and then clustering is performed in that space. First a two-stage clustering method applying concept space is developed; where Classification Expectation-Maximization (CEM) algorithm is used to find the word concepts and Multinomial Mixture (MM) model is then applied in this reduced concept space for document clustering. Then, three clustering approaches based on PLSA are developed. Ext-PLSA model supplements the previous approach by combining two stages in a process. CS-PLSA algorithm allows an effective model selection for clustering. Finally, voted-PLSA provides a successful multi-view clustering procedure on a multilingual collection.

In section 2, frequent-itemset based methods are explored. In [11], a new technique based on frequent concepts for document clustering is proposed. Frequent Concepts based Document Clustering (FCDC) algorithm utilizes the semantic relationship between words, explored using WordNet ontology, to create concepts. This process creates low dimensional feature vector giving an efficient clustering algorithm. From this feature vector frequent concepts are found and then hierarchical approach is used to cluster text documents having common concepts. FCDC is found to be more accurate, scalable and effective compared to BKM, UPGMA and FIHC.

Two semantic based document clustering algorithms are studied and compared in [30]. The first method, Hybrid Scheme for Text Clustering (HSTC), uses a hybrid approach to combine pattern recognition algorithms with semantic driven processes. Here, content-based distance measure is used as similarity measure. The second algorithm, Text Clustering with Feature Selection (TCFS), uses ontology based feature selection for clustering. This method also calculates term weight and semantic weight during preprocessing stage. Both these techniques are efficient in clustering process, but the quality of clustering is slightly better for TCFS. The problem of TCFS is that it is slow. In future, both these methods are proposed to combine to take advantage of each one of these.

According to the result provided by Hai-Tao Zheng, et.al, 2009, Hypernymy gives better semantical relationship than Hyponymy, Meronymy and Holonymy. So, Hypernymy is used to broaden the search. [31] utilizes hypernymy to identify semantic relation by using the WordNet. It acts as background knowledge of the Query and provides its synonymic terms. The new term-document matrix called Query based document vector model, which is constructed using query with two terms and its hypernymy is proposed.

Taxonomy is simplified version of ontology. Taxonomy describes hyponymy between concepts. So taxonomy is very useful in many NLP applications. A methodology for learning taxonomy from set of text documents each with one concept is presented in [32]. The taxonomy is obtained by clustering the concept definition documents with a hierarchical approach to SOM. Three different feature extraction methods: a combination of rule-based stemming and fuzzy logic-based feature weighting and selection, statistical stemming together with statistical key-phrase extraction, and rule-based stemming with the traditional tf-idf term weighting are compared.

Table 1 highlights all these methodologies with different parameters like the semantic approach applied, datasets used, evaluation parameters applied, limitations and future work for easy and quick reference.

Table 1. Comparison of various semantic driven document clustering methods

Paper referred	Semantic Approach	Clustering Algorithm	Dataset	Evaluation parameters	Limitation	Future work
B. Choudhary P. Bhattacharya [5]	Universal Networking Language (UNL): Semantic graph	Kohonen Self Organizing Maps	26 documents	Term Frequency UNL Link UNL Relation	Dataset too small	More semantic based approach and scalability
Andreas Hotho, Alexander Maedche, and Steffen Staab [17]	COSA (Concept Selection and Aggregation): ontology-based heuristics	K-means	document set from the tourism domain (2234 HTML documents)	silhouette coefficient	Dataset is not standard text documents, but real-world telecomm customer database (24,156)	Compare the results with another low dimensional baseline taken from latent semantic indexing.
Andreas Hotho, Steffen Staab, Gerd Stumme [2]	Wordnet (word sense disambiguation and feature weighting)	Bisecting K-Means	Reuters-21578 news corpus	Purity and inverse purity	All 21578 documents are not used. Restricted to 12344 documents	More advanced word sense disambiguation and feature weighting schemes to improve effectiveness
Vladimir Dobrynin, David Patterson, Niall Rooney	Distributional clustering, Narrow context	assign documents to the cluster with the closest centroid (JS-	ModApte split of the Reuters-21578 and 20-NG data	Precision and recall	Comparison with standard clustering algorithm is not done	Use in IR systems, Parallelism to improve performance, personalization technique, interactively browsing the

[18]	identificatio n	divergence)				query results
Farial Shahnaz, Michael W. Berry, V. Paul Pauca, Robert J. Plemmons [19]	hybrid technique for Non-negative matrix factorizatio n	GD-CLS (gradient descent with constrained least squares)	the Reuters Document Corpus and TDT2 (7919)	the regularizatio n parameter to control the sparsity, accuracy measure	Comparison with standard clustering algorithm is not done, documents associated with only one topic are used and topics with cluster sizes smaller than five are discarded	Efficiently updating features and clusters as documents are added to a text collection, Use in IR systems, Applying NMF to bioinformatics
Chihli Hung, Stefan Wermter, Peter Smith [20]	WordNet	Guided SOM	Reuters corpus, RCV1 (806,791 news article)	classification accuracy (CA) and average quantization error (AQE)	considers each word as a symbol i.e. uses VSM model; thus the sequences of words in sentences are ignored	Integrating some natural language processing techniques, such as tagging, parsing, and word sense disambiguation
Stanislaw Osi´nski and Dawid Weiss [3]	common phrase discovery and latent semantic indexing	SVD	test data from the Open Directory Project (a human-collected directory of Web pages)	cluster contaminatio n measure	VSM is used, time complexity is high, more memory requirements	improving the algorithm's efficiency and investigating possibilities for distributing the processing, creating a hierarchical structure of clusters, improving the document-to-cluster assignment phase, improving the phrase-selection method to prune elliptical or ambiguous phrases
Elena Montañés, Irene Díaz, José Ranilla, Elías F. Combarro, and Javier Fernández [21]	ML-based scoring measures for conducting Feature selection	Support Vector Machine	Reuters-21578 (Apte split11: 9,805 documents) and the Ohsumed corpora	information gain, Expected cross entropy for text (CET), ML measure, precision and recall	TF (term frequency) approach is used, proposed measures are more dependent on some statistical properties of the corpora	To apply other Clustering algorithm. Apply measures to other corpora
Niall Rooney, David Patterson, Mykola Galushka, Vladimir Dobrynin [22]	Distributional clustering, Narrow context identificatio n, Minimum spanning tree	Narrow context word discovery	Reuters Corpus Volume 1 (RCV1) collection (35 times Reuters-21578 collection)	Adjacent document similarity, Cluster homogeneity , micro-averaged precision, recall and F-measure, Context/clust er stability over time	Comparison with standard clustering algorithm is not done, the results are poorer for the Industries category set	Subdividing (larger) clusters into sub-clusters using knowledge inherent within the MST, to cluster larger size documents and related to more than one context, apply to soft clustering
Deng Cai, Xiaofei He, and Jiawei Han [23]	Locality Preserving Indexing (LPI), SVD	k-means	Reuters-21578 (8,067 documents)a nd TDT2 (9,394 documents)	accuracy (AC) and the normalized mutual information metric	discarded those documents with multiple category labels, Dimensionality reduction using LPI lacks a strong theoretical foundation	To obtain optimal projection by preserving locality and separating the data points with different labels, existence of a better approximation is unclear
Giansalvator e Mecca, Salvatore Raunich, Alessandro Pappalardo [24]	dynamic SVD clustering based on LSI	Noodles search engines' clustering algorithm	Google's top-ranked search results, Open Directory Project (DMOZ): 12 different datasets (16 100 in each)	F-measure, Grouper Quality Function, Lingo's cluster contaminatio n	Dataset is small	Using in the Web document clustering, integrating with web search engines
Kevin Lind [25]	simplicial complex clustering, Hypergraph clustering	Term Association Discovery	The Reuters-21578 (Modified Lewis Split: 13,625 documents)	Precision	Compared with human categorization	Comparison with standard benchmark algorithms with graphical interpretable results
Yong Wang	WordNet	Word sense	1600	entropy and	dataset small,	Performing syntactic

and Julia Hodges [14]	semantic network	disambiguation method, semantic relatedness measures among senses: senseno method offset method	abstracts from journals belonging to ten different Areas	F-measure	documents are not full length	analysis to find the important word in a context, Combining semantic and syntactic analysis for further improvement
Loulwah AlSumait and Carlotta Domeniconi [4]	Subspace clustering, Semantic LAC	Locally Adaptive Clustering (LAC)	Email-1431, Ling-Spam (453 spam Messages, 561 messages), 20NewsGroup, classic3	average error rate, standard deviation, and the minimum error rate	Evaluation parameters are different, dataset documents are short length	More experiments using different datasets and various feature selection approaches, semantic smoothing of VSM, LSK, and diffusion kernels, an analysis of the distribution of the terms' weights produced by Semantic LAC to identify the keywords to represent semantic topics
Yanjun Li, Soon M. Chung, John D. Holt [6]	sequence of words (meanings), Generalized Suffix Tree (GST), WordNet	Clustering based on Frequent Word Sequences (CFWS) and Clustering based on Frequent Word Meaning Sequences (CFWMS)	Reuters-21578, Classic data set, and a corpus of the Text Retrieval Conference (TREC)	F-measure and purity	Trade-off between the construction time and the memory space requirement when the GST is large.	To use efficient in-memory GST construction algorithms for very large data sets and large GST
Wei Song, Soon Cheol Park [26]	Latent semantic indexing (LSI) and SVD	genetic algorithm based on a latent semantic model (GAL)	Reuters-21578 (only 1600 documents)	F-measure	Dataset is small	To refine the program by reducing the evolving time
Wei Song, Cheng Hua Li, Soon Cheol Park [15]	thesaurus-based and corpus-based ontology: A transformed latent semantic indexing (LSI) model	self-organized genetic algorithm	Reuters-21578 corpus (200 + 600 documents)	multidimensional scaling (MDS) method and SStress criterion for dissimilarity calculation, precision and recall	Dataset is small	more elaborate thesaurus for text clustering as WordNet may not be precise enough to assess the semantic similarities in some specialized domains
Hai-Tao Zheng, Charles Borchert, Hong-Gee Kim [27]	latent semantic analysis (LSA) and gene ontology (GO)	ontological clustering method	PubMed abstract (biomed) dataset (324 + 334)	F-measure	Domain specific, performance depends on the comprehensiveness of the ontology used	more NLP methods, user navigation of abstract, more ML algorithms to estimate the parameter values, other biomedical ontology
Hai-Tao Zheng, Bo-Yeong Kang, Hong-Gee Kim [12]	hypernymy, hyponymy, holonymy, and meronymy	detection of noun phrases with the use of WordNet as background knowledge	Reuters-21578 (200 documents), subset of 20-newsgroup, Reuters-21578 test collection (8,654 documents with one category)	Purity and entropy	performance depends on the comprehensiveness of the noun phrases inspected and the ontology used	more NLP methods to identify noun phrases more accurately, weighing different hypernyms, combine the four relationships with weights to build the feature vectors for documents, domain- and task-specific ontologies
Chun-Ling Chen, Frank S.C. Tseng, Tyne Liang [7]	hypernyms of WordNet	Fuzzy-based Multi-label Document Clustering (FMDC): Association rule mining	Classic, Re0, R8, and WebKB	Overall F-measure	WordNet decreases the clustering accuracy on Re0 and R8 datasets	Combining the syntactic analysis, Incrementally updating the cluster tree
Shady Shehata,	sentence-based,	Concept based mining model,	23,115 ACM abstracts,	concept-based	Comparison with standard benchmark	to link this work to Web document clustering, the

Fakhri Karray, and Mohamed S. Kamel [28]	document-based, corpus-based concept-analysis	Hierarchical Agglomerative Clustering (HAC), Single-Pass Clustering, and, k-Nearest Neighbor	reuters (12,902), Brown, 20 newsgroup	similarity measure, F-measure, entropy, precision	algorithms with graphical interpretable results is not done	usage of the model on other corpora and its effect on classification
Chun-Ling Chen, Frank S. C. Tseng, Tyne Liang [8]	WordNet	Fuzzy Frequent Itemset-based Document Clustering (F^2IDC): fuzzy association rule mining	Classic4 (7094), Re0 (1504), R8 (7674), and WebKB (4199)	Overall F-measure	WordNet for F^2IDC is not likely to work well for text (Re0 and Re8 datasets)	To generate overlapping clusters, Efficient incremental clustering algorithm, To consider the abundant structural relation within Wikipedia
Muhammad Rafi, M. Shahid Shaikh, Amir Farooq [16]	Topic Map representation of the documents	agglomerative hierarchical clustering, Document Clustering based on Topic Maps (TMHC)	Reuters-21578, test collection of Distribution 1.0, NEWS20, OHSUMED	F-measure, purity and entropy		
Young-Min Kim [29]	probabilistic models and topic models	Two-stage PLSA, Ext-PLSA, CS-PLSA, voted-PLSA	Reuters-21578 (4335), Reuters RCV2-French (5000), 20Newsgroups (16010) and WebKB (4196)	Micro-averaged precision and recall, Normalized Mutual Information	More semantic driven methods can be applied	The effect of the length of voting patterns and the number of latent variables in view-specific PLSA models, to expand the area of model selection, unified clustering model which comprises two different types of topics for multilingual corpus (language-specific topic and language-free topic)
Rekha Baghel, Dr. Renu Dhir [11]	frequent concepts: semantic relationship between words	Frequent Concepts based document clustering (FCDC)	Classic (7094), Wap (1560), Re0 of Reuter-21578 (1504)	F-measure, Recall, precision	User provides number of clusters	soft clustering, to generate titles for a document or a group of document
S. C. Punitha, M. Punithavalli [30]	semantic driven methods, ontology-based approach	hybrid scheme for text clustering (HSTC), Text Clustering with Feature Selection (TCFS)	Reuters-21578	F-measure, execution time	TCFS performance is slightly better in quality than HSTC, but it is slow	To combine both these methods to take advantage of quality clustering in a fast manner
S.Vijayalakshmi, Dr.D.Manimegalai [31]	Hypernymy using WordNet, Query based document vector model	K-means	20 Newsgroups (200)	Cluster Accuracy	User must give effective keyword, preprocessing is time consuming process	To identify sparse region and eliminate it to improve the cluster quality
Mari-Sanna Paukkeria, Alberto Pérez García-Plazab, Víctor Fresnob, Raquel Martínez Unanueb, Timo Honkelaa [32]	three feature extraction approaches: fuzzy logic-based, statistical key phrase extraction, and tf-idf	Self-Organizing Map, Hierarchical clustering	English, Finnish, Spanish dataset from Wikipedia (498 HTML and Media Wiki pages)	Precision, recall and F-measure	Articles about only animals in three languages	Apply to other domains than animals and other languages, use rule-based stemmer or to unsupervised stemming method for a larger data set for any language, large dataset, generation of full ontology, to learn more complex conceptual relationships

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 of document organization and management. But, clustering of documents according to semantic features is a challenging problem in text data mining. In this paper the semantic similarity measures are given. Also, a survey of various

clustering techniques with semantics into consideration is done. All these techniques are described in brief. The comparison of these techniques is shown in tabular format with various parameters like the semantic approach applied by author, datasets used, evaluation parameters applied, limitations and future work; which would be very easy for quick interpretation. This survey will be very useful for researchers in this area as there are still many issues that can be taken into consideration for further research.

7. REFERENCES

[1] David Sánchez, Montserrat Batet, David Isern, Aida Valls, "Ontology-based semantic similarity: A new feature-based approach," *Expert Systems with Applications*, Vol. 39, Issue 9, pp. 7718-7728, Jul. 2012

[2] Andreas Hotho , Steffen Staab , Gerd Stumme, "Wordnet improves Text Document Clustering," *In Proc. of the SIGIR 2003 Semantic Web Workshop*, 2003

[3] Stanislaw Osinski, Dawid Weiss, "A Concept-Driven Algorithm for Clustering Search Results," *in Journal of IEEE Intelligent Systems*, Vol. 20 Issue 3, pp. 48-54, May 2005

[4] Loulwah AlSumait, Carlotta Domeniconi, "Local Semantic Kernels for Text Document Clustering," *In Workshop on Text Mining, SIAM International Conference on Data Mining*, 2007

[5] B. Choudhary, P. Bhattacharyya, "Text clustering using semantics," *in Proc of the 11th International World Wide Web Conference*, 2002

[6] Yanjun Li, Soon M. Chung, John D. Holt, "Text document clustering based on frequent word meaning sequences," *Journal of Data and Knowledge Engineering*, Vol. 64, Issue 1, Jan. 2008, pp. 381-404

[7] Chun-Ling Chen, Frank S.C. Tseng, Tyne Liang, "An integration of WordNet and fuzzy association rule mining for multi-label document clustering," *Journal of Data and Knowledge Engineering*, Vol. 69, Issue 11, pp. 1208-1226, Nov. 2010

[8] Chun-Ling Chen, Frank S. Tseng, Tyne Liang, "An Integration of Fuzzy Association Rules and WordNet for Document Clustering," *In Proc. of the 13th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining, PAKDD*, pp. 147-159, 2009

[9] O. Zamir, O. Etzioni, O. Madani, R.M. Karp, "Fast and intuitive clustering of web documents," *in Proc. of the 3rd International Conference on Knowledge Discovery and Data Mining*, pp. 287–290, 1997

[10] O. Zamir, O. Etzioni, "Web document clustering: a feasibility demonstration," *in Proc. of Annual ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 46–54, 1998

[11] Rekha Baghel, Dr. Renu Dhir, "A Frequent Concepts Based Document Clustering Algorithm," *International Journal of Computer Applications (0975 – 8887)*, Vol. 4 **–** No.5, Jul. 2010

[12] Hai-Tao Zheng, Bo-Yeong Kang, Hong-Gee Kim, "Exploiting noun phrases and semantic relationships for text document clustering," *Journal of Information Sciences*, Vol. 179, Issue 13, pp. 2249-2262, Jun. 2009

[13] G.A. Miller, WordNet: a lexical database for English, Commun. ACM 38 (11), 39‑41, 1995.

[14] Yong Wang, Julia Hodges, "Document Clustering with Semantic Analysis," *In Proc. of the 39th Annual Hawaii International Conference on System Sciences*, HICSS, Vol. 03, pp. 54.3, 2006

[15] Wei Song, Cheng Hua Li, Soon Cheol Park, "Genetic algorithm for text clustering using ontology and evaluating the validity of various semantic similarity," *Journal of Expert Systems with Applications*, Vol. 36, Issue 5, pp. 9095-9104, Jul. 2009

[16] Muhammad Rafi, M. Shahid Shaikh, Amir Farooq, "Document Clustering based on Topic Maps," *International Journal of Computer Applications (0975 – 8887)*, Vol. 12– No.1, Dec. 2010

[17] Andreas Hotho, Alexander Maedche, Steffen Staab, "Ontology-based Text Document Clustering," *K¨unstliche Intelligenz*, Vol. 16, No. 4, pp. 48–54, April 2002

[18] Vladimir Dobrynin, David Patterson, Niall Rooney, "Contextual Document Clustering," *In 26th European Conference on IR Research*, ECIR, pp. 167-180, 2004

[19] Farial Shahnaz, Michael W. Berry, V.Paul Pauca, Robert J. Plemmons, "Document clustering using nonnegative matrix factorization," *Information Processing and Management*, Vol. 42, Issue 2, pp. 373-386, Mar. 2006

[20] Chihli Hung, Stefan Wermter, Peter Smith, "Hybrid Neural Document Clustering Using Guided Self-Organization and WordNet," Journal IEEE Intelligent Systems archive, Vol. 19 Issue 2, pp. 68-77, Mar. 2004

[21] Elena Montañés, Irene Díaz, José Ranilla, Elías F. Combarro, and Javier Fernández, "Scoring and Selecting Terms for Text Categorization," *Journal of IEEE Intelligent Systems*, Vol. 20 Issue 3, pp. 40-47, May 2005

[22] Niall Rooney, David Patterson, Mykola Galushka, Vladimir Dobrynin, "A scaleable document clustering approach for large document corpora," *Journal of Information Processing & Management*, Vol. 42, Issue 5, pp. 1163-1175, Sep. 2006

[23] Deng Cai, Xiaofei He, and Jiawei Han, Senior Member, "Document Clustering Using Locality Preserving Indexing," *IEEE Transactions on Knowledge and Data Engineering*, Vol. 17, No. 12, Dec. 2005

[24] Giansalvatore Mecca, Salvatore Raunich, Alessandro Pappalardo, "A new algorithm for clustering search results," *Journal of Data and Knowledge Engineering*, Vol. 62, Issue 3, pp. 504-522, Sep. 2007

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

[26] Wei Song, Soon Cheol Park, "Genetic algorithm for text clustering based on latent semantic indexing," *Computers and Mathematics with Applications*, Vol. 57, Issues 11–12, pp. 1901-1907, Jun. 2009

[27] Hai-Tao Zheng, Charles Borchert, Hong-Gee Kim, "GOClonto: An ontological clustering approach for conceptualizing PubMed abstracts," *Journal of Biomedical Informatics*, Vol. 43, Issue 1, pp. 31-40, Feb. 2010

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

[29] Young-Min Kim, "Document Clustering in a Learned Concept Space," Ph.D. Thesis, Dec. 2010

[30] S. C. Punitha, M. Punithavalli, "Performance Evaluation of Semantic Based and Ontology Based Text Document Clustering Techniques," *Procedia Engineering*, Vol. 30, pp. 100-106, 2012

[31] S.Vijayalakshmi, Dr.D.Manimegalai, "Query based Text Document Clustering using its Hypernymy Relation," *International Journal of Computer Applications* 23(1):13–16, Jun. 2011

[32] Mari-Sanna Paukkeri, Alberto Pérez García-Plaza, Víctor Fresno, Raquel Martínez Unanue, Timo Honkela, "Learning a taxonomy from a set of text documents," *Applied Soft Computing*, Vol. 12, Issue 3, pp. 1138-1148, Mar. 2012