CHAPTER 4

NON-CLASS ELEMENT BASED ITERATIVE TEXT CLUSTERING ALGORITHM FOR IMPROVED CLUSTERING ACCURACY USING SEMANTIC ONTOLOGY

4.1 INTRODUCTION

Clustering is a widely studied data mining problem in the text domains. The problem finds numerous applications in customer segmentation, classification, collaborative filtering, visualization, document organization, and indexing. In this chapter is provided a detailed survey of the problem of text clustering. The key challenges of the clustering problem, as it apply to the text domain. Then on-class key methods are used to improve the mining analysis of text clustering.

Text classification is one of the first applications of machine learning, that applies to the general problem of supervised inductive learning: given a set of training documents, classified to one or more predefined categories, learn to automatically classify new documents. Automated text classification has been used in a number of different applications: automatic indexing, content management, filtering and routing, word sense disambiguation, and keyword search space categorization.

The sparse and high dimensional representation of the different documents necessitate the design of text-specific algorithms for document representation and processing, a topic heavily intent to mining in the information retrieval on graph non class techniques have been proposed to optimize document representation for improving the accuracy of matching a
document with a query. Most of these techniques can also be used to improve document representation for clustering. Compared to previous applications of clustering, three major challenges must be addressed for clustering (hyper)text databases.

- Very high dimensionality of the data: this requires the ability to deal with sparse data spaces or a method of dimensionality reduction.
- Very large size of the databases (in particular, of the World Wide Web): therefore, the clustering algorithms must be very efficient and scalable to large databases.
- Understandable description of the clusters: the cluster descriptions guide the user in the process of browsing the clustering and, therefore, they must be understandable also to non-experts.

4.2 PROBLEM DEFINITION

There are number of clustering methods has been discussed for the problem of text clustering and this section discuss about some of them.

- Retrieving accurate information for users in Search Engine faces a lot of problems. This is due to accurately measuring the semantic similarity between words is an important problem.
- The word “apple” consists of two meaning one indicates the fruit apple and the other is the apple company. So retrieving accurate information to users to such kind of similar words is challenging.
Undesired initiative clusters method to measure semantic similarity between words which consists of snippets, page-counts and support vector machine.

An approach to compute the semantic similarity between words or entities using text snippets. But in this they are going to implement and compute the semantic similarity between words in Search engine without using Snippets or Support Vector Machines.

Because using Snippets or Support Vector Machines makes the job of finding similarity easier. It is implemented the same concept without using snippets or support Vector machines.

All the methods discussed above has the problem of poor clustering accuracy and produces higher false classification ratio.

4.3 NON-CLASS BASED ITERATIVE TEXT CLUSTERING

The method identifies the list of terms and computes the semantic closeness measure. Based on computed closeness measure the method identifies the class of the text using the semantic ontology. Then for each term being identified the method computes the non-class weight for each class. Using both the measures the method assigns the class for the document. The entire document clustering has been split into number of stages namely preprocessing, semantic bound measure and clustering. Each will be discussed in detail in this section.
4.3.1 Preprocessing Text Document

In this stage, the method reads the text document and extract the textual terms from the document. Then for each term extracted, the method performs the stop word removal and stemming process. The method maintains the list of stop words which has no meaning and from the stop word removed content, the method removes the end tokens. Finally the method identifies the list of pure nouns by applying the tagging process. To perform tagging the method use the Stanford part of speech tagger.
Algorithm: 4.1

Input : Document D

Output : Term set ts

Start

Read Document D.

Read document text \( T = \sum Text \in D \)

Identify term set \( Ts = \int Split(T, space) \)

Read stop word list Sw

Remove stop words.

\[ Ts = \sum (Terms(Ts) \in sw) \cap Ts \]

For each term Ti from Ts

Perform part of speech tagging.

\( Ti = \text{Tag}(Ti) \).

If Ti==Noun then

Else

Remove term from set.

End

End

Stop.

The above discussed algorithm identifies the list of pure nouns by performing stop word removal and stemming process.
4.3.2 Semantic Bound Measure

The method maintains the list of semantic ontology which is the collection of properties and classes. For each class, the ontology file contains list of terms and their properties. From the terms identified in the preprocessing stage, the method verifies the presence on each class. For each term the method has been verified for its presence. This is performed for each class of terms. Finally for each class, the method computes the semantic bound measure. Then the method compute the semantic closeness measure based on the bound computed.

Algorithm: 4.2

**Input** : Term set Ts, Semantic Ontology O

**Output** : Semantic Bound Measure SBM, Semantic Closeness measure, Non class elements.

Start

Read Term set ts.

Read semantic ontology O.

For each term Ti from Ts

Compute semantic bound measure $sbm = \frac{Nc}{Tn}$.

End

For each class C

Compute semantic closeness measure $scm$

End

Choose the top closure class $C = O(\text{Max}(Scm))$
Identify non class elements $Ne = \sum Terms(Ts) \not\in O(c)$

Stop

The above discussed algorithm computes the semantic bound measure, semantic closeness measure and identifies the list of non-class elements.

### 4.3.3 Clustering Documentation

The clustering of the text documents is performed by computing the semantic bound and semantic closeness measure. Then the method selects the top class according to the semantic closeness measure. From the selected class, the method identifies the list of terms and with the term set identified from the document text, the method identifies the non-common elements. Using the non-class elements identified, the method computes the frequency of the terms towards each class. Based on the frequency computed and the semantic closeness measure, the method computes the truth weight for each class. Based on truth weight computed the document is assigned with the selected class.

**Algorithm: 4.3**

**Input**: Document D, Ontology O.

**Output**: Class C

Start

Read Document D.

Term set $Ts = \text{Preprocess Document D.}$

$[Sbm, Scm, Ne] = \text{Compute semantic bound measure(Ts).}$

For each class $ci$

Compute term frequency $Tf = \frac{\sum Terms(Ts) \in O(c)}{\text{Number of terms of } c}$

Compute truth weight $Tw = Scm \times Tf$
End

Choose the class with higher truth weight.

Assign class to the document.

Stop.

The above discussed algorithm computes the truth weight to identify the class of the document being submitted.

4.4 SIMULATED RESULTS

In this section, the proposed Method produced a text mining in document processing and the results are displayed below. Figure 4.1 shows the Non-Class Element Based Iterative Text Clustering Algorithm for Improved Clustering Accuracy Using Semantic Ontology. Also, the results of various qualities of service parameters are discussed.
Figure 4.2 shows the result of Non-Class Element Based Iterative Text Clustering estimation produced by the proposed method. The result shows that the semantic bound measure of text clustering is classified from document dataset.
Figure 4.3 Non class clustering SBM accuracy.

Figure 4.3, shows the result of Non-Class Element Based Iterative Text Clustering estimation produced by the proposed method. The result shows that semantic bound measure (SBM) of the clustering accuracy is processed from document dataset.
Figure 4.4, shows the result of Non-Class Element Based Iterative Text Clustering estimation produced by the proposed method. The result shows that the Combined weightage of text clustering is classified from document dataset.

4.5 SUMMARY

In this chapter we discussed and evaluate the presence of non-class elements are used as the key for text clustering. First the method computes the
semantic closeness measure for each class towards the document given. Then for each class, the method identifies the non-class elements. Then the method computes the truth weight for each class based on the number of non-class element and semantic closeness measure. Based on truth weight computed the method selects a class for the input document. The method has improved the clustering accuracy up to 96.4% and reduces the false rate.