CHAPTER 5

GRAPH BASED RELATIONAL DEPTHNESS
ESTIMATION BASED TEXT CLUSTERING USING
SEMANTIC ONTOLOGY

5.1 INTRODUCTION

With the extensive application graph of text mining technology, Rationality and effectiveness of text representation analyses the relational depthness prerequisite to ensure the quality of text mining that are correlated with the text clustering. The ontology term refers. One of representative statistical method is of ontology preference, which builds a statistical vector model based on word frequency. However, the research results have found that the deep semantic and internal structure based on the relational information, it usually correlate to identify a lot of semantic information by evaluating clustering accuracy. Therefore, the semantic text representation is a semantic oriented of identifying relationship cluster evaluation model. It is based on the semantic relationship among the concepts.

The first category, syntax-based service discovery process primarily depends on selecting appropriate keywords and making a query that matches the selected keywords with the Web service descriptions. Because keywords are unable to capture the underlying semantics of the relational text clustering, they may miss some results and may return many irrelevant results. Second, many users would like to express their requests more precisely than is possible using keywords alone, the underlying reason that users want to search for Web services is that they need to find operations that offer a particular functionality. Current mining text discovery documents usually provide details about these
service operations. Hence, service consumers must browse these returned results individually to determine whether each service meets their requirements. Thus, the greater the number of mining results use ontology that emerge, the heavier the burden that to reduce the pattern hold to cluster the resultant.

Hierarchal Document clustering are used in a quantity of relational text that are matched the neighbor that are of text mining and to retrieve semantic information retrieval. Text clustering enhance document relational meaning that improve the information retrieval systems and word of matching case to find the nearest neighbors of a document. More recently, text clustering has been projected for use in observing a set of documents to normalize the results returned by examination of search engine in response to a user’s query. Text Document clustering has also been used to generate hierarchical clusters of documents. Document clustering is measured using hierarchal form in the mining data logs. It’s a essential procedure in mining fundamental structures in text data sets.

![Hierarchical clustering semantic word pair](image)

**Figure 5.1 Hierarchical clustering semantic word pair**
From the above Figure 5.1 hierarchical Text clustering means finding the groups that are related to each other to form word match pair. These groups are collected together in an unstructured formal document to relevant node graphs. In fact, clustering becomes structured relational sets for its competence to analyze the graph of hierarchical word pair in match case model to generalize relational information in case of match pair to form relevant identification.

A directed graph is made up of nodes, the directed edges of connection nodes and the feature weight values of reflecting nodes and edges. If a graph structure is applied to document representation, all feature items in a document will be represented by the nodes, the adjacency and positional relationship between all feature-items will be represented by the directed edges. Then a document may be represented by a graph structure, in which all these nodes will reflect text feature information, the directed edges will reflect the adjacency and positional relationship between all feature items, and the edge weights will reflect semantic relevance between feature items. Therefore, we can see that the text representation based on graph structure can not only well reflect the structure information of document from the definition of graph structure, but also retains semantic feature information of document.

By the hierarchal representation the text clustering analyses lexical patterns documents into various category or groups is essential formation relevant scores of text data. Hierarchical mining be formalize the applicable resource to improve the data mining relevant searching which spread over to virtually every ground where there are large volumes of material that need for group. Extracting apposite feature and representative it in a expressive way is measured the relevance of important in document clustering. Clustering Text documents are often characterized as high-dimensional, sparse vectors and complex semantics.
The semantic cluster relationship in classification and the hierarchal relationship between the components of noun relevant and among nominal in general are not easy to categorize rigorously. So the need to find a document feature where the entities in the same document that related linguistically that are shown in figure 5.2 to respectively other are demonstrated. In these cases, the cooperative entity related with a link, and that the designation declared in the same text is joined by exploiting the interdependence amongst them. The research shows that finding semantic relationship can lead to improvement in the entity linking accuracy.

A relational identification are carried out graph based text clustering by document clustering is projected to solve the semantic relational difficult because preceding studies established that it is able to detention the relative among words and recover the presentation of a comparison measure between iterated texts. In calculation hierarchal assembly advances the semantic cluster precision and complete enhancement to have great performance.
The main imperative is hierarchal that relates the relational information extraction holds irrelevant semantic problem associated with semantic Representation can be reduced. In directive relational identification to semantic address their sub-clusters are formulated and to represent text into graph based picture to reduce the clustering information of relational and capture the semantics of text. A graph based text representation method produce good clustering results.

Simultaneously, the operations and their parameters provide more comprehensive interaction details than the description itself. Therefore, the most effective information to express the interaction devolves to the names of operations and their parameters. Many researchers consider operation names and parameters as important objects for semantic mining. However, the information contained in the names is quite limited, making it difficult to mine and apply their semantics. Associated with other types of text illustration identification which the relevant scheme, Relational Graph method is supplementary to lexical analyze since it is hieratically categorized by its facility to capture the relation among words in text. Graphs are scientific constructions which comprise of nodes and edges which link nodes organized. This form of depiction embraces formal examination of impressions and graphs that relate to identification by conceptual groping relational text. The basic idea is to convert documents to its syntactic structure. Semantic Sentence structure are generalized the semantic level are hierarchal represented by similar dependency graphs. The document is categorized as a graph with opinions as nodes and connection as arc. In case of resemblance between documents in dissimilar clusters and edge performs among two nodes. If the documents that are limited in the cluster are extremely related, the edges in the similar cluster will consider more than the boundaries crossways clusters.
Text clustering Dependency graph can also be well-defined as a directed graph representative of identifying relational of several objects near each other. It is probable to derive an relational evaluation order or group of structural hint order that compliments the given dependences from the relevant identification graph. Hierarchal relational for carrying information that this type of text clustering. The competence to contemporary information more professionally than other type of text illustration given below.

![Graph feature formation of Semantic cluster](image)

**Figure 5.3 graph feature formation of Semantic cluster**

The same time the text representation of graph structure value added semantic identity formation of cluster mining. In Figure 5.3 shows the text representation, if clustering adopt the semantic cluster relative match case and the word co-occurrence relation of text are measured the identity of graph structure the graph construction is not only intelligent to characterize text semantic, but also is a respectable relational representation model for text information. Therefore this hierarchical using a grouping of text semantic evidence and the features of word co-occurrence to hypothesis a graph construction of document, and the technique is practical in interpersonal text clustering in order to recover the clustering effort.
5.2 PROBLEM DEFINITION

- The problem faced in clustering high-dimensional data is sparsity of this data. Many researchers have tried for segmentation reveals.

- The unformatted clustering problem where added sub cluster of dictionary representation step to the k-means group clustering development to spontaneously calculate the weights of all scopes of relations in each cluster.

- Addressing problem begins sparsity issues to ensemble the cluster. Cluster members are not well as similarity measurement to object meaning as boundary points.

- The data on both synthetic and real that the new algorithm can generate results clustering not well edge mapping consolidation.

- Bag of words are not sufficient for extract content similarity.

- Textual content holder forms cluster based sub band similarity measure heavenly mismatch relational content word pairs.

- Depthless estimation need the graph estimation point. Semantic evaluation that doesn’t measure the semantic similarity word pairs.

5.3 GRAPH BASED RELATIONAL DEPTHLESS ESTIMATION

The problem of text clustering has been approached using semantic graph in this method. The method uses the semantic ontology which contains the collection of classes and their properties. The properties and the classes are arranged in different hierarchical level. Using the semantic ontology the method generates the semantic graph and for the input text document the
method extract the terms using the preprocessing algorithms. Then for each term from the term set, the method computes the depthless estimation for each class. Based on the depthless estimated, the method selects the class of the document.

![Graph based relational Estimation model](image)

**Figure 5.4 Graph based relational Estimation model**

### 5.3.1 Preprocessing Relational Text

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 ontology relevant information selection process. The method maintains relevant information search which has semantic feature selection relational word and from the relational word removed content; the method removes the end tokens. Finally the method identifies the list of pure nouns by applying the tagging process for ontology representation of information gathering.
Relational process can be identified by the following equation,

\[ RT(t) = \frac{\text{term to searching the data in document}}{\text{total number of documents in dataset}} \times (t-\text{in terms of time frequency}) \]

To finding semantic relational text,

\[ SRT = \frac{\log(\sum \text{total relational process}(RT))}{\text{total number of Terms in document}} \]

Relational information mining from document,

\[ RI = \frac{RT(t) \times SRT}{\text{Total time by document}} \]

### 5.3.2 Semantic Graph Generation

The semantic graph is generated using the ontology available. First the method initializes the root tree. Then for each class a single node is generated under the tree. Then for each properties present and class present under each semantic class, a leaf is generated and number of nodes is generated based on the properties presented. This will be iterated for each property class present under the semantic class. Generated semantic graph will be used to perform clustering.

**Algorithm: 5.1**

The clustering algorithm of relational semantic graph structure.

Input: The graph set is \( G = \{G_1, G_2 \ldots, G_m\} \) and the clustering number is \( k \).

Output: The clustering of every graph is \( 'C_j \ (j \in [1, k]) \) and the clustering center is \( M_j (j\in [1,k]) \).
Step 1: The structure graph $G$ random

Step 2: Calculate the clustering center $M_j(\sum_{1,k})$ of each clustering $C_j(\sum_{1,k})$.

Step 3: Respectively calculate the distance between every graph $G_i(\sum_{1,n})$ and $M_j(\sum_{1,k})$ in $G$. $G_i(\sum_{1,n})$ will be subsumed into the clustering of closer distance. And then obtain new clustering as $C_j(\sum_{1,k})$.

Step 4: Then calculate the center $M_j(\sum_{1,k})$ of every new clustering $C_j(\sum_{1,k})$.

Step 5: Repeat step 3 to step 4, until $M_j(\sum_{1,k})$ does not change.

The Minimal structural diagram with an average distance of other diagrams is called as median graph. Median figure is actually a center of clustering, the calculation formula is as follow.

$$g = \text{semantic min} \left( \frac{1}{G} \sum_{m=1}^{G} s(\text{Rule } R), Sm \right)$$

The set of graph structure is $G = \{G1, G2, ..., G\}$ and there is a graph structure $g$ that meets $g > G$. If the average distance between $g$ and all of the elements in $G$ is the smallest value,

### 5.3.3 Semantic Depthness Measure

The semantic depthness is estimated using the semantic ontology. First the method counts the number of properties present in the each level and the number of properties gets matched. Then based on the count produced, the method computes the frequency of presence. Using the frequency computed for
each level, the method computes the semantic depth measure. Computed value will be used to perform clustering.

If $V_i$ and $V_j$ is synonymous relationship in knowledge network, so the semantic relevancy between $V_i$ and $V_j$ is 1, that is $\sum (V_i, V_j) = 1$. Otherwise, its semantic relevancy is calculated as follows

$$\sum (V_i, V_j) = \frac{(\text{distance}(V_i, V_j) + \beta \text{Seq} \times \mu \times (d(V_i) + d(V_j)))}{N(V_i, V_j) \times 2 \times \text{seq} \times \text{max}(d(V_i) - d(V_j))}$$

Where $d(V_i)$ and $d(V_j)$ respectively express the hierarchy of $V_i$ and $V_j$ corresponding to the node of ontology tree. Distance $(V_i, V_j)$ is the weight-value sum of all edges on the shortest path between $V_i$ and $V_j$ also $N(V_i, V_j)$ is the number sum of edges on the shortest path between $V_i$ and $V_j$. Seq is the maximum depth of ontology tree, $\beta$ is an adjustable parameter.

### 5.3.4 Clustering Text Documents

Text document clustering is a selection of text documents with the particular word(s) present. So each group of text documents called cluster of text documents of a particular word’s presence. Unsupervised relational learning is forms the text clustering and no direction for relevant means that nearby is base of case matching proficient that has allocated documents to cluster. In text document clustering, it is the distribution and makeup of the text documents as a group based on a particular word present in all the grouped text documents. Clustering is sometimes referred to as automatic classification.

The clustering of the text documents is performed by computing the semantic depthless measure. First the method performs preprocessing to obtain
the pure terms from the document. Then the method generates the semantic graph and using the graph generated, the method computes the semantic depthness measure for each class. Based on semantic depthness a single class is selected and assigned for the document.

**Algorithm: 5.2**

Input: k items

\[ D = (x_1, x_2, x_3 \ldots x_n) \]

Output: \( k = (C_1, C_2, C_3 \ldots C_k) \)

Step 1: Assign initial value for mean depth points

\[ A_1, A_2, A_3 \ldots A_K \] //k seeds

Repeat

Step 2: Assign each item \( x_i \) individual for the cluster which has the closest mean;

Step 3: Calculate new mean depth for each cluster;

Until the stemming was performed;

Step 4: int result=0;

Step 5: for each \( D \) in intermediate values;

Step 6: result += \( (D) \);

Step 7: Compact docs (String (result));

From the above algorithm text represent \( X \) terms as input document and \( C \) as clustering, a group of words are used on a collection of text documents to calculate mean. A for discovering such text documents having with the given set of words. Further such discovered text documents for the given set of words are grouped into that many cluster of text documents.
5.3.5 Query Clarity Feature Selection

On the other hand, Feature selection algorithms aims in reducing the high dimensionality of the feature vector into a lower dimensional space by selecting the best subset features from the original feature set. Most of this methods have been Features are derived from resource selection methods developed in the context of distributed semantic learning. Query performance prediction approaches can also be recycled for type prediction by transmitting a higher score for the gathering with higher predicted resultant information performance be gathered for relevant identification. Among such methods, to employ Query Clarity, which predicts performance using the Knowledge learning based on match case query evaluation and relevance classification.

5.3.6 Dictionary Based Matching

In relevant match case users searches semantic based on direct signs about relational measure of information which category they intended to search the documents, by including terms. While these terms may not happen in a popular of queries, they can be a durable indication of type relationship for a given query of document clustering. They built a dictionary for match case of words by using the documents of the information collection and metadata identified fields.

5.3.7 Field-based Collection Query Likelihood

Although some of existing type prediction methods use the collection term statistics, none use the field structure of documents available for personal information collection. Considering that the retrieval effectiveness of semi-structured document collections has been improved by exploiting the field based collecting this structure and expect similar benefits for the type prediction problem.
5.3.8 Relational Information of Iterative Grid Search

Information extracted from iterative grid search means the semantics features extraction be extracted at the mean evaluation to be mutual data are feasible to achieve a neighbor graph preference of semantic values that exploit the variance of performance which has training set of queries. Definitely, To find the optimal value for each parameter in turns while fixing the values of the parameters beforehand found, and retelling the whole process until the information carrying document cluster.

5.3.9 Combining Type Prediction Methods

Predicting information familiarized to preprocess the syntactic pattern type prediction approaches introduced to combined supervised classification or as semantic search case, it is plausible that to can get further performance benefits by combining individual methods in a linear model where weights are found using learning methods. The classification that carried different methods with different objective functions: relational grid search of attribute values, rank-learning method and a multi-class classifier.

5.3.10 Rank Learning Method

Alternatively, Rank prediction can carrying relational one cast the type prediction information as a ranking values can carried out relevant collections higher than non-relevant collections formation that are semantically retrieved. This method can be particularly beneficial of information for finding multiple documents with different types, since cluster evaluates the classification higher preference as rank to similarity retrieval of information have produced optimized results is model with typical multi-class classification methods.
5.4 SIMULATED RESULTS

In this section, the proposed Method produced a text mining in document processing and the results are displayed below. Figure 5.1 shows the Graph Based Relational Depthness Estimation Based Text Clustering Using Semantic Ontology. Also, the results of various qualities of service parameters are discussed.

![Figure 5.5 dataset initiated process in clustering model](image-url)
Figure 5.5, shows the result of Graph Based Relational Depthness Estimation produced by the proposed method. The result shows that the initiated process text clustering is classified from document dataset. Initially semnatic bounding measure is calculated at the ontology process. Additionally semantic bound measure was iterated as the possibility measure produce higher efficiency result.

![Text Clustering Using Semantic Ontology](image)

Figure 5.6 relational deplness Rrelational clustering accuracy.
Figure 5.6, shows the result of Graph Based Relational Depthness Estimation produced by the proposed method. The result shows that the Relational clustering accuracy is processed from document dataset.

![Graph Based Relational Depthness Estimation](image)

**TEXT CLUSTERING USING SEMANTIC ONTOLOGY**

<table>
<thead>
<tr>
<th>Category Name</th>
<th>Rel-weight</th>
<th>SDM</th>
<th>SCW</th>
<th>Combined We.</th>
</tr>
</thead>
<tbody>
<tr>
<td>acq</td>
<td>7080</td>
<td>264.0</td>
<td>7080.0</td>
<td>4808.0</td>
</tr>
<tr>
<td>corn</td>
<td>28</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>crude</td>
<td>1602</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>earn</td>
<td>4393</td>
<td>56.0</td>
<td>4393.0</td>
<td>2947.333333...</td>
</tr>
<tr>
<td>grain</td>
<td>1162</td>
<td>75.0</td>
<td>1162.0</td>
<td>798.666666...</td>
</tr>
<tr>
<td>interest</td>
<td>943</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>money-fx</td>
<td>1495</td>
<td>10.0</td>
<td>1495.0</td>
<td>1000.0</td>
</tr>
<tr>
<td>ship</td>
<td>595</td>
<td>7.0</td>
<td>595.0</td>
<td>390.0</td>
</tr>
<tr>
<td>trade</td>
<td>1212</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>acq</td>
<td>7021</td>
<td>61.0</td>
<td>7021.0</td>
<td>4701.0</td>
</tr>
<tr>
<td>corn</td>
<td>28</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>crude</td>
<td>1602</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>earn</td>
<td>4393</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>grain</td>
<td>1162</td>
<td>23.0</td>
<td>1162.0</td>
<td>782.333333...</td>
</tr>
<tr>
<td>interest</td>
<td>943</td>
<td>10.0</td>
<td>943.0</td>
<td>032.0</td>
</tr>
<tr>
<td>money-fx</td>
<td>1495</td>
<td>5.0</td>
<td>1495.0</td>
<td>998.333333...</td>
</tr>
<tr>
<td>ship</td>
<td>595</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>trade</td>
<td>1212</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
</tbody>
</table>

**Figure 5.7  Relational semantic depthness and clustering measure**

Figure 5.7, shows the result of Graph Based Relational Depthness Estimation produced by the proposed method. The result shows that semantic and clustering weightage of text clustering is classified from document dataset.
5.5 SUMMARY

In this chapter we discussed and evaluate the graph based relational depthness estimation based text clustering using semantic ontology. Initially semantic bound measure iteratively calculate the semantic linkage. to improve the document clustering by the evaluation of depthness measure based on the predictive information produce higher efficiency result produce 97.6% clustering accuracy compared to prior methodology.