1 ABSTRACT

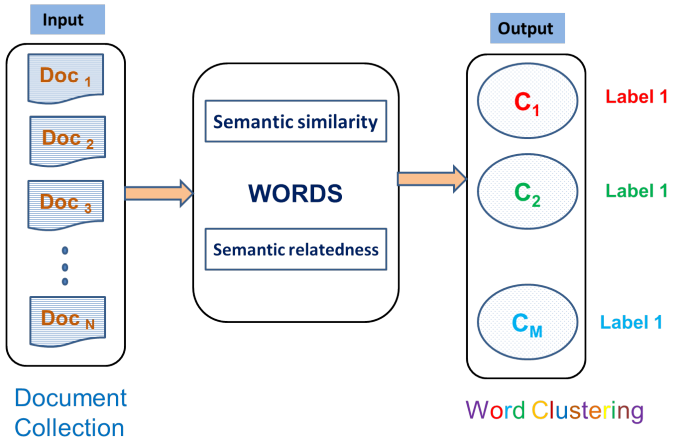
Word clustering is a technique for partitioning sets of words into subsets of semantically similar words and is increasingly becoming a major technique used in a number of Natural Language Processing tasks ranging from word sense or structural disambiguation to information retrieval and filtering. This project involves a set of sequence of preprocessing steps that represent the data better for clustering. This also demonstrates various clustering algorithms such as Affinity propagation, Agglomerative and Markov Clustering to infer the similarity among the words. It also includes labeling the word clusters automatically based on heuristics using WordNet ontology. This is experimented on synthetic data and clusters along with labels are shown in results.

2 INTRODUCTION

Word clustering represents similar words as groups according to their contextual, syntactic, or semantic similarity. It can be considered as the most important unsupervised learning problem. It deals with finding a structure in a collection of unlabelled data. In this project, the existing methodology[1] for document clustering has been modified in two-levels that adapt to word clustering. These enhancements focus on pre-processing the corpus that combines the traditional vector space model (VSM) and the term mutual information (TMI) based on ontology using WordNet. However, the preprocessed corpus results into term vector that is directly considered as input to the clustering process. Following are the contributions:

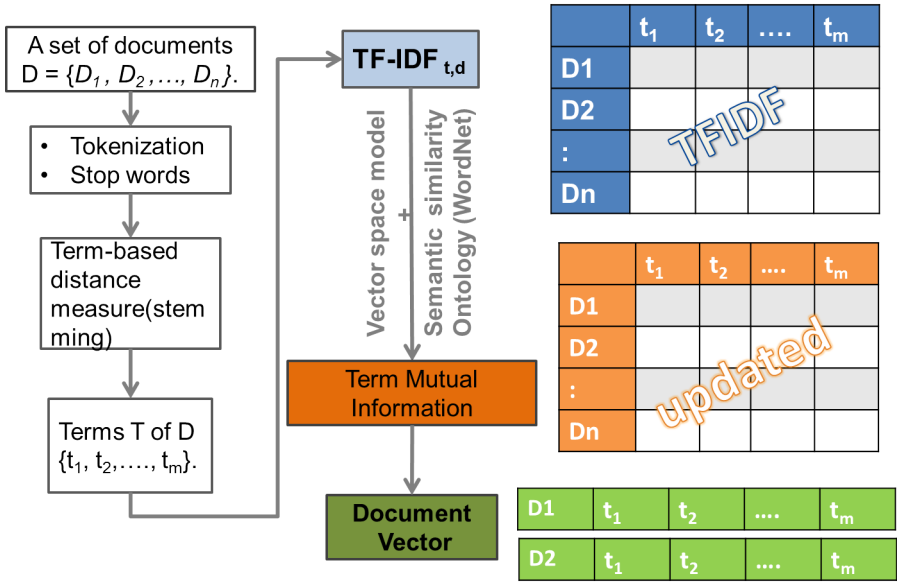
- In pre-processing the corpus:
  - Weighting scheme based on TF-IDF is used to represent document vector, and
  - Term vector space conversion from Document vector space and normalization.
- In clustering:
  - Employing Affinity propagation, Agglomerative, and Markov clustering.
  - Utility for better graphical representation of the clustered outputs.
- In Labeling:
  - A new heuristic based approach is proposed using WordNet Ontology.

3 PROBLEM DEFINITION



Given a Document collection, Partition words into clusters such that words that belong to the same cluster are semantically similar and related. Assign labels automatically to the clusters.

4 METHODOLOGY



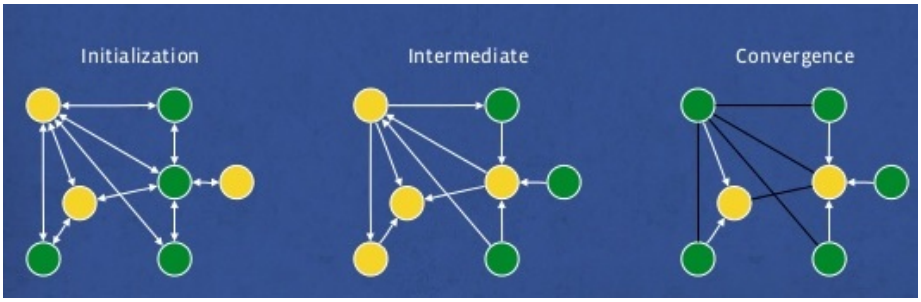
4.1 Preprocessing

- 1. Input: the traditional term-based vectors for the corpus, the number of terms (m) and the number of doc-uments (n), and parameter Îť to indicate the semantic information between two terms in WordNet;
- 2. Use WordNet modify the term-based VSM to ontologybased VSM in the linguistics preview.
  - For all terms in the vocabulary of the corpus, use WordNet to find the semantic relationship between each pair:
    - IF t_{i1} âĹĹ Synsets(t_{i2}) or t_{i2} âĹĹ Synsets(t_{i1}) ((t_{i1}, t_{i2}) makes sense), then set the semantic relationship $_{i1i2}$ between t_{i1} and t_{i2} to Îť;
    - ELSE the semantic relationship $_{i1i2}$ is assigned to zero.
- Find Term Mutual Information using WordNet
- Update semantic relatedness using summation of all the term.
- Transpose them to term vector and uses cosine similarity to find the Similarity Matrix between every pair of term.
- Pass Similarity Distance Matrix to Clustering Algorithms.

5 CLUSTERING

- **Agglomerative Clustering**
  - Compute the distance matrix between the input data points
  - Let each data point be a cluster
  - Repeat

    - Merge the two closest clusters.
    - Update the distance matrix.
  - Until only a single cluster remains.
  - Key operation is the computation of the distance between two clusters.
  - Different definitions of the distance between clusters lead to different algorithms.
- **Affinity Clustering**
  - It is a clustering algorithm based on the concept of "message passing" between data points.
  - Discover exemplars based on similarity.



- **Markov Clustering**

Expand: $M_{exp} = Expand(M) \stackrel{def}{=} M * M$

Inflate: $M_{inf}(i,j) \stackrel{def}{=} \frac{M(i,j)^r}{\sum_{k=1}^{n} M(k,j)^r}$

Prune: zero out the least elements in each column

Repeat until M converges

5.1 Labeling

- Labeling using Heuristics based on keywords
- Input: List of keywords from Document collection.
- Select top scored words from each cluster.
- Similarity measure between these top scored words in a cluster and keyword.
- Select the keyword with best similarity score.
- Obtain hypernym of the selected keyword and consider that as label for that cluster

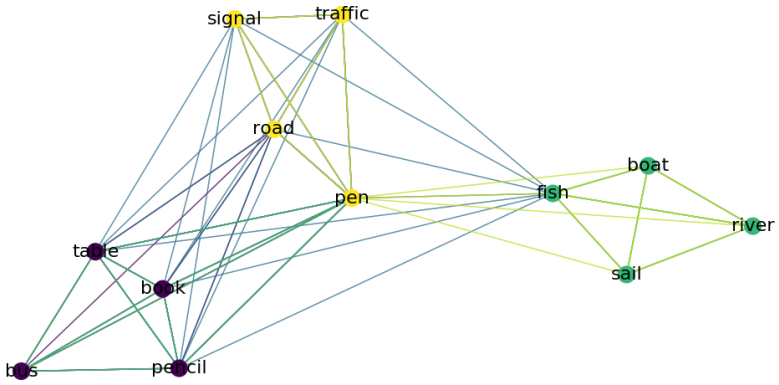
6 RESULT

6.1 Setup

- Documents 1: book,pencil,bus,table
- Documents 2: signal,road,pen,traffic
- Documents 3: fish,boat,river,sail
- Documents 4: pencil,pen

6.2 Clusters

- **Affinity Propogation**
  - Pencil
  - Bus
  - Signal
  - boat, book,fish,pen,river,road,table,traffic
- **Markov Clustering**



6.3 Labeling

7 CONCLUSION

From a corpus of specific domain, clustering of words that are similar and labeling the clusters automatically is a non trivial task in Natural Language processing. This project demonstrated the methodologies to find the similarity among the words using various clustering algorithms. It also includes a new heuristic based approach for labeling the word clusters automatically.This heuristics require suitable ontology that are completely depends on the domain. This project is experimented over a synthetic data. It also observed that preprocessing the corpus is the most computationally intensive step.