An Efficient Kernel Affinity Propagation Method for Document Clustering

S. Rathinaparimalam
PG Scholar
Department of Information Technology
Bannari Amman Institute of
Technology,Sathyamangalam, Tamilnadu

G. Srinitya
Assistant Professor (Sr.Grade)
Department of Information Technology
Bannari Amman Institute of
Technology,Sathyamangalam, Tamilnadu

ABSTRACT

Semi-supervised learning method is a new interesting direction of machine learning approach. It gives the computer a learning ability and makes good use of the obtained knowledge in the application. Semi-supervised learning performs the process of data analysis and mining effectively with the help of few exemplars or little pre-known information. A new Non-Euclidean Space similarity measurement contains the structure information, which is proposed in the Tri-Set computation method. The new similarity measurement not only attentions on the Euclidean Space constraint, but also gives the basic information about the text files. This method is named as Kernel Affinity Propagation (KAP). The method is applied to the benchmark data set Reuters-21578. Further the result is compared with the k-means algorithm and original Affinity Propagation algorithm. The comparison result shows that KAP improves the clustering execution time and provides the better clustering output.

Keywords

Document Clustering – Semi supervised learning-Smilarity measurement- Message Matrix Computation-Kernel Affinity Propagation method.

1. INTRODUCTION

Document clustering is the task of spontaneously organizing text document into meaning full cluster or set. In other words, the documents in one group share the same topic and the documents in different groups represent different topics. Clustering the documents is one of the most significant tasks in text mining. There are various number of techniques launched for clustering documents since there is quick growth in the field of internet and computational technologies, the field of text mining have a sudden growth, so that simple document clustering to more interesting task such as construction of granular classifications, sentiment analysis, and text summarization for the scope of devolving higher quality data from text[1].

The process of Document Clustering aims to discover ordinary groupings, and thus present a summary of the classes in a collection of documents. In the field of machine learning, it is known as unsupervised learning. In a clustering problem, the number, properties, or Relationship of classes is known in advance. Clustering text documents by classifying a subset of representative instances play an important role in recent text mining and information retrieval research. In fact, organizing a huge amount of objects into significant clusters is often used to browse a collection of objects and establish the results returned by a search engine[2][3].

After the grouping process, the obtained clusters are represented with models, which can include all part of the features that appear in the cluster fellows. During cluster-based query processing, simply those clusters that contain examples related to the query are considered for advance relationships with cluster members, e.g., documents. This scheme, occasionally called Cluster-Based Retrieval, is proposed to increase both efficiency and effectiveness of the document retrieval structures[4][5][6].

Figure 1 illustrates the overall steps involved in the document clustering process. It is a more exact technique for unsupervised document association, automatic topic removal and fast information retrieval. Clustering documents involves four stages. Each stage has multiple sub stages. The main four process involved in document clustering are

1) Preprocessing of raw data
2) Feature Selection/Extraction
3) Clustering Techniques/Algorithms
4) Clustering Results

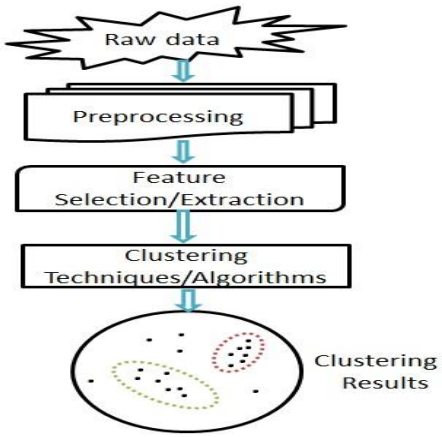

Fig 1: Overall Process in clustering

1.1 Challenges in Document Clustering

Document clustering is being studied from many decades but still it is far from a trivial and solved problem. Some of the challenges are mensioned below:

1. Choosing appropriate features of the documents that should be used for clustering.
2. Selecting an applicable similarity measure between documents.
3. Selecting an applicable clustering method utilizing the above similarity measure.
4. Implementing the clustering algorithm in an efficient way that makes it achievable in terms of required memory and CPU resources.
5. Finding ways of assessing the worth of the performed clustering.

2. SEED CONSTRUCTION METHOD

2.1 Preprocessing

Preprocessing of the data set is used to improve the quality of the clustering process. In this modules the following are carried out .Data are collected from Reuter's dataset. The data set will be enclosed with HTML tag and syntaxes. In pre-processing method want to remove the noise data from our dataset like

- Stripping process
- Splitting Word
- Stemming Word
- Stop Word

Stripping is designed to remove unnecessary tagging and untagging operation from automatically generated programs and make that for splitting operation. The Reuters data set contains special tags such as "<TOPICS>" and "<DATE>" etc. The stripping process strips the document from the special tags.

Split word is used to split the paragraph into word and this word is used for next Pre-Processing methods. The splitting phase of the data set cuts the files into single words.

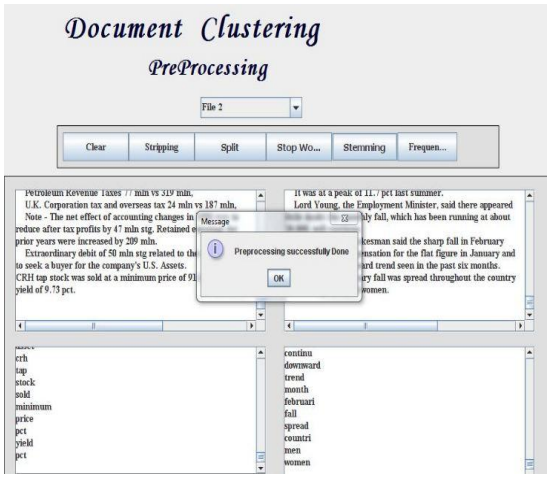

Fig 2: Preprocessing of Retures

Stop words are words which are filtered out earlier to, or after, processing of natural language data (text). It is organized by human input and not automated. Some tools specially avoid using them to support phrase search. Any set of arguments can be chosen as the stop words for a given determination. Stop word is used to filter out Prepositions, Conjunctions, and Pronouns Words that occur in the document. Such words have no standards for retrieval purpose, such as is, the, which, at and on.

Stemming is significant process for web pages and search engine optimization. The stemming procedure tries to eliminate inflectional and derivational suffixes in order to conflate word alternates into the same stem on root.A word is reduced to its root through a language dependent method stemming.It is used to replace all the variants of a word with the single stem of the word. Variants include plurals, third person suffixes, past tense suffixes; etc. Stemming increases the storage and search efficiency. In this assignment Porter algorithm is used for stemming process. Porter algorithm is a linear algorithm; exactly it has five phases applying rules within each step. In each step, if a suffix regulation coordinated to a word, then the conditions attached to that rule are verified on what would be the resulting stem, if that suffix was detached, in the way well-defined by the rule.

2.2 Extract Significant Features

Significant terms (or keywords) are set of meaningful terms in a text document that give a high-level description of the content for person who reads. Significant terms removal is a elementary step for several tasks of natural language processing. In these modules the following are carried out. These "most significant" features could be main phrases related with each document when available. In this, they could be all the words in the title of each document. Thus the most significant words are extracted. Basic Steps for extracting most significant words from the document is

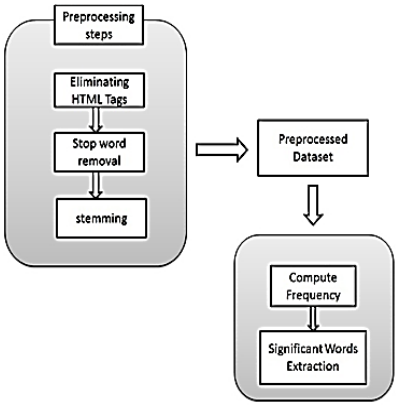

Fig 3: Preprocessing and Extract Significant Features

It is the process of calculating the amount of times a word appears in one or more text files. And it implements the word frequency examination and list the number of times each word appears. Finally it will be arranged by frequency.

Fig 4: Extract Most Significant Features

Extracting the feature is the form of transferring the input data into the set of features. Feature extraction projects a data set with greater dimensionality onto a lesser number of dimensions. Feature extraction can also be used to improve the speed and efficiency of learning process. From the process of frequency computation it rejects the least frequency word occur in the text files.

2.3 Seeds Construction

In this module the following are carried out in Affinity propagation algorithm[7]. Data points can be exemplar (cluster center) or non-exemplar (other data points). But in this paper exemplars are called as seeds. The seeds are constructed using the Tri-set computation approach.

2.4 Kernel Affinity Propagation Method

Based on AP method, here proposed a unique method called "Kernel Affinity Propagation". There are various new structure methods which that includes: Tri-Set computation, similarity measurement, seeds construction, and message transmission. i.e., three new feature sets, are constructed named by

- Cofeature Set (CFS),
- Independent Feature Set (IFS),
- Substantial Cofeature Set (SCS)

The structural information of the documents is involved in the novel similarity measurement. Then, original AP approach is extended as a semisupervised learning strategy.

2.4.1 Similarity Measurement

Similarity measurement plays an significant role in affinity Propagation clustering. In demand to give exact and actual similarity measurement for specific domain, i.e., to define these Tri-sets, a detail computation of the new features is done. In novel methodology, every term in text is still estimated as a feature and each document is still deemed as a vector. However, all the features and vectors are not calculated simultaneously.

Let D be a set of texts

$$D = \{d_1, d_2, ..., d_N\}$$

Assume that di and dj are two objects in D, they can be denoted using the following two subsets:

$$d_i = \left\{ \left\langle f_i^1, n_i^1 \right\rangle, \left\langle f_i^2, n_i^2 \right\rangle, ... \left\langle f_i^L, n_i^L \right\rangle \right\},$$

$$d_j = \left\{ \left\langle f_j^1, n_j^1 \right\rangle, \left\langle f_j^2, n_j^2 \right\rangle, ... \left\langle f_j^L, n_j^L \right\rangle \right\},$$

Let Fi and Fj be the feature sets of the two objects,

$$F_i = \left\{ f_i^1, f_i^2 \cdots f_i^L \right\}$$

Respectively

$F_j = \left\{ f_j^1, f_j^2 \cdots f_j^M \right\}$. Let the set composed of the "most significant" types of dj. "That is capable of representing crucial aspects of the document. These "most significant" features could be key phrases associated with each document when available. Or, as used in the experiments, they could be all the words in the name of each document.

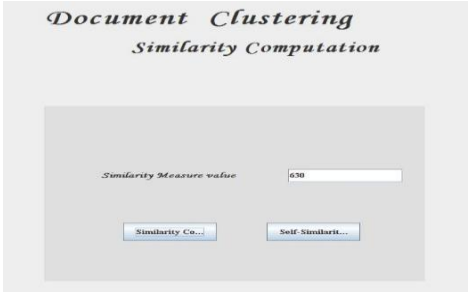

Fig 5: Similarity Computation For the input

2.4.2 Compute the Tri-set

1. Cofeature Set.

Let d_i and d^j be two objects in a data set. Suppose that some features of d_i, also belong to d^j. Consequently, new two-tuples subsets consisting of these features and their values in dj are constructed.

Cofeature Set between di and dj

$$\left\langle f_m, n_m \right\rangle \in CFS_{(i,j)}, \text{iff } f_m \in F_{(i,j)}, \text{and } \left\langle f_m, n_m \right\rangle \in d_j$$

$$F_{(i,j)} = F_i \cap F_j$$

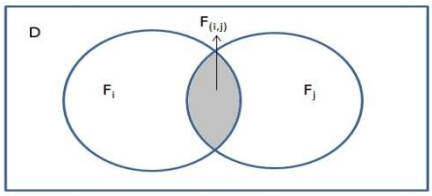

Fig 6: Venn diagram of $F_{(i,j)}$

2. Independent Feature Set.

Suppose that some features of d_i, do not belong to d_j. Subsequently, we can construct a new two-tuples subset containing of these features and their values in d_i.

It is defined as

$$d_j : \left\langle f_p, n_p \right\rangle \in IFS_{(i,j)} \text{ and } \left\langle f_m, n_m \right\rangle \in d_j$$

$$\bar{F}_{(i,j)} = F_i - F_{(i,j)}$$

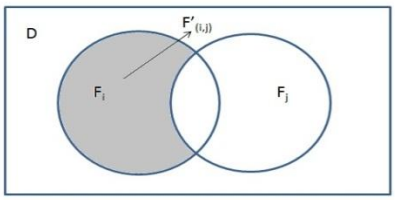

Fig 7: Venn diagram of $F'_{(i,j)}$

3. Substantial Cofeature Set

Suppose that some features of d_i, also belong to the most significant features of d_j. Consequently a novel two-tuples subset consisting of these types are constructed and their values as the most significant features in d_j.

$$\left\langle f_q, n_q \right\rangle \in SCS_{(i,j)} \text{ , where } f_p \in \hat{F}_{(i,j)}, n_q$$

$$\hat{F}_{(i,j)} = F_{(i,j)} \cap DF_j$$

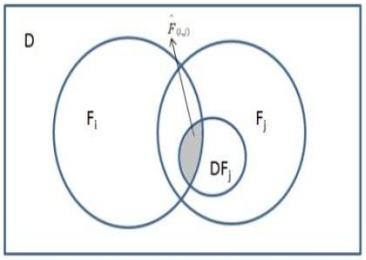

Fig 8: Venn diagram of $\hat{F}_{(i,j)}$

Fig 9: Tri-Set Computation

2.4.3 Message computation

The KAP approach computes two kinds of messages exchanged between data points. The first one is called "responsibility" r(i,j) : it is sent from data point 'i' to candidate exemplar point 'j' and it reflects the accumulated evidence for point 'j' is to serve as the exemplar for point 'i'. The second message is called "availability" a(i,j): it is sent from data point 'j' to applicant exemplar point 'i' and it reflects the accumulated evidence for how appropriate point 'i' to choose point 'j'as it exemplar At the opening, the availabilities are initialized to zero a(i,j) = 0.

In this module the AP algorithm is used partially for Prediction of cluster .In AP algorithm there are some representations.

Exemplar:

A data point that is well representative of itself and some other data points.

Data Points:

One item of data; affinity propagation clusters data points To evaluate the correct exemplars and data points using certain matrix are computed.

Similarity measurement :

The similarity s(i,j) indicates how well x_j is suited to be the exemplar for x_i. for instance , it can be initialized to

$$s(i, j) = -\left\| x_i - x_j \right\|^2, \; i \neq j$$

Responsibility r(i,k) is computed

A non-exemplar data point informs each candidate exemplar whether it is suitable for joining as a member for this a measurement is used.

$$r(i, j) = s(i, j) - \max_{j' \neq j} \left\{ a(i, j') + s(i, j') \right\}$$

Availability a (i,j) is Evaluated

A candidate exemplar data point informs other data points whether it is a good exemplar. This metrics intimates the availability of itself within that cluster.

$$a(i, j) = \begin{Bmatrix} \min \left\{ 0, r(i, j) + \sum_{i' \neq i, j} \max \left\{ 0, r(i', j') \right\} \right\}, i \neq j \\ \sum_{i' \neq i} \max \left\{ 0, r(i', j) \right\}, \qquad i = j \end{Bmatrix}$$

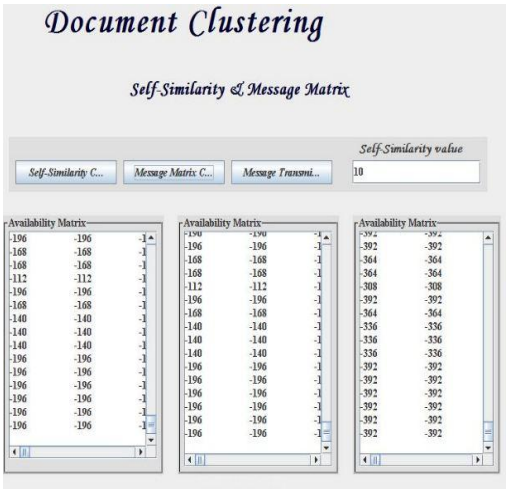

Fig 10: Self Similarity and Message matrix Computation

2.4.4 CLUSTER FORMING

The main knowledge of clustering is to find which documents may have words in common and place the documents with the most words in common into the same groups. For the text clustering it is needed to evaluate certain data points and exemplars which are called as seeds. In this module to identify exemplars (Seeds) among data point's responsibility and availability matrices are used consequently the clusters are formed considering all the data points around the exemplars. It operates by concurrently considering all data point as potential exemplars and swapping messages between data points until a good set of exemplars and clusters emerges. In order to identify the correct exemplars the responsibility and availability values are added. The results are dispatched accordingly. The correct clusters with its data points are grouped accordingly.

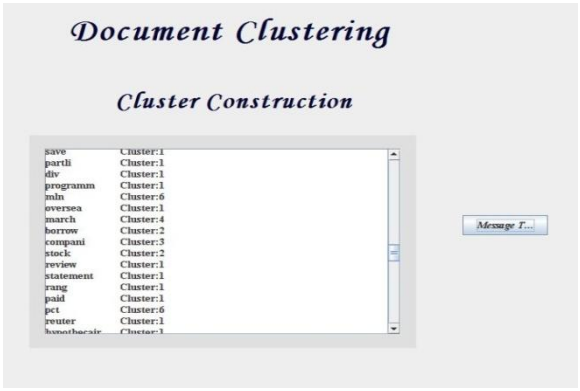

Fig 11: Cluster Construction

3. CONCLUSION AND FUTURE ENHANCEMENT

3.1 Conclusion

In this paper, proposed a similarity measure which is extended from Cosine coefficient using structural information on the basis of Cofeature Set, Independent Feature Set, and Substantial Cofeature Set. Different features at different points of texts can be represented by these three sets. Their physical information increases the clustering results. The novel similarity measurement can be used to compute the

asymmetric similarity directly, which is not limited to the symmetric space. Furthermore, a new clustering algorithm which combines Affinity Propagation with semi supervised learning, namely, the Kernel Affinity Propagation algorithm is proposed. KAP is realistic to full text clustering which extends the application of Affinity Propagation

3.2 Future Enhancement

It makes an important improvement in text clustering tasks. In addition, it is believed that since KAP is based on a complete similarity measurement and on a generic seeds construction strategy, it can be usually applied to other clustering problem domains. In future this algorithm will be still enhanced such that it can be extended for data sampling, fuzzy modeling web mining etc.

4. REFERENCES

[1] Nicholas O. Andrews and Edward A. Fox "Recent Developments in Document Clustering", International Journal of Information Technology and Computer Science(IJITCS) , Vol.2, No.2, pages:1-25, Dec. 2010.

[2] Y.J. Li, C. Luo, and S.M. Chung, "Text Clustering with Feature Selection by Using Statistical Data," IEEE Trans. Knowledge and Data Eng., vol. 20, no. 5, pp. 641-652, May 2008.

[3] C. Buckley and A.F. Lewit, "Optimizations of Inverted Vector Searches," Proc. Ann. ACM SIGIR, pp. 97-110, 1985.

[4] G.Salton and M.J. McGill," Introduction to Modern Information Retrieval". McGraw Hill Book Co., 1983.

[5] G. Salton, "Dynamic Information and Library Processing". Prentice Hall, Inc., 1975.

[6] N. Jardin and C.J. van Rijsbergen, "The Use of Hierarchic Clustering in Information Retrieval," Information Storage and Retrieval, vol. 7, no. 5, pp. 217-240, 1971.

[7] Renchu Guan , Xiaohu Shi , Maurizio Marchese , ChenYang and Yanchun Liang , "Text Clustering with Seeds Affinity Propagation '' , IEEE Transactions on Knowledge and Data Engineering, Vol. 23 , No.4 ,pges.627-637,APRIL,2011.