# An Efficient Text Clustering Approach using Biased Affinity Propagation

Isha Sharma
TIEIT (Bhopal)
Department of CSE

Mahak Motwani
TIEIT (Bhopal)
Department of CSE

## ABSTRACT
Based on an effective clustering algorithm Seeds affinity propagation- in this paper an efficient clustering approach is presented which uses one dimension for the group of the words representing the similar area of interest with that we have also considered the uneven weighting of each dimension depending upon the categorical bias during clustering. After creating the vector the clustering is performed using seeds-affinity clustering technique. Finally to study the performance of the presented algorithm, it is applied to the benchmark data set Reuters-21578 and compared it for F-measure, with k-means algorithm and the original AP (affinity propagation) algorithm the results shows that the presented algorithm outperforms the others by acceptable margin.

### Keywords
Affinity Propagation, Text Mining, Clustering.

## 1. INTRODUCTION
We are facing an ever increasing volume of text documents. The abundant texts flowing over the Internet, huge collections of documents in digital libraries & repositories, & digitized personal information such as blog articles and Emails are piling up quickly every day. These have brought Challenges for the effective and efficient organization of text documents. Clustering in general is an important and useful technique that automatically organizes a collection with a substantial number of data objects into a much smaller number of coherent groups. The searching by clusters reduces the searching time. Cluster-based retrieval is based on the hypothesis that similar documents will match the same information needs In document-based retrieval, an information retrieval (IR) system matches the query against documents in the collection and returns a ranked list of documents to the user. Groups documents into clusters & returns a list of documents based on the clusters that they come from. The task for the retrieval system is to match the query against clusters of documents instead of individual documents, and rank clusters based on their similarity to the query. Any document from a cluster that is ranked higher is considered more likely to be relevant than any document from a cluster ranked lower on the list. Cluster-Based Retrieval system is the clustering algorithm used k-means which required number of clusters before starting which is exactly not possible for the non labeled data and secondly they highly depends upon the initial selection of centroids. To overcome these limitations of the algorithm the affinity propagation clustering algorithm is preferred preferred which searches the set of exemplars on the basis of responsibility and availability of each member.

## 2. AFFINITY PROPAGATION
In k-mean clustering an exemplar learning algorithm that takes as input an initial set of exemplars (often randomly-selected) and then iteratively refines that set while changing the clusters to match the set of exemplars. This drawback overcome by An algorithm (Affinity Propagation) that identifies exemplars among data points and forms clusters of data points around these exemplars. It operates by simultaneously considering all data point as potential exemplars and exchanging messages between data points until a good set of exemplars and clusters emerges. a single iteration involves computing all responsibility messages based on the current availability messages, the input similarities and the input preferences, and then computing all availability messages based on the responsibility messages, which were just updated. The "responsibility" matrix which is presented by r(i, k). Which specify relative measure of suitability of vector $x_k$ as an exemplar for $x_i$.

$$r(i,k) \leftarrow s(i,k) - \max_{k' \neq k}\{a(i,k') + s(i,k')\}$$

The "availability" matrix which is presented by $a(i,k)$. Which specify relative measure of suitability of vector $x_i$for $x_k$ for being the member $x_k$.

$$a(i,k) \leftarrow \min\left(0, r(k,k) + \sum_{i' \notin \{i,k\}} \max\left(0, r(i',k)\right)\right) for\ k$$

$$a(k,k) \leftarrow \sum_{i^1 \neq k}^{\neq k} \max\left(0, r(i',k)\right) for\ k = k$$

## 3. CONSTRUCTION OF SEEDS
For semi-supervised clustering, the main aim is to efficiently cluster a large number of unlabeled objects using only a small number of provided labeled objects. Starting with a few initial labeled objects, the construction of efficient initial "seeds" for our Affinity Propagation clustering algorithm is proposed in [9]. The initial seeds for the AP guarantees the precision cluster formation and avoids the random search which frequently causes imbalance errors. The specific seeds construction method that is named Mean Features Selection [9], can be described as:

Let$N^O, N^F, N^D$ *and* $F^C$represent, respectively, the object number, the feature number, the most significant feature number, and the feature set of cluster $c$ in the labeled set (which are searched by viewing each object in cluster $c$).

Let$F$ is the feature set and $D^F$ is the most significant feature set of seed $c$ (for example, $D^F$ of this manuscript could be all the words (except stop words) in the title, i.e., {text, clustering, seed, Affinity, and Propagation}).

Let $f_k \in F_C$, $f_{k'} \in F_C$their values in cluster $c$ are $n_k$ and $n_{k'}$, the values of being the most significant feature are $n_{DK}(0 \leq n_{DK} \leq n_k)$and $n_{DK'}(0 \leq n_{DK'} \leq n_{k'})$.The seeds construction method is prescribed as

$$n_{k'} \geq \frac{\sum_{k=1}^{N^F} n_k}{N^0}, f_{k'} \in F$$

$$n_{DK'} \geq \frac{\sum_{k=1}^{N^D} n_{DK}}{N^0}, f_{k'} \in DF$$

This method can quickly find out the representative features in labeled objects. The seeds are made up of these features and their values in different clusters. Accordingly, they should be more representative and discriminative than normal objects. In addition, for seeds, their self-similarities are set to $+\infty$ to ensure that the seeds will be chosen as exemplars and help the algorithm to get the exact cluster number.

## 4. PROPOSED ALGORITHM

The text files are converted into the Vector space on the basis of word list provided for each category. The weight is assigned to each dimension of the vector on the basis of bias required. Now the seeds are estimated using the process as described in section 3. Taking the initial seeds a self-similarity matrix is calculated by using the method presented. Finally the number of clusters emerges automatically as the inherent property of affinity propagation clustering.

## 5. EVALUATION MEASURES

To evaluate the performance of clustering, three kinds of measures, F-measure, entropy, and CPU execution time, are used to compare the generated clusters with the set of categories created manually in Reuters. This paper introduce only f-measure The F-measure is a harmonic combination of the precision and recall values used in information retrieval. Due to the higher accuracy of the clusters mapping to the original classes, the larger the F-measure, the better the clustering performance.

P (i,j)=$N_{ij}/N_j$

R (i,j)= $N_{ij}/N_i$

Where Nij is the number of objects of classes i in cluster j, Nj is the number of objects of cluster j, and Ni is the number of objects of class i. The corresponding F-measure F (i, j) is defined as

F(i, j)= $\frac{(2P(i,j)\,R(i,j))}{P(i,,j)+R(i,,j)}$

The global F-measure for the whole clustering result is defined as
F(i,j)=$\sum(Ni/N)$ max $(F(i,j))$

## 6. SIMULATION RESULTS

The performance of the proposed algorithm is evaluated using publicly available Reuters-21578 (Reuters) data set against the manually pre classified labels. The original Reuters data consist of 21,578 documents in 22 files in each file the different documents and their properties are defined by special tags such as "<TOPICS>" and "<DATE>"among others. The text are converted into all capital letters to save it in separate files. For the evaluation of the algorithm the documents are taken as groups of 100, 200, 300, 400 and 500.

**Table 1: showing the results for previous (SAP) algorithm**

| Total Doc. | Precession | Recall | F-Measure |
|---|---|---|---|
| 100 | 0.6987 | 0.4051 | 0.2334 |
| 200 | 0.773 | 0.2548 | 0.1295 |
| 300 | 0.757 | 0.3038 | 0.173 |
| 400 | 0.6976 | 0.3273 | 0.1699 |
| 500 | 0.6571 | 0.2942 | 0.1585 |

**Table 2: showing the results for proposed algorithm**

| Total Doc. | Precession | Recall | F-Measure |
|---|---|---|---|
| 100 | 0.7373 | 0.7962 | 0.7656 |
| 200 | 0.7867 | 0.6789 | 0.7288 |
| 300 | 0.7265 | 0.7451 | 0.7357 |
| 400 | 0.6589 | 0.7508 | 0.7018 |
| 500 | 0.7013 | 0.7511 | 0.7254 |

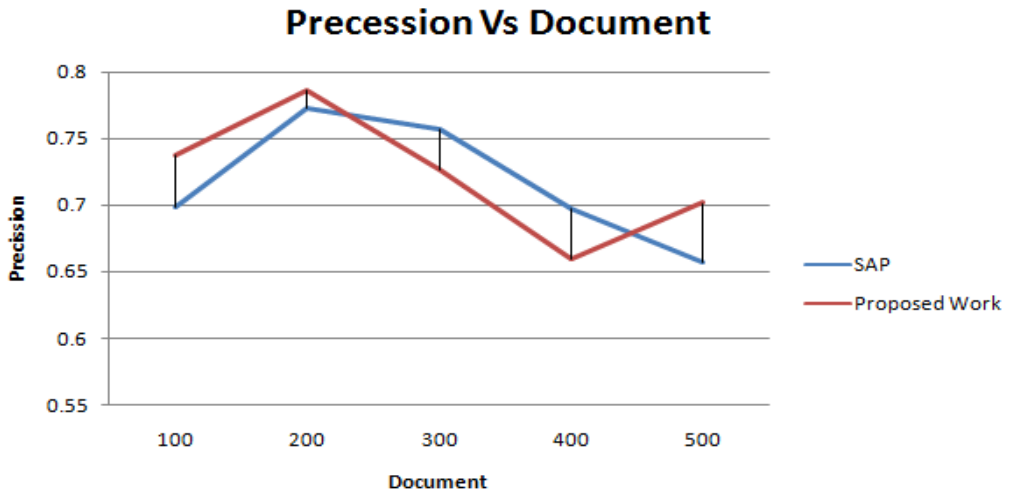


**Figure 1: comparison plot for precision of the clustered documents for different document data size**

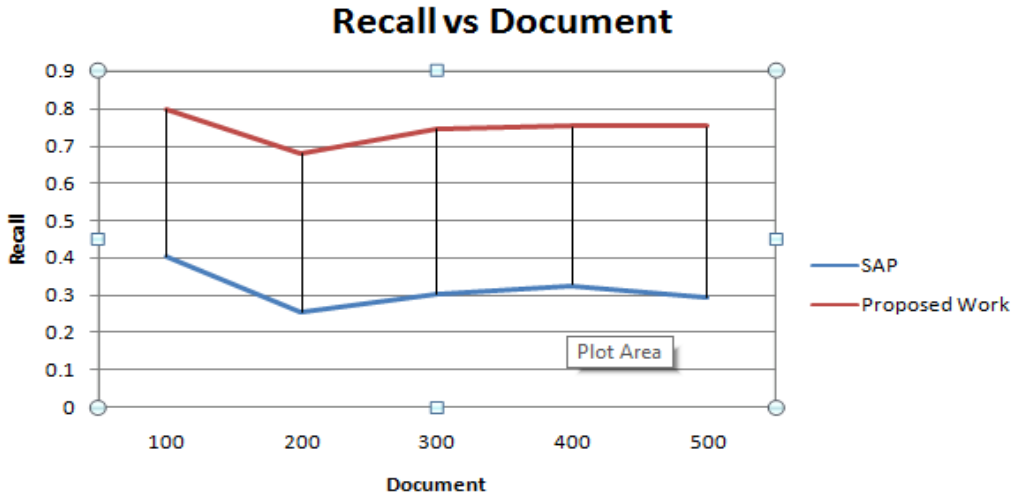


**Figure 2: Comparison plot for recall of the clustered documents for different document data size.**

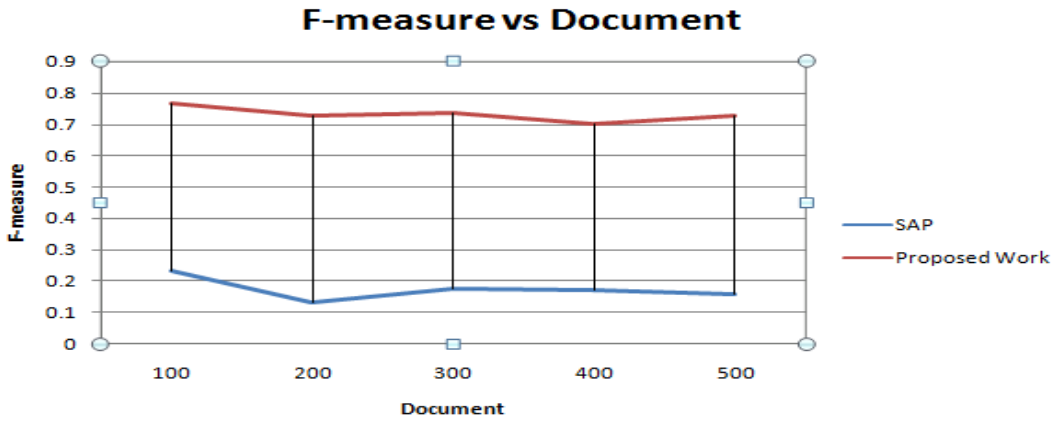


**Figure 3: comparison plot for F-measure of the clustered documents for different document data size.**

## 7. CONCLUSION

In this paper , provide weight which is assigned to each dimension of the vector on the basis of bias required. Which is extended from SAP using categorical information, This algorithm improves the accuracy but also effectively avoids being random initialization and trapped in local minimum The algorithm can also be used for clustering problems in different domains which can be analyzed in future work.

## 8. REFERENCES

[1] Shifei Ding and Hui Li "Quotient Space Granularity Selection Based Affinity Propagation Clustering Algorithm", Journal of Computational Information Systems 10: 6 (2014) 2425–2433

[2] Erik Cambria, Björn Schuller, Yunqing Xia and Catherine Havasi "New Avenues in Opinion Mining and Sentiment Analysis", Knowledge-Based approaches toconcept-level sentiment analysis.

[3] Chen Yang and Renchu Guan "A Feature-Metric-Based Affinity Propagation Technique for Feature Selection in Hyper-spectral Image Classification", Geoscience and Remote Sensing Letters, IEEE,Sept. 2013.

[4] Qingyao Wu, Michael K. Ng and Yunming Ye "Co-Transfer Learning Using Coupled Markov Chains withRestart", IEEE Intelligent Systems, 08 March 2013. IEEE computer Society Digital Library.

[5] Wei Chen, Qichong Tian, Xiaorong Jiang, Zhibo Tang, Caihua Guo, Xinzheng Xu, Hong Zhuand Shifei Ding "Domain Knowledge Blended Affinity Propagation", Appl. Math. Inf. Sci. 7, No. 2, 717-723 (2013).

[6]Vicenc Quera, FrancescS. Beltran, InmarE. Givoni and Ruth Dolado "Determining shoal membership using affnity propagation", Behavioural Brain Research 241 (2013) 38–49.

[7] Stevan Rudinac, Alan Hanjalic and Martha Larson "Generating Visual Summaries of Geographic AreasUsing Community-Contributed Images", IEEE TRANSACTIONS ON MULTIMEDIA, VOL. 15, NO. 4, JUNE 2013.

[8] Teng Li,Bin Cheng, Xinyu Wu and Jun Wu "Low-Rank Affinity Based Local-Driven Multilabel Propagation", Mathematical Problems in EngineeringVolume 2013, Article ID 323481.

[9] Renchu Guan, Xiaohu Shi, Maurizio Marchese, Chen Yang, and Yanchun Liang "Text Clustering with Seeds Affinity Propagation", IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 23, NO. 4, APRIL 2011.