An Efficient Text Clustering Approach using Affinity Propagation with weight modification

Isha Sharma
Department of Computer Science & Engineering
Truba Institute of engineering & Information technology
Bhopal Madhya Pradesh INDIA
ishasharma0701@gmail.com

Prof.mahak motwani
Department of Computer Science & Engineering
Truba Institute of engineering & Information technology
Bhopal Madhya Pradesh INDIA
mahak.motwani@trubainstitute.ac.in

Abstract

Recently the text mining has emerged as one of the most important fields of data mining because of most of the searching in the web is done on the basis of provided text, also the increasing use of social web network uses the text as major component and extracting the effective information directly or indirectly requires an efficient grouping algorithm which should be capable of providing efficient clustering. The most widely used techniques use vector space model to find equivalent vector of the text for clustering. The vector space model represents the text on the form of n-tuples numeric array (vector) where each dimension represents the unique word and the value is the weight of that word on the basis of term frequency-inverse document frequency (tf-idf), the problem of the technique is that the unique words count in any document may be very large which will create the similarly long vectors whose processing will require large memory with processing power secondly analysis may be required a bias categorical grouping which not addressed in the above technique. Hence 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, Entropy and Execution time 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

The text clustering is the branch of data mining where the data samples are grouped on the basis of text they contain. The field is gaining interest mainly for information retrieval which can then be used to store the data in the form of abstract information (vector) and used for quick searching. The searching by clusters reduces the searching time because instead of matching with each individual samples only the cluster representatives (i.e. centroids) are used and after that the matching is performed to the members (exemplar's) of the closest cluster representative. The initial requirement of the Cluster-Based Retrieval system is the clustering algorithm one of the most commonly used k-means which is a heuristic search algorithm. The problem with such algorithms are first that they need the 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 which searches the set of exemplars on the basis of iteratively calling of responsibility and availability of each member. To further improve quality and convergence time of the affinity propagation seeds propagation is proposed in which the affine propagation is started with biased initial preferences (seeds) for all exemplars on the basis of initial similarity. In this paper an efficient vectoring technique with seeds affinity propagation is presented. The rest of this paper is organized as follows. Section II provides the related work to the study. In Section III, some standard techniques are discussed followed by the working of the Seeds Affinity Propagation Algorithm in next section V presents the proposed method Section VI shows the simulation results in various different scenarios. Finally, Section VII presents our conclusions and further research directions.

2. Text Clustering Review

In the text data mining clustering is used for grouping the similar documents as a technique to reduce the searching time in information retrieval systems. Initially the searching was implemented by performing a serial scan (in which the matching of a query against each individual document of a database) and thus the number of query document matching calculations could take too much time. however a clustered search, in which a query is matched with database which is arranged in clusters, may achieve better matching since the formation of one representative vector (centroid) of a group could show some hidden relations which exist between the documents in the database.

For all cases the clustering techniques can be broadly divided into two categories.

1. Non-Heuristic
2. Heuristic

Non-Hierarchic Clustering: In this technique grouping or partition describes a classification in which group elements or members are estimated using some well and pre-defined rules and there are no hierarchic relationships is considered between the elements.

Heuristic Clustering: The problem with Non-Heuristic that it cannot be applied where the rules are not pre-defined. In such cases the partitioning rule need to be estimated by the analysis of the dataset.

Several non-heuristic and heuristic procedures have been used for document clustering, although in this paper our main emphasis is on the use of heuristic clustering procedures that can achieve a grouping of the documents at lower computational cost, hence the review presents only the literatures related to heuristic approach. An improved AP clustering algorithm based on the quotient space granularity selection is proposed by Shifei Ding et al [1], they introduced the quotient space concept to the AP clustering analysis, which can find an optimal granularity from all possible granularities. Affinity Propagation using Feature Metric is proposed in [3], it finds a transformation matrix of the feature space using equivalence constraints. Qingyao Wu et al [4] addressed the machine learning problem called transfer learning. They focus on how to use labeled data of different feature spaces to enhance the classification of different learning spaces simultaneously. For that they used the concept of coupled Markov chain with restart, because the transition probabilities in the coupled Markov chain can be constructed by using the intra-relationships based on affinity metric among candidates in the same space, and the interrelationships based on co-occurrence information among candidates from different spaces. The algorithm computes ranking of labels to indicate the importance of a set of labels to an instance by propagating the ranking score of labeled instances via the coupled Markov chain with restart .Blended Affinity Propagation for domain knowledge in Case of weaker separability feature s pace is proposed by Wei Chen et al [5]. Their algorithm combines the domain knowledge function and the similarity measure of the AP algorithm, the algorithm makes iterating to obtain the clustering result. Affinity propagation for determining shoal membership in [6], the paper presents an extension AP of, denoted by STAP, that can be applied to shoals that fusion and fission across time. STAP incorporates into AP as of temporal constraint that takes cluster dynamics into account, encouraging partitions obtained at successive time steps to be consistent with each other. The low-rank affinity based local-driven algorithm to robustly propagate the multi labels from training images to test images is proposed by Teng Li et al [8], The multitask low-rank affinity, which jointly seeks the sparsity-consistent low-rank affinities from multiple feature matrices, is applied to compute the edge weights between graph vertices. A semi supervised text clustering algorithm, called Seeds Affinity Propagation (SAP) with seed construction method to improve the semi supervised clustering process is proposed in [9].

3. Affinity Propagation Clustering

Affinity propagation (AP) is a clustering algorithm utilizes the concept of "message passing" amongst the data vectors. Unlike the other clustering algorithms such as k-means or c-means, AP clustering finds the vectors that are most likely belongs to similar group called exemplar's and the number of groups cannot be estimated or provided before running the algorithm.

The formal definition of the affinity propagation can be described as:

Let x_1 through x_n be a set of data vectors.

Let s be a function that quantifies the similarity between any two points, like $S(x_i, x_k)$ which represents the similarity between $x_i$and x_k. The similarity can be considered as inversely proportional to distance, hence the higher similarity represents the closeness of the vectors.

The algorithm evolves by iteratively and alternatively two message passing steps called "responsibility" and "availability' matrices:

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.

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.

Initially both matrices are stored with zeroes, the algorithm then performs the following updates iteratively:

First, responsibility updates are calculated by:

$r(i,k) \leftarrow s(i,k)\text{-}\max_{k \neq k'}\{a(i,k') + s(i,k')\}$ [1]

Then, availability is updated per

$a(i,k) \leftarrow \min\left(0, r(k,k) + \sum_{i' \notin (i,k)} max\left(0, r(i',k)\right)\right)$ for k≠ k [2]

$a(k,k) \leftarrow \sum_{i \neq i} max\left(0, r(i',k)\right)$ for k=k [3]

3.1 Seeds Construction

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 guarantee the precision cluster formation and avoid 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^Crepresent, 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

Firstly 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.

$$v_{i,new} = v_i * w_i$$

$where\ v_{i,new}\ is\ the\ i-th\ dimention\ of\ the\ vector\ after\ biasing,$

$v_i\ is\ the\ i-th\ dimention\ of\ the\ original\ vector,$

$w_i\ is\ the\ weight\ of\ the\ corresponding\ dimention$

Now the seeds are estimated using the process as described in section 3.1.

Taking the initial seeds a self-similarity matrix is calculated by using the method presented in section 3.

Finally the number of clusters emerges automatically as the inherent property of affinity propagation clustering.

5. 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 pre classification labels are eliminated before the clustering processes, and is used to evaluate the clustering accuracy of each clustering algorithm at the end of the execution.

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.

Before using it for clustering some preprocessing operations are performed which separates the main text information then after stop words removal, special character removal andword stemming is performed. Finally

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	Entropy	Time
100	0.6987	0.4051	0.2334	0.5265	0.1909
200	0.773	0.2548	0.1295	0.3015	0.4832
300	0.757	0.3038	0.173	0.3442	0.9865
400	0.6976	0.3273	0.1699	0.3657	1.7568
500	0.6571	0.2942	0.1585	0.3213	2.6458

Table 2: showing the results for proposed algorithm

Total Doc.	Precession	Recall	F-Measure	Entropy	Time
100	0.7373	0.7962	0.7656	0.6161	0.1921
200	0.7867	0.6789	0.7288	0.2012	0.4898
300	0.7265	0.7451	0.7357	0.2354	0.9818
400	0.6589	0.7508	0.7018	0.3241	1.7333
500	0.7013	0.7511	0.7254	0.1768	2.6301

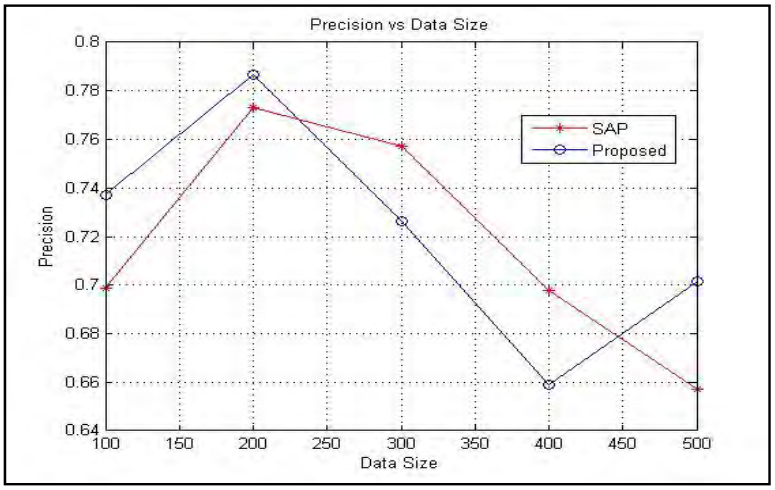

Figure 1: comparison plot for precision of the clustered documents for different data size.

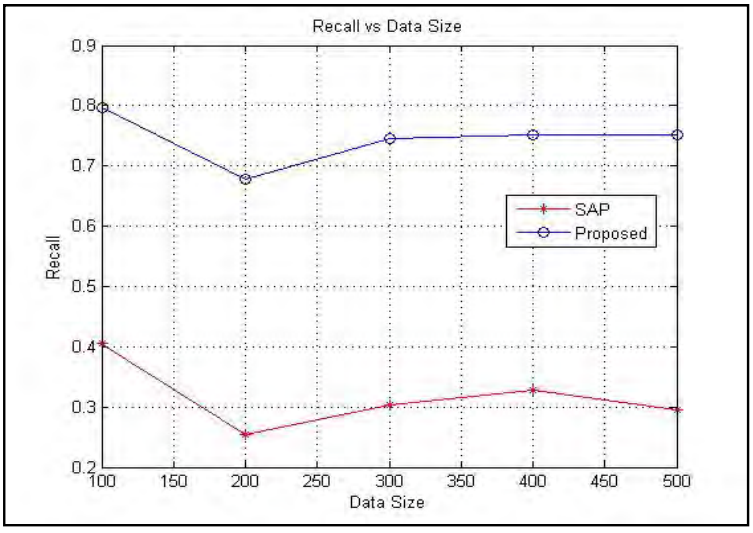

Figure 2: comparison plot for recall of the clustered documents for different data size.

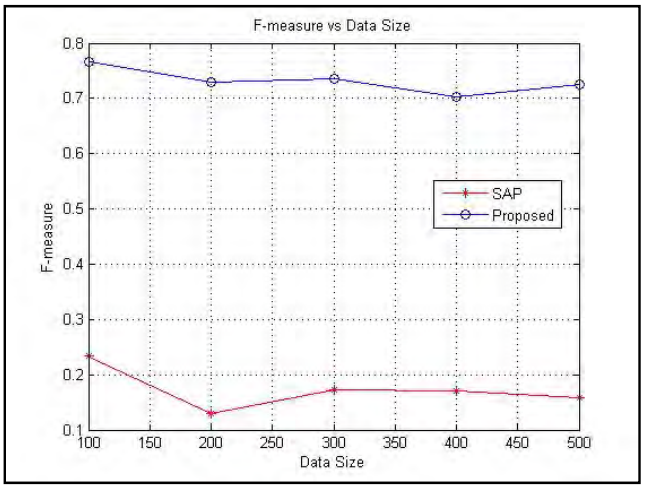

Figure 3: comparison plot for F-measure of the clustered documents for different data size.

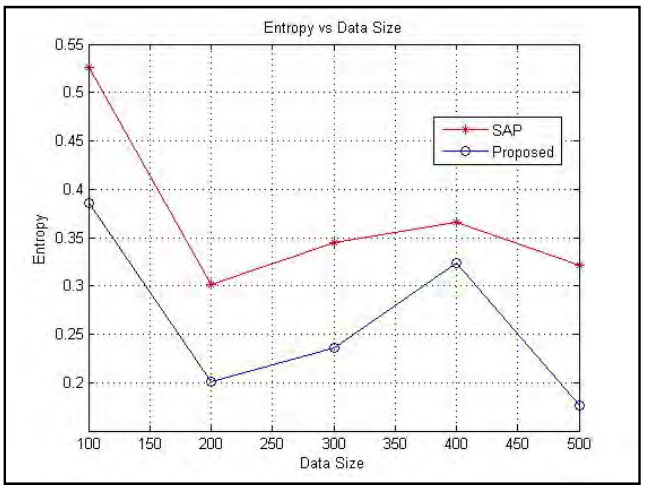

Figure 4: comparison plot for Entropy of the clustered documents for different data size.

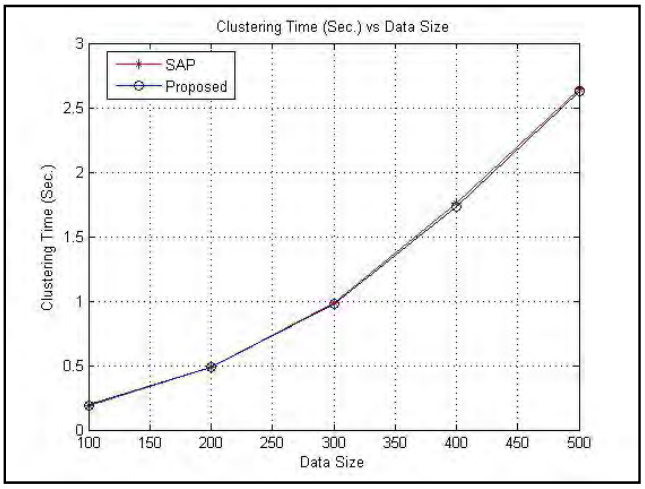

Figure 5: comparison plot for clustering time for different data size.

6. Conclusion and Future Scope

This paper proposes a biased affinity based text clustering algorithm which is extended from SAP using categorical information as the basis of biasing dataset vectors dimension, and reducing the feature vector size by taking a group of words belong to similar impression based on categorical evaluation presently these sets are limited to five. These information improves the clustering results. The simulation results shows that the proposed algorithm not only improves the overall performance (refer to table 1 and 2) without increasing the computing complexity. The seeding effectively reduces the chances of poor performance of AP leads by random initiation. The algorithm can also be used for clustering problems in different domains which can be analyzed in future work.

Acknowledgment

Our thanks to the experts who have contributed towards development of the template.

References

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

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

[3] Qingyao Wu, Michael K. Ng and Yunming Ye, "Co-Transfer Learning Using Coupled Markov Chains with Restart", IEEE Intelligent Systems. IEEE computer Society Digita l Library, 08 March 2013.

[4] 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

[5] Shifei Ding and Hui Li, "Quotient Space Granularity Selection Based Affinity Propagation Clustering Algorithm", Journal of Computational Information Systems 12425–2433 (2014).

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

[7] Teng Li,Bin Cheng, Xinyu Wu and Jun Wu , "Low-Rank Affinity Based Local-Driven Multi label Propagation", Mathematical Problems in Engineering Volume 2013, Article ID 323481. 2013.

[8] Vicenc Quera, Francesc S. Beltran, Inmar E. Givoni and Ruth Dolado, "Determining shoal member ship using affinity propagation", Behavioural Brain Research 241 38–49, (2013) .

[9] 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).

AUTHORS PROFILE

Isha sharma received B.E Degree in Information technology from Rajiv Gandhi University, and currently Pursuing M.tech in the field of Computer Science from RGPV, Bhopal.