IOSR Journal of Electronics and Communication Engineering (IOSR-JECE)
e-ISSN: 2278-2834,p-ISSN: 2278-8735
PP 72-79
www.iosrjournals.org

A New Fangled Seed Construction Algorithm for Text Clustering

Ms.Shamna.B
Computer Science and Engineering Department, Muslim Association College Of Engineering ,Venjaramoodu

***Abstract**: Clustering is a division of data into groups of similar objects, each group called cluster, consists of objects that are similar between themselves and dissimilar to objects of other groups. In other words, the goal of a good document clustering scheme is to minimize intra-cluster distances between documents ,while maximizing inter-cluster distances. A distance measure(or, dually, similarity measure)thus lies at the heart of document clustering. Text Clustering is a category in clustering where the information which is logically similar is combined together. In the previous approaches K-means algorithm was used. Fore coming is the Affinity Propagation algorithm. This is prevailing as an important algorithm in novel semi supervised text clustering. Affinity Propagation algorithm has been enhanced by merging the concept of constructing seeds and called as Seeds Affinity Propagation. In this algorithm Triset computation is used. In proposed approach Jaccard Based Similarity index is used where the clustering quality is increased on adopting this index. Result shows that using our proposed algorithm yields higher F-measure when compared with K-Means and original AP algorithm ,the execution time is also faster moreover provides robustness.*

***Keywords:** Seed propagation, content clustering, cofeature set, unilateral feature set, major cofeature set, Jaccard index, seed creation.*

I. Introduction

Text withdrawal, sometimes alternately referred to as text data mining, roughly equivalent to text analytics, refers to the process of deriving high-quality information from text. High-class information is typically derived through the devising of patterns and trends through means such as statistical pattern learning. Text mining usually involves the process of structuring the input text (usually parsing, along with the addition of some derived linguistic features and the removal of others, and subsequent insertion into a database), deriving patterns within the structured data, and finally evaluation and interpretation of the output. 'High quality' in text mining usually refers combination of relevance, novelty, and interestingness. Typical text mining tasks include text categorization, text clustering, concept/entity extraction, production of coarse taxonomies, sentiment analysis, document summarization, and entity relation modeling.

Document clustering (also referred to as Text clustering) is closely associated to the concept of data clustering. File clustering is a more specific technique for unsupervised document organization, automatic topic extraction and fast information retrieval or filtering. Document retrieval based on Cluster-Based Analysis, is anticipated to improve both fitness and expediency of the document clustering and salvage systems [1], [2], [3], [4]. Clustering is used to arrange a large amount of documents into meaningful clusters. This process helps the browsing of netting information and organizes the results returned by a search engine [5], [6].Time-honored approaches for clustering data are based on metric similarities.

The clustering performance depends on the similarity measure and message updating frequency. For its simplicity, general applicability, and good performance, AP has already been used in text clustering. Wang et al. combined AP with a similar policy for e-learning property clustering [7]. Ma et al developed an incremental method for document clustering [20]. Modern approaches, like Affinity Propagation (AP) algorithm [8], are used for non metric similarities. Affinity Propagation is a resultant of an application of the max-sum algorithm on the basis of message passing between data points [8]. Affinity propagation takes as input a gathering of real-valued similarity between data points, and outputs the clustered data by identifying delegate example called pattern for each data point. As an alternative of using the inventive affinity propagation directly, e-learning adapted it in Map Reduce framework to make it applicable to large-scale data [7]. However, they used AP only as an unsubstantiated algorithm and did not consider any structural information derived from the specific documents.

In many realistic applications, completely unsupervised learning is lacking pertinent information. On the other hand, supervised learning wants a preliminary large number of information, which requires luxurious human attempt and instance[26]. Therefore, in recent years, semi supervised learning has captured a great deal of attentions. A machine learning paradigm in which the model is constructed using both labeled and unlabeled data is a semi supervised learning [9],[10],[13][14]. For training a little amount of labeled data and a large amount of unlabeled data a new approach is used [10]. We present a new clustering algorithm by extending Seed Affinity Propagation with 1) a narrative asymmetric similarity measurement that captures the structural information of texts, and 2) a semi supervised erudition approach, that make use of the awareness from a few labeled objects versus a large number of unlabeled ones [22]. In information salvage, there are several dimensions of similarities are used. The simplest of all resemblance measures, namely, easy matching coefficient, is including the number of common expressions in two documents.

S(A,B)=│A∩B│, where │D│ gives the size of a set D. This coefficient does not take into description the sizes of A and B . A more dominant coefficient called Cosine coefficient takes this information into description [11]:

$$S(A,B) = \frac{|A \cap B|}{|A|^{1/2}|B|^{1/2}} \qquad (1)$$

Two measures are symmetric. In this paper, we recommend an asymmetric similarity measurement using Jaccard similarity coefficient for two different credentials, which is different from the conventional symmetric measurements [11], [20]. An asymmetric similarity measurement called Tri-Set method is thus proposed based on three characteristic sets. Lastly, we present and examine the description of precise primary values for the clustering algorithm, that we named Seeds, to bootstrap the initial phases of the new clustering algorithm. We thus propose a narrative semi supervised clustering algorithm: Seeds Affinity Propagation (SAP). To observe the efficacy of the proposed method, we have applied it to the yardstick data set Reuters-21578. In order to examine the performance of the new algorithm we have performed a detail evaluation with four clustering methods on the identical data set, namely,

1. k-means approach [21].
2. The original Affinity Propagation algorithm
3. Modified Affinity Propagation technique, which combines AP with the new-fangled similarity measurement(AP(Tri-Set)); and
4. a customized Affinity Propagation method which combines AP with the new seed creation semi supervised technique (SAP(CC)).

The rest of the editorial is prearranged as follows: we start in Section 2 with a brief assessment of the Affinity Propagation clustering approach and associated work.

II. Related Work

SAP was planned as a new-fangled and dominant technique for pattern learning. In the subsequent sections , we specify in concise the geometric model of the AP approach. At beginning, AP takes in input a gathering of real-valued similarities between data points. For an N data point's data set, ti and tj are two objects in it. The similarity s(i, j) indicates how well tj is suited to be the pattern for ti. For instance, it can be initialized as

$$\mathrm{S(i,j)} = \frac{\sum_{\mathrm{i,j=1,i\neq j}}^{\mathrm{N}} \mathrm{S(i,j)}}{\mathrm{N*(N-1)}}, 1 \leq 1 \leq \mathrm{N} \qquad (2)$$

where N is the number of documents taken.

The AP approach calculates two types of messages exchanged between data points [8]. The first one is called responsibility r(i, j) is sent from data point i to contender pattern point j and it reflects the accumulated confirmation for how well-matched point j is to provide as the pattern for point i. The second message is called availability a(i,j) is sent from contender exemplar point j to point i and it reflects the accumulated evidence for how appropriate it would be for point i to choose point j as its pattern. At the beginning, the availabilities are initialized to zero:

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

(3)

$$\mathrm{a(i,j)} = \begin{cases} \min\{0, \mathrm{r(j,j)} + \sum_{\mathrm{i' \neq i,j}} \max\{0, \mathrm{r(i',j)}\}\}, \mathrm{i} \neq \mathrm{j} \\ \sum_{\mathrm{i' \neq i}} \max\{0, \mathrm{r(i',j)}\}, \mathrm{i} = \mathrm{j} \end{cases}$$

(4)

III. Affinity Propagation Algorithm

We propose a narrative method called "Seeds Affinity Propagation" based on AP method. The major new features of the new algorithm are: Tri-Set calculation, similarity calculation, seeds creation, and messages transmission. I start the presentation of the algorithm by explaining the basic similarity measurement used in our approach, i.e., three new feature sets, named by Cofeature Set (CFS),Unilateral Feature Set (UFS), and Significant Cofeature Set(SCS) [24]. The structural information of the text documents is included into the new resemblance dimensions. Affinity Propagation was used to solve various clustering problems such as image processing, detecting genes and individual preferences predictions [12][19].Then, I present how we extend the original AP approach with semi supervised learning policy[22].

3.1. Measuring Similarity

Resemblance measurement has got an important responsibility in Affinity Propagation clustering. In order to give specific and effectual similarity measurement for our scrupulous area, i.e., text document, we set up the subsequent feature sets: the Co feature Set, the Unilateral Feature Set, and the Significant Co feature Set. We first specify the computations of the new features to describe these sets . In our approach, each term in text is deemed as a feature and each document is deemed as a vector [1],[13]. Though, all the features and vectors are not calculated at the same time, but one at a time.

Consider T(t1; t2; . . . tn) be a set of texts. Suppose that ti and tj are two items in T. Let j i represent the xth and yth feature of ti and tj, correspondingly. Consider now, the set MFj composed of the "most significant" features of tj. "Most significant" means features that are capable of representing central aspects of the document. These "most significant" features could be key phrases or tag related with each document when they are available.

Cofeature Set can be explained as:

.

Let ti and tj be two items in a data set. Suppose that some features of ti, also belong to tj. Thus, we can create a new two-row division comprising of these features and their values in tj. We characterize it as the Cofeature Set between ti and tj and is represented as:

<fx, nx> £ CFS(i,j) if f £ F where F is the feature set.

Unilateral Feature Set can be explained as:

Some features of ti do not fit in to tj. As a result, we can create a new two-row division. We characterize it as the Unilateral Feature Set between ti and tj and is represented as:

<fy, ny> £ UFS(i,j) if f £ F where F is the feature set.

Significant Cofeature Set can be explained as:

Suppose that some features of ti , also belong to the most significant features of tj . As a result, we can create a new two-row division consisting of these features and their values as the most significant features in tj.

<fz, nz> £ SCS(i,j) if f £ F where F is the feature set.

riset Similarity is calculated as follows:

$$S(i,j) = \alpha \sum_{x=1}^{CFS} nx + \beta \sum_{y=1}^{SCS} ny - \gamma \sum_{z=1}^{UFS} nz \qquad (5)$$

3.2 Creating Seeds

In semi supervised clustering, the main intention is to competently cluster a large amount of unlabeled items starting from a comparatively small number of first labeled items [14]. Given a few first labeled items, construct well-organized initial "seeds" for Affinity Propagation clustering algorithm. Let FC is the feature set and MF is the most significant feature set and c be a seed in the cluster formed. Let fk £ FC and fk'£ FC.

The seeds' construction method is given as

1. nk'$\geq \frac{\sum_{k=1}^{NF} nk}{N0}$, fk'£ FC;
2. nMk'$\geq \frac{\sum_{k=1}^{ND} nDk}{N0}$,fk'£ MF; where F is the feature set and MF is the most significant feature set.

3.3 System Architecture

The architecture diagram of SAP algorithm used for document clustering is diagrammatically represented as follows as Fig 1. The architecture is fully stressed on the type of data set used for document clustering.

3.4 Seed Affinity Propagation Algorithm

Steps included in this algorithm are Initialization, Seeds construction, Tri-Set computation, Self-Similarity computation, Initialize messages, Message matrix computation, Exemplar selection, Updating the messages.

Here the Datasets are collected, preprocessed and Intrinsic words are captured. Similarity is computed for each words in the documents. Tri-set computation is performed on the documents. Seeds Affinity Propagation Algorithm is used with the options for similarity Computation ,Seed construction and Message computation (Uses AP Partially)and identifies exemplars among data points and forms clusters of data points around the exemplars.

The Jaccard index, also known as the Jaccard similarity coefficient is a statistic used for comparing the similarity of sample sets.

Steps:

3.4.1. Preprocessing

Data are collected from reuter dataset. The data set will be enclosed with HTML tag and syntaxes The steps done in this process are:

Striping: Eliminate the tags in the document.

Stopwords Removal: short function words, such as the, is, at, which, on etc are removed.

Words Stemming : Eliminating the Prepositions articles, etc.

3.4.2. Extract Significant Features: Key terms extraction is a basic step for various tasks of natural language processing. In these modules the following are carried out. These “most significant” features could be key phrases or tags associated with each document when available. This is the basic step for seeds creation.

3.4.3.Seeds creation: The seeds are constructed using the Tri set computation approach.

3.4.4. Tri Set calculation : Computing UFS (Unilateral Feature Set): features present in one document but not present in other document. Computing CFS (Co Feature Set): features present in both documents. Computing SCS (Significant co feature Set): some features present in one document belong to the most significant feature in other document called as SCS.

3.4.5 . Similarity calculation: Compute similarities using Eqn (5).

3.4.6. Message matrix Computation: In this module the AP algorithm is used partially for Prediction of cluster . Exemplar and Data Points are the two representations of documents. To evaluate the correct exemplars and data points certain matrix are computed. For data points i and k Responsibility r(i,k) is computed and Availability a(i,k) is evaluated as follows:

a(i,k)=0, $r(i,k) = s(i,k) - max\{s(i,k')\}$ (6)

The availability and responsibility messages are passed between data points [18].

3.4.7. Adding the two matrices and select the pattern as r(i,k)+a(i,k).

3.4.8.Cluster Identification :The main idea 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. 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. Repeat the steps 5,6,7 and 8 until best matching clusters are formed.

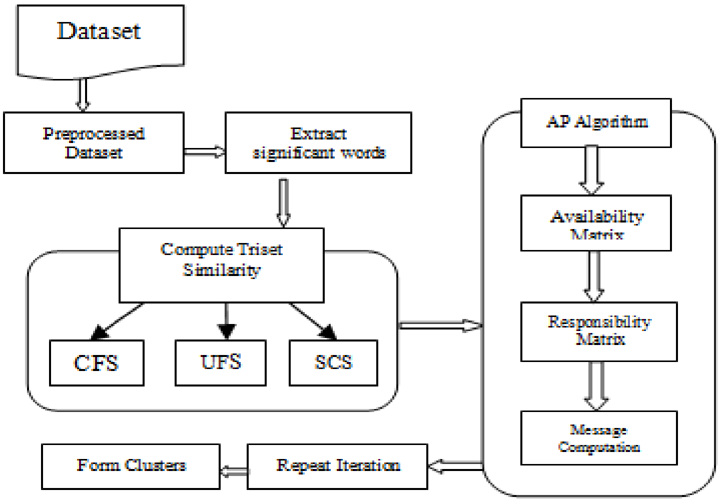

Figure . 1 System Architecture for Text Clustering

IV. Experimental Results

To determine the performance of SAP algorithm, we used text data Reuters-21578 [15],[16],[17]. The experimental results shows that SAP using Jaccard for finding best clusters obtained better accuracy. For analyzing the accuracy of obtained results, we acquired three measures namely F-measure, Entropy and CPU execution time.

4.1Data Sets

We used different data sets extracted from Reuter-21578 [16]. Sample data sets denoted by CACM, MED, EXC, PEO, TOP. These data sets belong to different classes. Experiment using Sap used top ten classes namely “acq”, ”corn”, ”crude”, ”grain interest”, ”trade”, ”money-fx”, ”wheat”, ”ship” extracted from Reuters data set [23]. Table 1 summarizes the characteristics of different data sets.

4.2Evaluation Methods

Three evaluation measures are used to compare the newly developed algorithm with previous approaches..They are F-Measure, Entropy, CPU execution time. We can use the terms precision P(i,k) and recall R(i,k) for defining accuracy of clusters. The data points used are i and k.

$$P(i,k) = {}^{Nik}/_{Nk} \tag{7}$$

$$R(i,k) = {}^{Nik}/_{Ni} \tag{8}$$

F-Measure is calculated as follows:

$$F = \sum_i \frac{Ni}{N} \max(F(i,k)) \tag{9}$$

Entropy is calculated as follows:

$$Ek = -\sum_i pik \log(pik)$$

(10)

The previous approaches for clustering early mentioned were applied to Reuter data set and obtained different strategies. Table 2 listed the strategies used for different clustering algorithms.

Table 1 Summary Of Data Sets

DataSet	Num of documens	Num of classes	Min class size	Max class size	Num of unique terms	Avg doc length
CACM	842	43	11	51	3225	59
MED	287	9	26	39	4255	77
EXC	334	7	28	97	3258	67
PEO	694	15	11	143	5046	102
TOP	2279	7	23	750	10719	113

Conventional approaches for clustering used different similarity calculations for finding clusters.
K-Means algorithm [21] adopted the following equation.

$$S(X,Y) = \frac{\sum_{i=1}^{n} xiyi}{\sum_{i=1}^{n} \sqrt[2]{xi*xi} \sum_{i=1}^{n} \sqrt[2]{yi*yi}} \tag{11}$$

AP(CC) and SAP(CC) used cosine coefficients as the similarity between two documents.
SAP(Jaccard) and AP(TriSet) used eqn (5) for finding similarity between two documents. Self-similarity is computed as follows:

$$s(p,p) = \frac{\sum_{I,J=1,I\neq J}^{N} s(i,j)}{N\times(N-1)} - \psi \ \max\{S(i,j)\} \tag{12}$$

Where ψ is an adjustive factor. The parameter ψ is correlated with data set size .Frey et al pointed out that the higher values of this parameter caused to find more clusters using AP algorithm.

Table 2 Different Strategies For Different Clustering Algorithms

	K-Means	AP(CC)	AP(TriSet)	SAP(CC)	SAP(Jaccard)
TriSet Similarity	Not present	Not present	present	Not present	Present
SemiSupervision	Not present	Not present	Not present	present	Present

The experimental evaluation measures used to improve clustering accuracy like F-measure and Entropy can be diagrammatically represented as follows for the different clustering algorithms [25]. The bar chart diagrams for the evaluation measures are given below. Fig 2 denotes F-Measure and Fig 3 denotes Entropy.

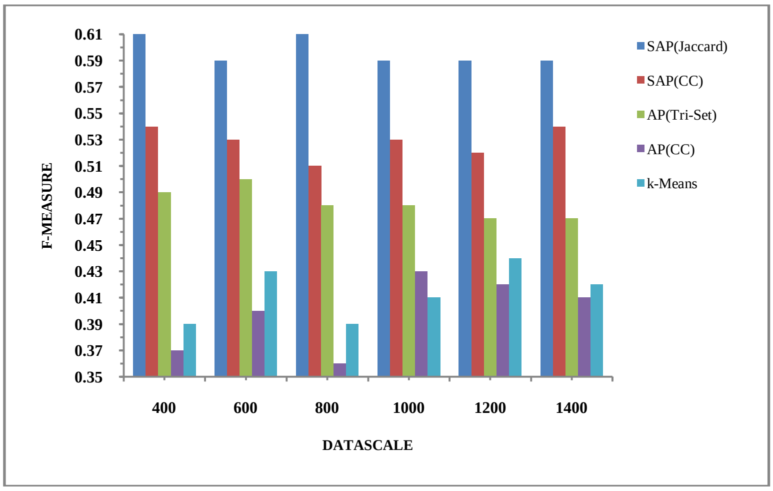

Fig. 2. F-measure comparison of clustering algorithms.

In Fig 2 the mean F-measure value of AP(CC) is close to that of K-means[25], while the mean F-measure values of AP(TriSet), SAP(CC) and SAP(Jaccard) are 16.8,27.6 and 43.9 percent higher than that of K-means. Table 3 listed as follows.

Table 3 Mean F-Measure Values

	Mean F-measure
SAP(Jaccard)	.599
SAP(CC)	.531
AP(TriSet)	.486
AP(CC)	.403
k-Means	.416

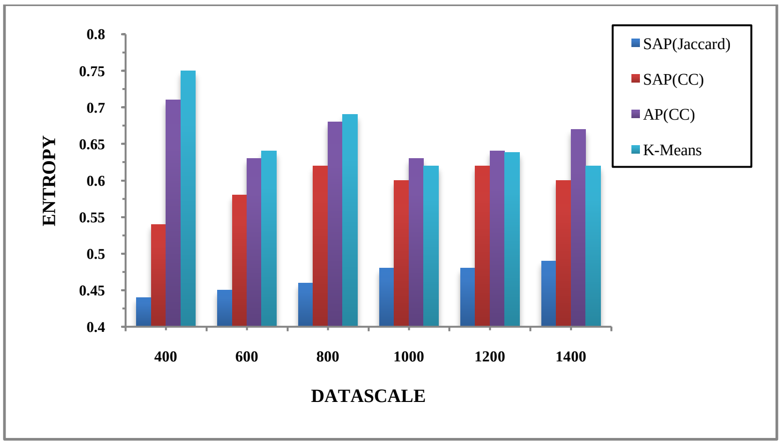

Fig.3. Entropy comparison of clustering algorithms.

In Fig 3, SAP has the lowest entropy percent compared to other algorithms. Table 4 summarizes entropy values of different clustering algorithms.

Table 4 Average Entropy Values

	Mean Entropy
SAP(Jaccard)	.472
SAP(CC)	.592
AP(Triset)	.610
AP(CC)	.657
K-Means	.658

Apart from the two comparison measures, F-measure and entropy another important comparison measurement is robustness. It can be used to improve the clustering accuracy of SAP clustering algorithm.

The Figure 4 shows the allocation of different documents belong to the Reuter data set. The number of seeds percentage varies with the number of documents present in the data set.

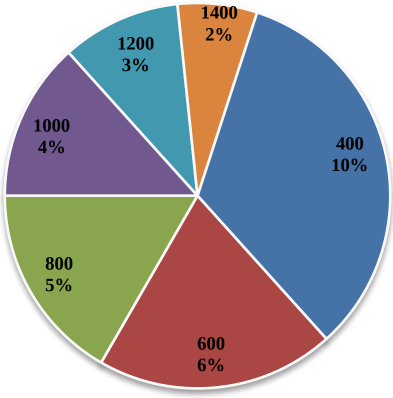

Figure. 4 Relationship between documents and seeds;

Based on all the experimental results performed earlier an original analysis can be given: K-Means is based on the objective function and searches for minimum number of clusters [25]. But K-Means can be easily ensnared into local minimum. On other hand, SAP calculates similarity using Tri-Set method in which the structural information was taken into account. Thus SAP using Jaccard coefficients can be effectively work out the problem of finding clusters.

V. Conclusions And Future Scope

In this paper, a new similarity measurement algorithm which is extended from the use of cosine coefficients are used to cluster the documents. The new clustering algorithm named Seed Affinity Propagation (SAP) is based on the structural information of documents. Three feature sets called UFS, CFS, and SCS were applied to the SAP algorithm and improved the clustering accuracy and efficiency.SAP is more robust than all conventional approaches of document clustering. Due to the merits of SAP , this can be applied to different application areas like Segmentation, Newsgroups etc. As a future work we have to explore more algorithms that are more accurate , efficient and robust than SAP using Jaccard coefficients.

References

[1]. 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.

[2]. G. Salton, Dynamic Information and Library Processing. Prentice-Hall, Inc., 1975.

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

[4]. E.M. Voorhees, “The Efficiency of Inverted Index and Cluster Searches,” Proc. Ann. Int’l ACM SIGIR, pp. 164-174, 1986.

[5]. 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.

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

[7]. W.H. Wang, H.W. Zhang, F. Wu, and Y.T. Zhuang, "Large Scale of E-Learning Resources Clustering with Parallel Affinity Propaga- tion," Proc. Int'l Conf. Hybrid Learning 2008 (ICHL '08), pp.1-10,Aug. 2008.

[8]. B.J. Frey and D. Dueck, "Clustering by Passing Messages between Data Points," Science, vol. 315, no. 5814, pp. 972-976, Feb. 2007.

[9]. B.J. Frey and D. Dueck, "Non-Metric Affinity Propagation for Un- Supervised Image Categorization," Proc. 11th IEEE Int'l Conf. Computer Vision (ICCV '07), pp. 1-8, Oct. 2007.

[10]. T Z.H. Zhou and M. Li, "Semi-Supervised Regression with Co- Training Style Algorithms," IEEE Trans. Knowledge and Data Eng., vol. 19, no. 11, pp. 1479-1493, Aug. 2007.

[11]. J. MacQUEEN, "Some Methods for Classification and Analysis of Multivariate Observations," Proc. Fifth Berkeley Symp. Math. Statistics and Probability, pp. 281-297, 1967.

[12]. Y. Jiang and A. Tuzhilin, "Dynamic Micro Targeting: Fitness- Based Approach to Predicting Individual Preferences," Proc. Seventh IEEE Int'l Conf. Data Mining (ICDM '07), pp. 173-182, Oct. 2007

[13]. F. Sebastiani, "Machine Learning in Automated Text Categoriza- tion," ACM Computing Surveys, vol. 34, pp. 1-47, 2002.

[14]. Z.H. Zhou and M. Li, "Tri-Training: Exploiting Unlabeled Data Using Three Classifiers," IEEE Trans. Knowledge and Data Eng., vol. 17, no. 11, pp. 1529-1541, Nov. 2005.

[15]. Z.H. Zhou and M. Li, "Distributional Features for Text Categor- ization," IEEE Trans. Knowledge and Data Eng., vol. 21, no. 3, pp. 428-442, Mar. 2009.

[16]. D.D. Lewis, "Reuters-21578 Text Categorization Test Collection," htt p: //www.da viddle w is.com/re s ources/te s tco llect ions/ reuters21578, May 2004.

[17]. A. Estabrooks, T. Jo, and N. Japkowicz, "A Multiple Resampling Method for Learning from Imbalanced Data Sets," Computational Intelligence, vol. 2, no. 1, pp. 18-36, 2004.

[18]. B.J. Frey and D. Dueck, "Response to Comment on 'Clustering by Passing Messages Between Data Points,'" Science, vol. 319, no. 5864, p. 726d, Feb. 2008.

[19]. L. Michele, Sumedha, and W. Martin, "Clustering by Soft- Constraint Affinity Propagation Applications to Gene-Expression Data," Bioinformatics, vol. 23, no. 20, pp. 2708-2715, Sept. 2007.

[20]. Text Clustering with Seeds Affinity Propagation,Renchu Guan, Xiaohu Shi, Maurizio Marchese, Chen Yang, and Yanchun Liang.

[21]. H. Zha, C. Ding, M. Gu, X. He, and H.D. Simon, "Spectral Relaxation for K-Means Clustering," Advances in Neural Information Processing Systems, vol. 14, pp. 1057-1064, MIT Press, 2001.

[22]. Z.H. Zhou, D.C. Zhan, and Q. Yang, "Semi-Supervised Learning with Very Few Labeled Training Examples," Proc. 22nd AAAI Conf. Artificial Intelligence, pp. 675-680, 2007.

[23]. S. Dumais, J. Platt, D. Heckerman, and M. Sahami, "Inductive Learning Algorithms and Representations for Text Categoriza- tion," Proc. Seventh Int'l Conf. Information and Knowledge Manage- ment, pp. 148-155, 1998.

[24]. S. Huang, Z. Chen, Y. Yu, and W.Y. Ma, "Multitype Features Coselection for Web Document Clustering," IEEE Trans. Knowl- edge and Data Eng., vol. 18, no. 4, pp. 448-458, Apr. 2006.

[25]. L.P. Jing, M.K. Ng, and J.Z. Huang, "An Entropy Weighting K- Means Algorithm for Subspace Clustering of High-Dimensional Sparse Data," IEEE Trans. Knowledge and Data Eng., vol. 19, no. 8, pp. 1026-1041, Aug. 2007.

[26]. [26] F. Wang and C.S. Zhang, "Label Propagation through Linear

[27]. Neighbourhoods," IEEE Trans. Knowledge and Data Eng., vol. 20,

[28]. no. 1, pp. 55-67, Jan. 2008.