Constrained Clustering with Seeds and Term Weighting Scheme

Uraiwan Buatoom
Sirindhorn International Institute of Technology, Thammasat University, PathumThani, Thailand
Email: uraiwanb31@gmail.com, uraiwan.buatoom@student.siit.tu.ac.th

Waree Kongprawechnon
Sirindhorn International Institute of Technology, Thammasat University, PathumThani, Thailand
Email: waree@siit.tu.ac.th

Thanaruk Theeramunkong
Sirindhorn International Institute of Technology, Thammasat University, PathumThani, Thailand
Associate Fellow, The Royal Society of Thailand, Bangkok, Thailand
Email: thanaruk@siit.tu.ac.th

***Abstract*—While traditional unsupervised learning is blind and the performance relies on the choice of initial seeds. The idea of constrained clustering can use a small number of labeled instances to partly guide a large number of unlabeled instances. It focuses on a set of predefined classes with an aim is to increase the performance of supervised and unsupervised learning using constraints. This paper proposes a new idea of semi-supervised learning based on particularly seeded constrained clustering, where the clustering guidance comes from the statistics of a small set of labeled data. In contrast with existing approaches in seeded K-Means where the labeled instances are specified. However, the proposed work investigates how weighting obtained from a training set affects the seeded-clustering results. Experimental results are demonstrated on three groups of term-weighting statistics; in-collection, intra-class, and inter-class based on frequencies/distributions and an ambiguity class pass entropy value. Text datasets is studied in our experiment. The result also depicts that the term weighting scheme is a potential mean to control/guide the initial and clustering process over a standard normal term weighting scheme.**

***Index Terms*—Semi-supervised, Term weighting, Distribution class, Ambiguity class and Seeded k-means.**

I. Introduction

Generally, classification (supervised learning), and clustering (unsupervised learning) are two complementary mining tasks where the former uses a set of labeled objects as examples, but the latter uses simple groups of similar objects without labeled information. However, due to the costly construction of labeled data (objects), several researchers applied a relatively small set of labeled data (objects) to create a predictive model, next used this model to label a large number of unlabeled data and then revised the initial predictive model using the automatically-labeled data. The method which is known as semi-supervised learning, was shown to improve the classification performance in several literatures [1]–[3].

On the other hand, clustering works with data without predefined labels by grouping them based on a sort of similarity. Recently, instead of blind grouping natural of clustering, there have been a number of works that incorporate a set of constraints to control the clustering towards the user desire [4], [5]. Constraints clustering focuses on information of a small labeled data to aid a bias in the clustering process, which generate from the labeled-level [6], [7], the instance-level, whereas the instance-level represent in the two types of pair-wise constraints; namely CANNOT links and MUST links [4], [8] and the cluster-level [9].

According to the resultant clusters were accuracy depends on the initial centroid. Most of them [6], [10], [11] approach work on incorporating labeled constraints into clustering methods, that uses the labeled data to generate seed clusters for reducing the chances of poor local optimal. The result showed that the proposed method can obtain better performance than the method which used a random seed. Nowadays, there is a challenging task on how to represent the constraints during the clustering process. Most research still focuses on the constrained-based approach by pairwise constraints, which provides two types of constraints between two data instances of clustering. This consists of pairwise associated instances in the same cluster called "MUST links" and a pairwise unrelated class of instances called "CANNOT links". However, the background knowledge is specific on the instances and also expresses the attribute/term level as discussed in [12], [13]. Even now, the term distribution that was introduced by inter-class (ICSD), intra-class (CSD), and in-collection (SD) is useful to improve the accuracy of centroid-based categorization [14]. Additionally, the term probability (entropy) is also offered. The both of terms of distribution can be applied to the heuristics to be a local weighting scheme for guidance on the clustering near the clustering solution.

In this research work, we propose a constraints in the form of statistics extracted from classes to weight terms as well as a framework of clustering to parameterize the unsupervised learning, especially clustering and weighting schemes. The effects of weighting schemes on the clustering is studied for the principles of guided weighting from the background knowledge of a dataset with a known labeled. Moreover, the terms in an unknown label dataset are re-assembled by promoting or demoting following on the term distribution and ambiguity of weighting scheme. In the proposed method, the quality of clusters is measured by three ways; class based (i.e., accuracy, f-measure, and conclusion with Geo-mean), cluster-based (i.e., purity), and similarity-based (i.e., inter-class similarity cosine). The efficiency of experiments based

on the term-weighting statistics is also studied by using text datasets which are balance and unbalance classes. Finally, the result depicts that the term weighting scheme can guide the clustering, followed by the expected clustering, by learning the characteristic term distribution of a class.

This paper is organized as follows: Section II describes the mathematical formulation of related work, including constraint clustering, the basic vector space model, and term weighting. In Section III, the proposed method and simulation model is illustrated. Section IV presents the experimental settings and performance measures, belonging to the properties of the class-based, cluster-based, and similarity-based. In Section V, the experimental result and error analysis are discussed. Finally, the conclusion is discussed in Section VI.

II. Related work

This section starts with an introduction of related work and discusses their involvement with our proposed work.

A. Clustering vs Seeded Constraint clustering

The most known traditional clustering is k-means. The k-means method is an unsupervised clustering task, where documents are partitioned into sub-groups. Let $D = \{d_1, d_2, ..., d_M\}$ be a set of M documents, where document $d_m = \{tf_{m1}, tf_{m2}, ..., tf_{mN}\}$ is a set of N keywords, so the group of instances represented by the M documents and N keywords. It proceeds by selecting the number (K) for the initial set of clusters $W_j = \{w_1, w_2, ..., w_K\}$. Let $W_j \subset D$ be the initial first centroid $IC_j = \{ic_1, ic_2, ..., ic_K\}$ occurring by randomly assigning each instance (d_i) to a cluster (w_K). Each instance is assigned to the closest cluster by measurement $D(d_i, ic_j)$, i.e., the distance between the instance with each centroid. The partition clustering is iteratively refined until the centroids stabilize. The k-means clustering method, as discussed above, has a drawback that it initial centroid and selects group documents without any knowledge of categories. Thus, the majority of existing methods are not sensitive enough to describe the effectiveness of term. The performance of k-means clustering criterion is a highly sensitive from the initial seed selection. One work is approached to keep them stuck in the poor local optimal point based on k-means. This initialization technique is created to preservers diversity of seeds while being robust to outliers based on simple probabilistic

The next on augmenting guide information with a small labeled of a semi-supervised method based on centroid, The centroid-based algorithm is a basic idea that is used some mean of a labeled vector for the representation of each class. Recently, Constraint clustering can be viewed as a class of semi-supervised learning algorithms with two types of constraints between two data instances of clustering. This algorithm is proposed to solve the blind data in clustering. It works on the concept of unsupervised learning (without any labeled training data), and supervised learning (with completely labeled training data). Basu et al. propose seeded k-means, that use the labeled data based on document frequency to generate the first initial centroid cluster but not keep during convergence clustering [6].

Many researchers [4], [5], [12] found that, not all constraint sets do not cover the feasibility issue and affected to improve the accuracy learning with a large amount of unlabeled data. However, we found a distinguishing term that can be expressed without specific instances of a given dataset. Therefore, we can represent the constraints in the form of term weightings based on the term distribution. This concept is applied to improve the accuracy of the centroid, based on the categorization by the background of labeled instance groups with inheritance dominant terms for an unlabeled data group, followed by the term distribution of the document, collection, and class in the labeled instance group [14]. This pattern keeps the flexible predicted class of the clustering k-means model and also improves the maintaining integrity of the specifications of the dataset. Thus, this paper proposes encapsulated dominant characteristics of instances to be a weighting term for term distribution, which propagate to the unlabeled instance group by promoting or demoting through terms.

B. Feature weighting

The terms in the single document cannot give the dominant of different class, so we need an extra effective set of the term that can represent a distinguishing term of each categorization. According to the bias-variance value of documents can normalize the term frequency to balance the dataset. In past, Lertnattee and Theeramunkong improved term word weight based on TFIDF adding the inner impact to disturb intra-class dispersion [14]. However, their methodology considers only the distribution of documents. This research proposes to improve the term weighting by considering the scattering, to observe the words within their distribution characteristics followed by in-collection, inter, and intra views of terms by class.

1) Basic Feature: The clustering method uses the vector space model to represent a document. Term weights are key parameters to help partition the label dataset into groups of similar follows based dimensions.

$$\text{TFIDF}(t) = \text{TF}(t, d) \times \text{IDF}(t) \tag{1}$$

TF is the frequency of a term (t), which appears in each document (d). IDF is used to avoid a part of solving a text classification. It is used to eliminate a general term. This considers the distribution of words in all document.

$$\text{IDF}(t) = log(1 + \frac{M}{m_{\text{t}}}) \tag{2}$$

where M represents the total number of documents and m_{t} is the number of documents that include the term t.

2) Term Weighting Using Term Distribution and Ambiguity: The feature/term selection base on a filter model which is independence any predictor but relies on the general characteristics of the training data that focus on relevance between term character and class. The usual term distribution in class is popular to explore the relation between optimal term subset

selection and relevance. In view of statistics, the distribution of term can be calculated on discrete and probability instead of absolute information. For this research, the standard deviation (discrete data) and entropy (probability data) can be defined for measurement. In order to modify the term values by using the term distribution within a class, we have used the vector space model for making it weighted. The term distribution weighting, used to adjust the weights of terms follows inter class and intra class characteristic instances. It introduces the term distribution from the variance of the term frequency values (tf). The term distribution is a proposed term weighting technique used to find the distinguishing terms. The important terms have properties as below:

- Terms of word should not appear in a whole collection.
- Distinguishing/dominant terms should mostly appear in a certain class and less in others.
- Distribution terms should not be different among instance in each class.
- Among classes, there should an altered term distribution.

On the other hand, the ambiguity properties are discussed by entropy as below:

- The number of occurrences of all terms named as, term and not term in whole document should not be balanced.
- The balance between terms named as, term and not term should not appear balance among classes.
- Dominant terms should have very few balance between terms and not terms in each class.

3) Standard deviation based on class views: For the properties which are mentioned overall, we can improve the term distribution to be term weighting, and then let it be re-weighting the vector space model of the term "n" of the document "m" in the cluster. The formal definition of term distribution weighting is organized by standard deviation based on in-collection, intra class, and inter class as follows:

a) Standard deviation of term (SD):

$$\text{SD}_n = \sigma_n = \sqrt{\frac{1}{M}\sum_{m=1}^{M}(tf_{mn} - \mu_n)^p}, \mu_n = \frac{1}{M}\sum_{m=1}^{M} tf_{mn} \tag{3}$$

where p is an arbitrary positive power

b) Inter-class standard deviation (ICSD):

$$\text{ICSD}_n = \sigma_{\text{ICSD}} = \sqrt{\frac{1}{K}\sum_{k=1}^{K}(\mu_{nk} - \mu_{nK})^p} \tag{4}$$

$$\mu_{nK} = \frac{1}{K}\sum_{k=1}^{K}\mu_{nk} \quad and \quad \mu_{nk} = \frac{1}{J}\sum_{j=1}^{J} tf_{jn}$$

K is the number of clusters to initial a cluster set $C_K = \{c_{k1}, c_{k2}, ..., c_k\}$, J is the number of documents in the classc_k.

c) Class standard deviation (CSD):

$$\text{CSD}_{nk} = \sigma_{\text{CSD(nk)}} = \sqrt{\frac{1}{J}\sum_{j=1}^{J}(tf_{jnk} - \mu_{nk})^p} \tag{5}$$

where tf_{jnk} is the value of documents j of term n in class k.

d) Average class standard deviation (ACSD):

$$\text{ACSD}_{nk} = \sigma_{\text{ACSD}(nk)} = \frac{1}{K}\sum_{k=1}^{K}\sigma_{\text{CSD}(nk)} \tag{6}$$

In addition, The bias-variance value of term distribution weighting can apply the p-Norm for minimized outliers. The efficacy of standard deviation is compared with power p, the Norm-2 has sensitive with outliers.

4) Entropy based on class views: This criterion is based on the distribution probability of the documents containing the term in the categories. The traditional entropy is in the range between 0 and 1 that nearly one mean the most impurity. The formal definition of term distribution weighting is organized by entropy based on in-collection, inter and intra class as follows:

a) Entropy of term (E):

$$\text{E}_n = (-P(t_n)(log_2 P(t_n)) + (-P(\bar{t}_n)(log_2 P(\bar{t}_n))) \tag{7}$$

b) Class-based Entropy (CE):

$$\text{CE}_n = \Big(-\sum_{k=1}^{K} P(t_n, c_k)(log_2 P(t_n, c_k))\Big) + \Big(-\sum_{\text{k=1}}^{K} P(\bar{t}_n, c_k)(log_2 P(\bar{t}_n, c_k))\Big) \tag{8}$$

c) Class-based Conditional Entropy (CCE):

$$\text{CCE}_n = \sum_{k=1}^{K} P(c_k)\Bigg(\Big(-P(t_n|c_k)(log_2 P(t_n|c_k))\Big) + \Big(-P(\bar{t}_n|c_k)(log_2 P(\bar{t}_n|c_k))\Big)\Bigg) \tag{9}$$

where t_n represents the number of documents that term t_n occurs at least one time and $\bar{t}_n$ is the number of document which does not appear term t_n.

III. Proposed weighting term

Before we begin to cluster a dataset, we must select the represented term of a class. Different terms should have different weights. Thus, before the using clustering method, there is a need of term weighting approaches to guide the blind data. The challenge is to selecting appropriate terms of documents that should be used for clustering. Selecting a term in an efficient way can improve the vector-based model with the distribution of terms of documents. For this reason, each class should consist of the different distinguishing terms. Then, we can analyze the background of a document. A higher weigh means that a term appears in high frequency in a certain class and rarely appears among the various classes. This weighting is used to promote the characteristics of a document. From the idea of the high variance indicates that the discrimination ability of the term distribution is not strong. The model in Figure 1 shows the inter class-side should maximize the distance/dissimilarity between mean of two classes, and the

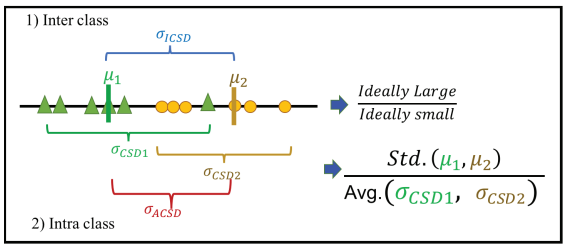

Figure 1: Term distribution weighting characteristics

intra class-side should minimize the variation (scatter) within each category. Moreover, for each cluster also should not balance term named as term and not term. The multiplicative model is also superior for a combination between the vector model and the characteristic weightings. Hence, our weighting model is as follows:

$$\text{TW} = \text{TFIDF} \odot (\text{STW}|\text{ETW}) \tag{10}$$

$$\text{STW} = \text{SD}^{\alpha} \times \text{ACSD}^{\beta} \times \text{ICSD}^{\gamma} \tag{11}$$

$$\text{ETW} = \text{E}^{\alpha} \times \text{CCE}^{\beta} \times \text{CE}^{\gamma} \tag{12}$$

The Eq. (10) presents how can work on incorporating labeled constrained into vector space by element-wise multiplication. The Eq. (11) shows how to contribute the term distribution weighting with the vector-based model of a document. Combing the characteristics of term weighting with the vector space model, these can act as a promoter (multiplier) or demoter (divisor). For ICSD and ACSD, these can represent the term of a class weight. Clustering the documents of the same group should be maximized. Other groups should be differently minimized. The deviation from ICSD is used to check the quality of an inter class. A high ICSD score means these terms are strongly represent among the class. Then, the ICSD acts as a promoter. ACSD is used to measure the quality of an intra class by measurement of variance in each class. A good term that represents a class should have a low score. Hence, ACSD should be a demoter. Furthermore, The SD score represents the terms of a collection factor. The high occurrence frequency of a term in a collection approaches a non-distribution term, which cannot be represent among classes. Thus, the high frequency of collection is scaled down scored by a factor that considers its collection frequency.

In addition,The Eq. (12) shows how to construct term weighting for ambiguity viewpoint. For E is the logarithm of the probability distribution is useful as a measurement of balance the number of occurrence term in a document. CE is used to consider a balance distribution term which uses the joint probability distribution that calculates the likelihood of two events occurring together and at the same even of class. The last, CCE is also used to check the balance distribution that chooses the conditional probability. This measurement calculates the probability occurrence of a term of an event class. Hence, all of them should promote at the lower score for avoid balance of terms in a document consider in a various viewpoint of class properties. The range of power weight $(\alpha, \beta, and \gamma)$ was between -1 to 1. This range is used to study the impact of term weighting for the guide vector of the dataset.

IV. Experiment Parameters and Performance Measurement

The experiment and performance measurement consist of dataset to stimulate term measurement, which is expressed as:

A. Dataset

To evaluate the proposed method, we utilize three datasets. In the first group, the "Thaireform" is short comments sentence from politics of Thailand (http://static.thaireform.org/). We select the second group from the "Amazon" dataset, which is a collection of reviews taken from Book, DVD, and Electronics domains in Amazon is used. The last one select from standard of WebKB dataset. For document-term representation, the TFIDF weighing is used. The dominant characteristics of the dataset are indicated in Table I.

Table I: Characteristics of three datasets

Dataset	Thaireform	Amazon	WebKB
No. of attributes	3549	6527	6527
No. of records	3000	6000	4161
Values MIN/MAX/ AVG./SD.	0/62/ 0.12/0.17	0/50/ 0.01/0.13	0/169/ 0.18/0.26
No. of classes	3	3	5
No. of docs/classes	1000/1000 /1000	2000/2000 /2000	221/3150 249/237/304

B. Quality of clustering measurement

The quality of clustering is evaluated using three criteria as follow:

1) Class-based measures: we use accuracy, f-measure, and balance score by geo-mean. Accuracy is used to consider the influence of term weighting. The effectiveness of clustering measurement is defined as the ratio of the total number of documents assigned with their correct classes (T_i) in all classes (K), compared with the total number of documents in the testing dataset (M).

$$\text{Accuracy} = A = \frac{\sum_{i=1}^{|K|} T_i}{M} \tag{13}$$

In General, multi-classification is evaluated by using a measurement, similar to the traditional measurement for evaluating a ranking based retrieval system, called precision (P) and recall (R). Mathematically, R_i represents the correctly identified document, which is the proportion between the number of retrieved correct documents (T_i) and the number of correct answers in each class (CM_i). P_i is also used to show true decisions, which define the proportion between the number of retrieved correct documents (T_i) and the number of retrieved answers in each class (RM_i).

$$\text{Recall} = R_i = \frac{T_i}{CM_i} \quad and \quad \text{Precision} = P_i = \frac{T_i}{RM_i} \quad (14)$$

The performance measurement may be misleading when examined alone. Normally, there is a measurement which represents the equal weight between R_i and P_i in each class, called f-measure. The f-measure is defined in two viewpoints, which are used for performance to indict per-class and all-class effectiveness. The average effectiveness of a classifier (F_i) is from pre-class trials. The macro-average $(\bar{F})$ is defined for measuring all-class performance, which is calculated by averaging the measurement over every class(F_i) on a testing dataset. Furthermore, the macro-average is given to the performance on all classes, regardless of how large the class is.

$$f\text{-measure} = F_i = \frac{2 \times R_i \times P_i}{R_i + P_i} \qquad and \qquad \bar{F} = \frac{\sum_{i=1}^{|K|} F_i}{|K|} \quad (15)$$

The geometric mean (GM) indicates the central tendency of a set of numbers by using the product of their values. Mathematically, it is defined as the n^{th} root of the product of 2 numbers$(n = 2)$, i.e., the GM of a dataset consists of accuracy and f-measure value is given by

$$\text{Geo-mean} = \text{GM} = \sqrt{A \times \bar{F}} \quad (16)$$

2) Cluster-based measure: The Purity is used to measure the percentage of one gold-standard partition which contain document primarily in a category, which is defined as

$$\text{purity} = \frac{1}{K} \sum_{k=1}^{K} Max|c_k \cap l_j| \quad (17)$$

From upper, c_k denote to be the set of cluster k for dataset then let $L_\text{j} = \{l_1, l_2, ..., l_\text{K}\}$represent the set of labeled categories.

3) Similarity-based measure : The average of cosine similarity between different clusters is aimed to get minimize for quality of cluster. The equation can evaluate as below.

$$\text{Inter} = \text{similarity}(A, B) = \frac{A.B}{\|A\|\|B\|} \quad (18)$$

For comparison, the range of inter value is clearly defined between 0-1 to 0-100.

V. Results and discussion

The experimental results demonstrate that the term weighting scheme is a potential means to control/guide the clustering process towards user intentions.

A. Effect of Single Additional Term Distribution Factors:

In the experiment, we conduct on 5-fold cross validation that dataset is split into 80% for training and 20% for testing.

Single additional term distribution factors from three term distribution factors are added one by one to the vector space model. The text datasets used TF, which Norm-1 the term frequency (TF) defined by $\frac{tf_{nm}}{\sum_{n' \in m} tf_{n'm}}$ to represent the vector space model. The text dataset is represented by the standard TF $\times$ IDF. The IDF weighting is used to take away some general words. The results in Table II show that, the bold numbers indicates the maximum accuracy value for the degree of power in each term distribution factor.

From Table II, the term distribution factor of predefined classes in many cases correlate the effectiveness with the ACSD and ICSD factors. The several datasets affect the negative degree of power for ACSD, and SD, but the ICSD is affected in a plus degree of power. In Table III is clear that all Entropt types affect in negative side. Normally, the big class can pull most instances into the same class. This table also shows that term weighting can improve the clustering, following the dominant character terms of a real class.

B. Analysis of Feature Distribution Factors with Different Power of Each Factor (based on multi additional term distribution factors)

In this experiment the multi additional term distribution factors consider the accuracy and f-measure by considering the geo-mean value, as shown in Table IV. The results in Table IV also depict the top ten effect of multi-additional term distribution factors that are categorized as best and worst scenario. The highest average value of geo-mean from three datasets is recorded at SD = -0.5, ACSD = -1, and ICSD = 0.5. The value of Centroid-Based (CB), Seeded k-mean (SK) are Thaireform = 96.12%, 95.56, AMAZON = 93.16%,92.7% and WebKB = 95.84%, 93.10% respectively. From the results it is clearly concluded that the set of query weighting of ICSD is promoted by plus side, ACSD and SD is demoted negative side. It is considered as a top affected. Furthermore, from Table III it is observed that in most cases of results E, CE, and CCE have the impact for a high power in the negative group hence the multiple also the same effect.

VI. Conclusion

For clustering, by using the term distribution weighting scheme, we can increase the accuracy by guiding the coefficient of term weight from expected results, and it can also reduce the cost of time to label. There are some parameters which may affect the performance of re-assemble by integrating with term distribution weighting terms. Furthermore, The most dominant factor is the effect of methodology for the cluster which follows intra-class and inter-class frequencies/distributions consider by term distribution and ambiguity of class. This paper also shows that we can rebuild a new vector which has potential means to control/guide the clustering process towards user intention. However, to improve the efficiency of document clustering we still need the flexible power of weighting and characteristics of a dataset. Finally, we would improve optimize the power of weighting, which could be studied in the future.

Acknowledgment

This research is financially supported under the Thammasat University′s research fund, Center of Excellence in Intelligent

Table II: The effect of the single term distribution factor (Variance)

Power of Weighting			Thaireform					AMAZON					WebKB					Avg.	Avg.	Avg.
$SD = \alpha$	$ACSD = \beta$	$ICSD = \gamma$	*ACC*	*F-1*	*GM*	*Purity*	*Inter*	*ACC*	*F-1*	*GM*	*Purity*	*Inter*	*ACC*	*F-1*	*GM*	*Purity*	*Inter*	GM	Purity	Inter
Panel I : Centroid-Based																				
-1	0	0	**94.90**	**94.95**	**94.93**	**94.90**	0.28	91.82	91.81	91.82	91.82	**0.26**	93.73	89.15	91.41	93.73	0.29	**92.72**	**93.49**	**0.2**8
1	0	0	75.84	75.86	75.85	75.84	33.25	84.67	84.77	84.72	84.67	3.60	58.76	56.16	57.44	75.73	5.75	72.67	78.75	14.20
0	-1	0	**95.17**	**95.22**	**95.20**	**95.17**	0.22	92.40	92.39	92.40	92.40	**0.20**	94.24	89.61	91.90	94.24	0.16	**93.17**	**93.94**	**0.20**
0	1	0	66.30	66.37	66.34	66.30	34.39	83.25	83.29	83.27	83.25	4.08	52.25	51.03	51.63	75.71	6.35	67.08	75.09	14.94
0	0	-1	79.60	80.08	79.84	79.60	**0.18**	75.82	76.63	76.22	75.82	0.26	84.15	74.80	79.33	84.15	0.23	78.47	79.86	**0.23**
0	0	1	85.10	**85.34**	85.22	85.10	35.86	80.74	81.21	80.97	80.74	3.31	**89.19**	82.83	**85.95**	**89.19**	3.51	**84.05**	**85.01**	14.23
0	0	0	93.70	93.78	93.74	93.70	3.92	90.99	90.97	90.98	90.99	1.38	86.55	81.09	83.78	86.57	2.02	89.5	90.42	2.44
Panel II : Seeded k-means																				
-1	0	0	**93.10**	**93.22**	**93.16**	**93.10**	**0.28**	91.01	90.98	90.99	91.01	0.26	90.58	84.23	87.35	90.58	0.28	**90.50**	**91.57**	**0.28**
1	0	0	52.70	55.43	54.05	52.7	30.89	71.14	73.78	72.45	71.14	3.25	36.60	26.65	31.23	75.71	5.25	52.58	66.52	13.13
0	-1	0	**94.01**	**94.10**	**94.05**	**94.01**	0.22	91.72	91.71	91.71	91.72	**0.20**	85.84	79.69	82.71	86.47	0.16	**89.49**	**90.74**	**0.20**
0	1	0	49.47	50.76	50.11	49.84	32.27	70.24	72.02	71.12	70.24	3.75	35.32	26.79	30.76	75.71	5.81	50.67	65.27	13.95
0	0	-1	68.34	67.5	67.92	68.40	0.16	44.34	50.25	47.2	44.34	0.14	66.36	59.15	62.65	77.15	**0.20**	59.26	63.30	**0.17**
0	0	1	76.31	79.85	78.06	76.31	33.37	74.07	76.18	75.12	74.07	3.12	**88.59**	**82.02**	**85.24**	**88.59**	3.49	**79.48**	**79.66**	13.33
0	0	0	92.91	93.12	93.01	92.91	3.88	90.35	90.36	90.36	90.35	1.38	77.20	72.63	74.88	84.19	1.87	86.09	89.15	2.38

Table III: The effect of the single term Ambiguity factor (Entropy)

Power of Weighting			Thaireform					AMAZON					WebKB					Avg.	Avg.	Avg.
$E = \alpha$	$CCE = \beta$	$CE = \gamma$	*ACC*	*F-1*	*GM*	*Purity*	*Inter*	*ACC*	*F-1*	*GM*	*Purity*	*Inter*	*ACC*	*F-1*	*GM*	*Purity*	*Inter*	GM	Purity	Inter
Panel I : Centroid-Based																				
-1	0	0	90.54	**90.61**	**90.57**	90.54	0.15	89.39	89.38	89.38	89.39	**0.11**	**91.21**	84.90	87.99	**91.21**	0.15	**89.32**	**90.38**	**0.14**
1	0	0	88.37	88.66	88.52	88.37	18.12	80.4	80.46	80.43	80.4	5.45	62.66	59.59	61.11	75.95	7.5	76.69	81.58	10.36
0	-1	0	93.04	**93.07**	**93.05**	93.04	0.15	89.75	89.74	89.75	89.75	**0.10**	**94.02**	89.26	91.61	**94.02**	0.14	**91.47**	**92.27**	**0.13**
0	1	0	83.67	83.78	83.73	83.67	18.63	79.54	79.54	79.54	79.54	5.61	49.20	46.78	47.98	75.71	7.74	70.42	79.64	10.66
0	0	-1	**94.07**	**94.15**	**94.11**	**94.07**	2.71	91.59	91.57	91.58	91.59	**1.11**	90.68	85.26	87.93	90.68	1.49	**91.21**	**92.12**	**1.77**
0	0	1	92.74	92.85	92.80	92.74	5.72	90.02	90.00	90.01	90.02	1.73	80.28	75.49	77.84	83.49	2.75	86.89	88.75	3.4
0	0	0	93.70	93.78	93.74	93.70	3.92	90.99	90.97	90.98	90.99	1.38	86.55	81.09	83.78	86.57	2.02	89.50	90.42	2.44
Panel II : Seeded k-means																				
-1	0	0	87.67	87.88	87.78	87.67	0.15	**88.11**	**88.08**	**88.09**	**88.11**	**0.11**	87.91	80.37	84.06	87.91	0.14	**86.65**	**87.90**	**0.14**
1	0	0	84.67	85.85	85.26	84.67	17.78	69.17	70.21	69.69	69.17	5.16	38.87	31.85	35.18	76.09	7.10	63.38	76.65	10.02
0	-1	0	85.95	86.97	86.46	85.95	0.15	88.57	**88.55**	**88.56**	88.57	**0.10**	**91.14**	84.97	88.00	**91.14**	0.13	**87.68**	**88.56**	**0.13**
0	1	0	74.73	77.18	75.94	74.73	17.92	64.09	64.81	64.45	64.09	5.24	34.99	28.20	31.41	75.71	7.29	57.27	71.51	10.15
0	0	-1	**93.67**	**93.86**	**93.77**	**93.67**	2.7	91.15	91.15	91.15	91.15	**1.11**	79.45	75.39	77.39	84.62	1.39	**87.44**	**89.82**	**1.74**
0	0	1	91.89	92.19	92.04	91.89	5.67	89.02	89.04	89.03	89.02	1.73	69.46	62.25	65.76	80.97	2.53	82.28	87.30	3.31
0	0	0	93.39	93.61	93.5	93.39	3.9	90.35	90.36	90.36	90.35	1.38	77.2	72.63	74.88	84.19	1.87	86.25	89.31	2.39

Table IV: The top-10 effect by the multi-term factors

Method	Power of Weighting					Total
	-1	*-0.5*	*0*	*0.5*	*1*	CB,SK
SD	3(0), 3(0)	4(0), 4(0)	3(0), 2(2)	0(3), 1(4)	0(7), 0(4)	10(10), 10(10)
ACSD	4(0), 4(0)	4(0), 4(0)	2(0), 2(0)	0(2), 0(4)	0(8), 0(6)	10(10), 10(10)
ICSD	0(3), 0(5)	0(3), 0(5)	5(2), 4(0)	5(1), 5(0)	0(1), 1(0)	10(10), 10(10)

Informatics, Speech and Language Technology and Service Innovation (CILS), and Intelligent Informatics and Service Innovation (IISI) Research Center, the Thailand Research Fund under grant number RTA6080013, the Thammasat University Fund on Research on Intelligent Informatics for Political Data Analysis,the Faculty Development Fund at Burapha University, Chanthaburi Campus,

REFERENCES

[1] K. Nigam, A. K. McCallum, S. Thrun, and T. Mitchell, "Text classification from labeled and unlabeled documents using em," *Machine learning*, vol. 39, no. 2, pp. 103–134, 2000.

[2] Z. Zhao, W. Qi, J. Han, Y. Zhang, and L. fa Bai, "Semi-supervised classification via discriminative sparse manifold regularization," *Signal Processing: Image Communication*, vol. 47, pp. 207 – 217, 2016.

[3] A. Dong, F. lai Chung, and S. Wang, "Semi-supervised classification method through oversampling and common hidden space," *Information Sciences*, vol. 349, pp. 216 – 228, 2016.

[4] K. Wagstaff, C. Cardie, S. Rogers, S. Schrödl *et al.*, "Constrained k-means clustering with background knowledge," in *ICML*, vol. 1, 2001, pp. 577–584.

[5] I. Davidson, K. L. Wagstaff, and S. Basu, "Measuring constraint-set utility for partitional clustering algorithms," in *European Conference on Principles of Data Mining and Knowledge Discovery*. Springer, 2006, pp. 115–126.

[6] S. Basu, A. Banerjee, and R. Mooney, "Semi-supervised clustering by seeding," in *In Proceedings of 19th International Conference on Machine Learning (ICML-2002*. Citeseer, 2002.

[7] Z. Zhang, J. T. Kwok, and D.-Y. Yeung, "Parametric distance metric learning with label information," in *IJCAI*. Citeseer, 2003, p. 1450.

[8] A. George, "Efficient high dimension data clustering using constraint-partitioning k-means algorithm." *Int. Arab J. Inf. Technol.*, vol. 10, no. 5, pp. 467–476, 2013.

[9] I. Davidson and S. Ravi, "Agglomerative hierarchical clustering with constraints: Theoretical and empirical results," in *European Conference on Principles of Data Mining and Knowledge Discovery*. Springer, 2005, pp. 59–70.

[10] F. Khan, "An initial seed selection algorithm for k-means clustering of georeferenced data to improve replicability of cluster assignments for mapping application," *Applied Soft Computing*, vol. 12, no. 11, pp. 3698–3700, 2012.

[11] X. Li, Y. Liang, and Y. Cai, "Optimizing initial centroids by density peak and entropy weighting method fork-means algorithm," *Journal of Bioinformatics and Intelligent Control*, vol. 4, no. 2, pp. 111–116, 2015.

[12] D. Klein, S. D. Kamvar, and C. D. Manning, "From instance-level constraints to space-level constraints: Making the most of prior knowledge in data clustering," Stanford, Tech. Rep., 2002.

[13] J. Schmidt, E. M. Brandle, and S. Kramer, "Clustering with attribute-level constraints," in *Data Mining (ICDM), 2011 IEEE 11th International Conference on*. IEEE, 2011, pp. 1206–1211.

[14] V. Lertnattee and T. Theeramunkong, "Effect of term distributions on centroid-based text categorization," *Information Sciences*, vol. 158, pp. 89–115, 2004.