PAPER Special Section on Intelligent Information and Communication Technology and its Applications to Creative Activity Support

Improving Seeded k-Means Clustering with Deviation- and Entropy-Based Term Weightings

Uraiwan BUATOOM[†a)], *Member*, Waree KONGPRAWECHNON[†b)], *Nonmember*, *and* Thanaruk THEERAMUNKONG[†,††c)], *Member*

SUMMARY The outcome of document clustering depends on the scheme used to assign a weight to each term in a document. While recent works have tried to use distributions related to class to enhance the discrimination ability. It is worth exploring whether a deviation approach or an entropy approach is more effective. This paper presents a comparison between deviation-based distribution and entropy-based distribution as constraints in term weighting. In addition, their potential combinations are investigated to find optimal solutions in guiding the clustering process. In the experiments, the seeded k-means method is used for clustering, and the performances of deviation-based, entropy-based, and hybrid approaches, are analyzed using two English and one Thai text datasets. The result showed that the deviation-based distribution outperformed the entropy-based distribution, and a suitable combination of these distributions increases the clustering accuracy by 10%.

key words: *term weighting, deviation-based weight, entropy-based weight, in-collection intra-class and inter-class information, k-means document clustering*

1. Introduction

Term weighting is a potential tool to guide/control a process and to improve classification (supervised learning) [1], [2] or clustering performance (unsupervised learning) [3], by assigning higher weights to more important terms. In the vector space model, term weighting can help to select informative terms/words and exclude common terms/words in order to construct a vector for a document. The traditional term weighting depends only on term frequency in the document and the inverse number of training documents that contain this term [4]. Instead of blind grouping, clustering can be extended to exploit constraints during the clustering process. Previously, the concept of controlled unsupervised learning has been studied by several works as a more specific task called constrained clustering [5]–[8]. In constrained clustering, the background knowledge or constraint can be in the form of either the labeled data (direct usage of class labels) [8], the instance-level constraint (must-link or cannot-link sets) [5], [9], or the cluster-level constraint (cluster size, lower/upper bounds on the radii, etc.) [10], [11]. However, most existing constrained methods ignore the effect of attributes (features) on the representation of each datum on the space (vector space model). As an exploration of attribute-level (feature-level) constraints in the clustering process, Schmidt et al. [12] proposed to use specific attribute values to control whether instances may or may not be assigned to the same group. The concept of the approach is to induce clusters of binary instances that satisfy constraints on the attribute level. They also showed how the well-established instance-level constraints, must-link and cannot-link, can be adapted to the attribute level. In this approach, rather than using instance-level constraints, they used information encoded in the form of characteristics of attributes that such instances hold. Although, this work seems an early work on attribute-level constraints, it handles only binary attributes.

In the past, for the classification task, some work used learning to weight for classification [33] and some applied the attribute-level constraints which are non-binary to improve classification performance. These works used different class-related statistics for weighting terms, such as chi-square, information gain, and gain ratio. Statistical confidence intervals can be used to guide clustering with such prior knowledge in the form of statistics. Two statistical approaches; namely the deviation approach [13], [14] and the entropy approach [15], were used to enhance discrimination with class information. For the former, Lertnatee and Theeramunkong [13] reported that using a term distribution (SD, CSD, and ICSD) in term weighting helped improve the accuracy of centroid-based categorization. The latter is to utilize probability, rather than distribution, to strengthen discrimination. Recently, Chai et al. (2016) [16] have applied the weighting coefficients, namely RELIEF, I-RELIEF, and B-LM2FW, to describe the relevant characteristics of different term/features. However, most works on these approaches focused on classification, not clustering. Some researchers suggested a combined term weighting to avoid bias from a single source of information [17]. While most works on term weighting focused on classification, it is still an open question on which statistical approaches are suitable for unsupervised learning, like the clustering task.

Based on the above background, this paper presents a method to improve the seeded k-means clustering by utilizing distribution term weighting as constraints for con-

Manuscript received March 27, 2019.
Manuscript revised October 30, 2019.
Manuscript publicized January 8, 2020.

[†]The authors are with the School of Information, Computer, and Communication Technology (ICT), Sirindhorn International Institute of Technology, Thammasat University, PathumThani, Thailand.

[††]The author is with The Royal Society of Thailand, Bangkok, Thailand.

a) E-mail: uraiwanb31@gmail.com (Corresponding author)
b) E-mail: waree@siit.tu.ac.th
c) E-mail: thanaruk@siit.tu.ac.th

DOI: 10.1587/transinf.2019IIP0017

trolling/guiding the process of clustering documents. Three types of the statistics, that is, in-collection, intra-class, and inter-class distributions, are used as term weighing, to express the behavioral of categories. Two schemes; alternatively deviation and entropy, are used to assign an appropriate weight to each term. Their performances are compared and analyzed. Even our previous work [18] first proposed utilization of deviation and entropy as term weighting for seeded k-means clustering, the investigation was done in a small scale and only preliminary experiments are conducted. In this paper, we further study the effect of deviation- and entropy-based term weighting in the quality of seeded k-means clustering, in terms of geo-mean of accuracy and f-measure, using three text datasets, i.e., two English and one Thai texts. The combination of deviation- and entropy-based term weightings is also explored. The rest of this paper is organized as follows. Section 2 describes previous works related to the distribution-based term weighting scheme using class information. In Sect. 3, the deviation- and entropy-based term weighting as well as their combined model, are presented. The document clustering using the proposed weights is given in Sect. 4. Section 5 provides the experimented settings and the performance measurement. In Sect. 6, the experimental results and error analysis are discussed. Finally, the discussion and conclusion are given in Sect. 7.

2. Related Works

This section describes existing works on term weighting that use class information, such as term distribution in a class or among classes, to improve performance. In the past, as the simplest approach, the inverse class frequency (ICF) was introduced to extend the conventional TFIDF with class information for text classification. Ren and Sohrab [19] introduced a so-called class-indexing-based term weighting approach with the inverse class frequency (ICF) function and a new inverse class space density frequency (ICS$_\delta$F), to improve classification. The ICF and ICS$_\delta$F was shown to help to emphasize or prefer words that occur only in a few classes, while it ignores words that appear in all classes.

Besides ICF, some previous works applied other statistical methods, particularly term distribution, to use class information for classification. Such approaches can be classified into two types: deviation- and entropy-based term weightings [13], [14], [20]. Some weighting methods used deviation-based distributions obtained from a set of labeled data to reflect the importance of a term in a certain class [21]. As a fuzzy approach, Lo et al. [22] proposed an objective weighting model based on the maximum deviation, and then integrated the interval number and distance function into the main structure to handle the uncertain information. Fattah [23] used the term weighting based on class density to reflect the relative importance of a certain term in a certain class. Most methods just measured the distribution of a term in a certain class relative to the whole documents in a class, but Lertnattee and Theeramunkong [13] proposed an improvement in the term weighting method that also addresses the deviations related to classes. They also applied the term distribution weighting scheme to adjust the weights of the terms that follow the collection, inter-class, and intra-class characteristics to improve classification. This term weighting technique is used to find distinguishing terms, and then to promote a specific term and/or to demote general terms.

Besides deviation, entropy is an alternative to the use of class information. In general, entropy is a measure of uncertainty or randomness. The higher the entropy of an object, the more uncertain the object's state is. Nigam et al. [24] proposed an approach to use the maximum entropy to measure the importance level of a term by estimating the constraints on the distribution of the class variable given to the document for text classification. As an alternative to the deviation-based approach, entropy-based weighting schemes can be used to improve the feature selection. Belonging to this approach, REMI used a robust measure of feature quality, called rank entropy, to compute the uncertainty in monotonic classification [20], [25]. Similarly, FS-JMIE applied a joint maximal information entropy method between features and class to define a metric for the efficiency of a feature subset evaluation [26]. Fragos et al. [27] extended the maximum entropy modeling with χ^2 to reduce the dimensionality of data and improve feature weighting. As a more recent work, Wu et al. [28] developed entropy to measure term distribution and term bias to control over-weighting and under-weighting. These methods presented three regularization techniques: (1) add-one smoothing for handling singular terms, (2) sublinear scaling and (3) bias term for shrinking the ratios between term weights. As a more specific topic, Lee et al. (2017) [29] showed how the entropy of the sentiments in the review texts characterized their influence and bias on the relationship between online word-of-mouth (WOM) and product sales.

3. Deviation- and Entropy-Based Distribution Term Weightings

This section describes deviation- and entropy-based distribution and their combination. To this end, the formulation of term weighting is shown as follows. Let $D = \{d_1, d_2, \ldots, d_{|D|}\}$ be a set of $|D|$ documents, $T = \{t_1, t_2, \ldots, t_{|T|}\}$ be a set of $|T|$ terms, and a weight can be given to each term t_n in a document d_m. The most commonly-used term weighting is term frequency (tf) and inverse document frequency (idf). Although there exist several variants of term frequency of the term t_n in the document d_m, one of the most popular settings is the norm-1 term frequency denoted by ntf_{mn} as shown in Eq. (1).

$$\text{ntf}_{mn} = \frac{\text{tf}_{mn}}{\sum_{s=1}^{|T|} \text{tf}_{ms}} \tag{1}$$

where tf_{mn} is the frequency of the n-th term in the m-th document. On the other hand, the inverse document frequency is often used to reduce effect of common terms that occur in most documents, such as articles or prepositions. The definition of the inverse document frequency of the term t_n, denoted by idf_n, can be defined as follows.

$$\text{idf}_n = \log\Big(\frac{|D|}{1+|D_n|}\Big) \tag{2}$$

$$D_n = \{d | d \in D \wedge in(t_n, d)\} \tag{3}$$

where $in(t_n, d)$ is the Boolean function representing the existence of the term t_n in the document d, D_n is the set of documents that include the term t_n, and $|D_n|$ is the number of documents in D_n.

3.1 Deviation-Based Term Weighting (DTW)

The standard deviation (SD) is a statistic that measures the dispersion of data values in a dataset, relative to its mean and is calculated as the square root of the variance. By this property, it is possible to use the SD in frequencies (occurrences) of a term occurring in a document in the document collection, the SD in frequencies of a term occurring in a class in the class system, and the SD in frequencies of a term occurring in a document in the focused class, for differentiate specific term and general terms. A term with a high standard deviation in its occurrences may be considered as more significant. It is also possible to consider the class-based deviation. Following the principles in [13], an important terms, i.e. terms that should be assigned with a high weight, tend to have the following properties.

- **In-collection deviation:** An important term (or word) tends to occur in a specific group of documents and they should not appear frequently in all documents in the whole collection.
- **Intra-class deviation:** An important term (or word) for a specific class, should occur almost equally in most documents in that class.
- **Inter-class deviation:** An important term (or word) should appear in a certain class and much fewer in other classes.

Although such properties were used in several works related to classification such as [13], [30], [31], it can be easily applied for clustering as follows. Let $C = \{c_1, c_2, \ldots, c_{|C|}\}$ be a set of $|C|$ clusters, then $C_k = \{d \mid d$ is a document that belongs to cluster $c_k\}$, where $\bigcup_{i=1}^{|C|} C_i = D$ and $C_i \cap C_j = \emptyset$. A value $\mathcal{F}$ (i.e., false) is assigned to $\langle d_i, c_k\rangle$ when document d_i does not belong to cluster c_k. Mathematically, the above three properties, i.e., the in-collection, inter-class, and intra-class distributions can be expressed by the standard deviation, inter-class standard deviation, and average class standard deviation as follows.

(1) In-collection Standard Deviation (sd_n):

$$\text{sd}_n = \sqrt{\frac{1}{|D|}\sum_{m=1}^{|D|}\Big(\text{tf}_{mn} - \mu_n\Big)^2} \tag{4}$$

$$\mu_n = \frac{1}{|D|}\sum_{m=1}^{|D|}\text{tf}_{mn} \tag{5}$$

The sd_n represents the standard deviation of a term t_n, calculated from its frequencies, denoted by tf_{mn}, in all documents d_m in the whole collection, and μ_n is the average term frequency of the term t_n in the documents in the collection. Note that this standard deviation is independent of a class but it relates to frequencies of the term in documents in the whole collection. While a term with a low sd_n is supposed to be an important term, it implies two situations: (1) when the term rarely appears, and (2) when the occurrences of the term are nearly equal in all documents for the whole collection. In contrast, a term with a high sd_n is trivial since it appears often or its frequencies are various among documents in the collection.

(2) Average Class Standard Deviation (acsd_n):

$$\text{acsd}_n = \frac{1}{|C|}\sum_{k}^{|C|}\sqrt{\frac{1}{|c_k|}\sum_{d_m\in c_k}\Big(\text{tf}_{mn} - \mu_{nk}\Big)^2} \tag{6}$$

The acsd_n represents the average standard deviation of the occurrences of the term t_n, with respect to all classes ($c_k \in C$). Implicitly, the acsd is an intra-class factor. A term has a low acsd when it is stable in its term frequencies among documents in the class. Then it is likely to be the representative of the class. Two typical situations are (1) when the occurrence of terms is nearly equal, and (2) when it rarely occurs for all documents in that class.

(3) Inter-class Standard Deviation (icsd_n):

$$\text{icsd}_n = \sqrt{\frac{1}{|C|}\sum_{n=1}^{|C|}\Big(\mu_{nk} - \mu_{n'}\Big)^2} \tag{7}$$

$$\mu_{n'} = \frac{1}{|C|}\sum_{k=1}^{|C|}\mu_{nk} \tag{8}$$

$$\mu_{nk} = \frac{1}{|c_k|}\sum_{d_m\in c_k}\text{tf}_{mn} \tag{9}$$

The icsd_n is the standard deviation of the t_n's average frequencies of the class μ_{nk} in Eq. (9). A term with a high icsd implies that the term occurs with dominantly different frequencies among the classes. Note that this factor, icsd, of a term is independent of classes.

3.2 Entropy-Based Term Weighting (ETW)

In the past, several works [15], [24], [26], [32] applied the entropy to identify important terms since the entropy is derived from the logarithm of a probability distribution and it can be used as an indicator for impurity level. Naturally an important term will have low entropy, that is low ambiguity in its probability to occur in a class or its probability to occur in the other classes. Similar to the above deviation-based term weighting, an important terms that have a high weight, should satisfy the following properties related to entropy.

- **In-collection entropy:** A term (or word) is considered to be prominent when there is somewhat balance between the number of documents which include that term, and the number of documents which do not include that term.

- **Intra-class entropy:** A term (or word) is considered to be prominent when there is somewhat imbalance between the number of documents in a class which include that term, and the number of documents in the class which do not include that term.
- **Inter-class entropy:** A term (or word) is considered to be prominent when there is somewhat imbalance between the number of documents in all classes which include that term, and the number of documents in all classes which do not include that term.

The formal definition of term distribution weighting is organized by entropy based on in-collection, inter-class, and intra-class as follows:

(1) In-collection Term Entropy (e_n)

The in-collection entropy of a term t_n, denoted by e_n, can be defined based on the probability that the document includes or does not include the term as shown below.

$$\mathrm{e}_n = -p(D_n)(\log_2 p(D_n) - p(\overline{D}_n)(\log_2 p(\overline{D}_n) \tag{10}$$

where D_n is the set of documents that include the term t_n, $\overline{D}_n$ is the set of documents that do not include the term t_n, and $p(D_n)+p(\overline{D}_n) = 1.0$. Intuitively, a term with somewhat high value for this type of entropy can be considered an important term. A term that either occurs in only few documents or occurs in almost all documents, will have a low entropy and it implies that such term is a not good representative terms/feature for classification or clustering.

(2) Class-based Term Conditional Entropy cce_n):

$$\begin{aligned}\mathrm{cce}_n &= \sum_{k=1}^{|C|} p(c_k)\Big(- p(D_n|c_k)\log_2 p(D_n|c_k) \\ &\qquad - p(\overline{D}_n|c_k)\log_2 p(\overline{D}_n|c_k)\Big) \qquad (11) \\ &= -\sum_{k=1}^{|C|} p(D_n, c_k)\log_2 p(D_n|c_k) \\ &\quad - \sum_{\mathrm{k}=1}^{|C|} p(\overline{D}_n, c_k)\log_2 p(\overline{D}_n|c_k) \qquad (12)\end{aligned}$$

The cce_n represents the summation of the entropy of the intra-class distribution on how likely documents in the class k includes the term t_n. That is, the summation of the entropy of binary distribution of the conditional probabilities; $p(D_n|c_k)$ and $p(\overline{D}_n|c_k)$ over the cluster set (corresponding to the class set) C. A term with a low cce value tends to be significant since the term either occurs in only few documents in the class, or occurs in almost all documents in the class and it implies that such term is a good representative feature for classification or clustering.

(3) Class-based Term Entropy (ce_n):

$$\begin{aligned}\mathrm{ce}_n &= -\sum_{k=1}^{|C|} p(D_n, c_k)\log_2 p(D_n, c_k) \\ &\quad - \sum_{\mathrm{k}=1}^{|C|} p(\overline{D}_n, c_k)\log_2 p(\overline{D}_n, c_k) \qquad (13)\end{aligned}$$

The ce_n represents the summation of the entropy of the inter-class distribution on how likely documents in the class k includes the term t_n. That is, the summation of the entropy of distribution of the joint probabilities; $p(D_n, c_k)$ and $p(\overline{D}_n, c_k)$ over the cluster set (corresponding to the class set) C. Note that $\sum_{k=1}^{|C|}(p(D_n, c_k) + p(\overline{D}_n, c_k)) = 1$. A term with a low ce value tends to be significant since the term has high variety in its distribution among classes.

3.3 Combined Term Weighting Scheme

This section presents the combined term weighting methods using the element-wise multiplication ($\odot$) with two weighting methods of three distributions, i.e., in-collection, inter-class, and intra-class deviation as well as in-collection, inter-class, and intra-class entropy. In this work, the weight of the term t_n in a document d_m, denoted by tw_{mn}, is designed to be in the form shown in Eq. (14).

$$\mathrm{tw}_{mn} = \mathrm{ntf}_{mn} \times \mathrm{idf}_n \times \mathrm{dtw}_n^{\alpha_1} \times \mathrm{etw}_n^{\alpha_2} \tag{14}$$

$$\mathrm{dtw}_n = \mathrm{sd}_n^{\beta_1} \times \mathrm{acsd}_n^{\beta_2} \times \mathrm{icsd}_n^{\beta_3} \tag{15}$$

$$\mathrm{etw}_n = \mathrm{e}_n^{\beta_4} \times \mathrm{cce}_n^{\beta_5} \times \mathrm{ce}_n^{\beta_6} \tag{16}$$

where α_1 and α_2 are the exponents or powers (weights) for dtw and etw; $\beta_1, \beta_2, \ldots, \beta_6$ are the exponents for sd, acsd, icsd, e, cce, and ce, respectively. In other words, the term weighting scheme (tw) is the multiplication of the normalized term frequency (ntf_{mn}), the inverse document frequency (idf_n), deviation-based term weights (DTW), and entropy-based term weights (ETW), where the DTW (dtw_n) and ETW (etw_n) components are weighted by α_1 and α_2, as shown in Eq. (14). The dtw_n are the multiplication of sd, acsd, and icsd weighted by β_1, β_2, and β_3, respectively. The etw_n are the multiplication of e, cce, and ce weighted by β_4, β_5, and β_6, respectively. Eq. (15) shows how distribution term weighting factors, i.e., sd, acsd, and icsd, contributes to weight a term (t_n). In the same way, Eq. (16) presents how entropy-based term weighting factors, i.e. e, cce, and ce, affects to weight a term (t_n). The weightings β_i's in the Eq. (15)-(16) can be set to positive values, zero values, or negative values, making the six factors (sd, acsd, icsd, e, cce, and ce) perform as a promoter, an inert, or a demotor.

4. Seeded *k*-Means Clustering Using Term Weighting

This section presents our framework of seeded k-means clustering where distribution statistics are extracted from the small set of C-classed labeled documents and used as constraints for clustering the large set of unlabeled documents

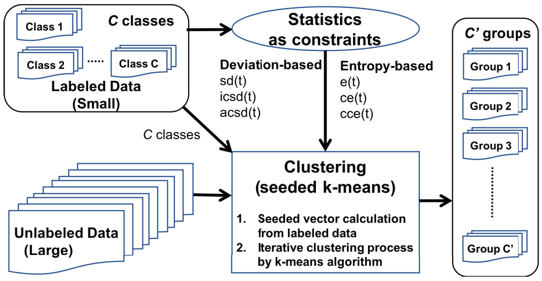


Fig. 1 The framework of seeded k-means clustering where distribution statistics are extracted and used as constraints for clustering.

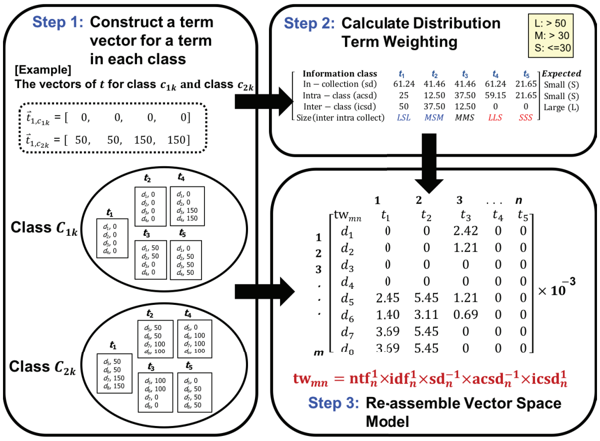

Step 1: Construct a term vector for a term in each class
[Example]
The vectors of t for class c_{1k} and class c_{2k}
$\vec{t}_{1,c_{1k}}$ = [0, 0, 0, 0]
$\vec{t}_{1,c_{2k}}$ = [50, 50, 150, 150]
Class C_{1k}
t_1: d_1, 0 d_2, 0 d_3, 0 d_4, 0
t_2: d_1, 0 d_2, 0 d_3, 0 d_4, 0
t_3: d_1, 50 d_2, 50 d_3, 0 d_4, 0
t_4: d_1, 0 d_2, 0 d_3, 150 d_4, 150
t_5: d_1, 0 d_2, 50 d_3, 50 d_4, 50
Class C_{2k}
t_1: d_5, 50 d_6, 50 d_7, 150 d_8, 150
t_2: d_5, 50 d_6, 50 d_7, 100 d_8, 100
t_3: d_5, 100 d_6, 100 d_7, 0 d_8, 0
t_4: d_5, 0 d_6, 100 d_7, 100 d_8, 100
t_5: d_5, 0 d_6, 50 d_7, 50 d_8, 50
Step 2: Calculate Distribution Term Weighting
L: > 50
M: > 30
S: <=30
Information class t_1 t_2 t_3 t_4 t_5 Expected
In – collection (sd) 61.24 41.46 41.46 61.24 21.65 Small (S)
Intra – class (acsd) 25 12.50 37.50 59.15 21.65 Small (S)
Inter – class (icsd) 50 37.50 12.50 0 0 Large (L)
Size (inter intra collect) LSL MSM MMS LLS SSS
1 2 3 . . . n
tw$_{mn}$ t_1 t_2 t_3 t_4 t_5
1 d_1 0 0 2.42 0 0
2 d_2 0 0 1.21 0 0
3 d_3 0 0 0 0 0
. d_4 0 0 0 0 0
. d_5 2.45 5.45 1.21 0 0
. d_6 1.40 3.11 0.69 0 0
d_7 3.69 5.45 0 0 0
m d_8 3.69 5.45 0 0 0
$\times 10^{-3}$
$\mathrm{tw}_{mn} = \mathrm{ntf}_n^1 \times \mathrm{idf}_n^1 \times \mathrm{sd}_n^{-1} \times \mathrm{acsd}_n^{-1} \times \mathrm{icsd}_n^1$
Step 3: Re-assemble Vector Space Model

Fig. 2 Example of term distribution weighting based on class information.

into C' groups, as shown in Fig. 1. Here, two types of the distribution statistics for term weighting are those shown in Sect. 3; i.e., deviation-based and entropy-based statistics.

In this figure, the unlabeled documents are encoded by conventional term weighting enhanced with distribution-based term weighting. At this stage, one can expect that appropriate term weighting will guide us to obtain good clusters of the unlabeled documents. In the past, commonly-used term weightings include binary frequency (bf), term frequency (tf), inverse document frequency (idf), and their combinations, such as tf × idf. In this work, we propose the term weighting scheme as constraints in the clustering process. The weights are calculated from the relatively small set of C-classed labeled documents The weights are of three types; i.e., in-collection, inter-class and intra-class statistics. The weighting are used to cluster the unlabeled data into C' groups. Figure 2 shows the clustering process that contains three main processes: (1) constructing a term vector for a term in each class, (2) calculating distribution term weighting, and (3) re-assembling the vector space model. In the step 1, the corpus is transformed into the bag-of-words, based on the vector space model. The vector space is modeled based on tf or ntf×idf, where tf represents the frequency of term features in each document, ntf represents normalize term, as shown in Eq. (1), and idf represents to settle the general words, as shown in Eq. (2). This example model can also be used in two groups, i.e., a vector of the labeled dataset based on term frequency tf (the upper part of the process) and the unlabeled vector based on term frequency with inverse document frequency (the lower part of process). In step 2, tf is used to encode the corpus for the distribution term weighting scheme. This term weighting encoding scheme is based on deviation-based and entropy-based approaches as mentioned in Eqs. (4)–(9) and Eqs. (10)–(13), respectively. In step 3, tf × idf is used as the based conventional term weighting for combining various distribution term weighting schemes from step 2. Note that the distribution term weighting is used for re-weighting the vector space model which is the formal definition of the term distribution weighting scheme. The conventional seeded k-means method assumes spherical-shaped clusters when the cosine similarity is applied for grouping documents.

5. Experiment Settings: Datasets and Measurements

5.1 Datasets

To evaluate the effectiveness of the proposed method three text datasets, i.e., Amazon, WebKB, and Thai-reform, are used for experiments, where the first two datasets are English corpora and the last one is a Thai corpus. Amazon is a collection of reviews, taken from Book, DVD, and Electronics domains from Amazon. WebKB consists of HTML-based texts and 4 popular items (student, faculty, course, and project) are selected from 7 classes, to categorize the data followed by 5 universities. Note that in this dataset 4161 documents from 4199 pages are selected that do not have only the structure of a web page. Thai-reform is the Thai comment which consists of 3,000 documents obtained from the Thai-reform webpage (http://static.thaireform.org/). However, some comments are in short sentences, while some are in lengthy detail, so 1,000 documents of each in the 3 largest classes are randomly selected. In the initial stage, we perform a pre-process of the two English language corpora, for example, we have transformed characters in the text into the lower case and then applied the Porter's Stemmer to make the words change to their root form. In practice, we omit the terms the frequency of which are less than 0.001 percent of the total number of words in the collection. Thes characteristics of three datasets (corpora) are summarized in Table 1.

5.2 Measurement

Our proposed clustering method is evaluated in three criteria: accuracy, f-measure, and geo-mean. Accuracy (A) is defined as the ratio of the number of documents assigned with their correct classes in all classes ($TP_1 + TP_2 + \ldots + TP_{|C|}$), compared with the total number of documents ($|D|$), where TP_i stands for the true positive of the class i and the number of classes is $|C|$.

$$A = \frac{\sum_{k=1}^{|C|} TP_k}{|D|} \tag{17}$$

The f-measure is defined in two viewpoints, which are used to evaluate the effectiveness of indict per-class and all-class new-points. The average effectiveness of a classifier (F_i) is from pre-class trials. The macro-average f-measure ($\bar{F}$) is used for measuring all-class performance, which is calculated by averaging the measurement over every class (f_k) on a testing dataset. Furthermore, the macro-average is calculated for the performance on all classes, regardless of the size (or length) of the class.

$$\bar{F} = \frac{\sum_{k=1}^{|C|} f_k}{|C|} \tag{18}$$

$$f_k = \frac{2 \times r_k \times p_k}{r_k + p_k} \tag{19}$$

Table 1 Characteristics of the experimental datasets.

Dataset	**Amazon**	**WebKB**	**Thai-reform**
General Information			
Language	English	English	Thai
No. of docs	6000	4161	3000
No. of classes	3	5	3
No. doc./class	2000each	221/237/249/304/3150	1000 each
No. of distinct terms	7614	6527	3549
redAvg. of distinct terms/doc	51.52	79.64	31.98
Avg. of distinct terms/class	6041.67 (6817/6933/4375)	4072.20 (3535/3380/3204/3807/6435)	2545.67 (2001/2835/2801)
Distribution Information			
No. of terms in a single class	205/312/233	16/23/10/24/936	161/384/301
No. of terms in the intersection of two classes	2722/252/243	0/3/6/366/4/3 335/1/312/496	203/863/252
No. of terms in the intersection of three classes	3647	0/0/189/2/211/291/0/146/281/206	1385
No. of terms in the intersection of four classes	NA	0/169/357/267/215	NA
No. of terms in the intersection of five classes	NA	1658	NA

$$r_k = \frac{TP_k}{TP_k + FP_k} \tag{20}$$

$$p_k = \frac{TP_k}{TP_k + FN_k} \tag{21}$$

where r_k is recall and p_k is precision of the class k, calculated from TP_k (true positive of the class k) and FP_k (false positive of the class k). While the measure A is used to evaluate all classes as one set, $\bar{F}$ assess the performance of each class separately and then combine the performances by averaging. These two measures have different evaluation properties and they can be complementary to each other. Therefore, their geometric mean (GM) is proposed to be the unified measure.

$$GM = \sqrt{A \times \bar{F}} \tag{22}$$

6. Experiment Results

6.1 Effect of Single Weighting Scheme

This first experiment surveys the outcome of an individual distribution factor on clustering quality by adding a term distribution-based weight (DW) one by one, where the weight is a deviation-based factor (either sd, acsd, or icsd) or an entropy-based factor (either e, cce, or ce) to the frequency-based weight (FW) as shown in Table 2. Note that the frequency-based weight(FW) is ntf × idf schemes, the norm-1 of term frequency (ntf) defined by Eq. (1). The quality of cluster evaluation is performed on seeded clustering with five-fold cross validation, where 80% of the data is used for training of distribution term weighting calculation and remaining 20% is used for testing. The result shows the distribution term factor of predefined classes that are correlated to the effectiveness of factors. The **Avg.** column shows the average performance of three datasets. It is used to in-

Table 2 Cluster quality in the form of geo-mean (accuracy, f-measure), when seeded k-means clustering with a single term distribution factor is used. (Panel I for deviation-based and Panel II for entropy-based)

TW			Amazon	WebKB	Thai-reform	Avg.
FW	⊙	DW				
ntf × idf	-	-	90.35 *(90.35,90.35)*	68.51 *(71.27,65.84)*	93.01 *(92.90,93.12)*	83.96 *(84.84,83.10)*
Panel I : Deviation-based weighting						
ntf × idf	/	sd	90.99 *(91.00,90.98)*	87.30 *(90.41,84.30)*	93.16 *(93.10,93.22)*	**90.48 *(91.50,89.50)***
ntf × idf	×	sd	72.44 *(71.13,73.77)*	31.41 *(38.61,25.55)*	54.04 *(52.70,55.42)*	52.63 *(54.15,51.58)*
ntf × idf	/	acsd	91.71 *(91.72,91.70)*	83.77 *(86.82,80.84)*	94.05 *(94.00,94.10)*	**89.84 *(90.85,88.88)***
ntf × idf	×	acsd	71.12 *(70.23,72.02)*	31.42 *(37.94,26.02)*	50.11 *(49.47,50.75)*	50.88 *(52.55,49.60)*
ntf × idf	/	icsd	52.68 *(53.60,51.78)*	63.15 *(65.31,61.06)*	70.62 *(71.57,69.69)*	62.15 *(63.49,60.84)*
ntf × idf	×	icsd	75.12 *(74.07,76.18)*	67.98 *(64.20,71.99)*	78.05 *(76.30,79.84)*	**73.72 *(71.52,76.00)***
Panel II : Entropy-based weighting						
ntf × idf	/	e	88.09 *(88.10,88.08)*	79.40 *(84.45,74.66)*	71.37 *(69.60,73.18)*	**79.62 *(80.72,78.64)***
ntf × idf	×	e	69.68 *(69.17,70.20)*	31.14 *(35.21,27.54)*	83.39 *(82.57,84.23)*	61.40 *(62.32,60.66)*
ntf × idf	/	cce	88.56 *(88.57,88.55)*	78.94 *(81.45,76.51)*	75.74 *(74.13,77.39)*	**81.08 *(81.38,80.82)***
ntf × idf	×	cce	64.45 *(64.08,64.81)*	30.01 *(34.00,26.49)*	74.28 *(72.97,75.62)*	56.25 *(57.02,55.64)*
ntf × idf	/	ce	91.15 *(91.15,91.14)*	75.71 *(77.29,74.16)*	93.49 *(93.40,93.58)*	**86.78 *(87.28,86.29)***
ntf × idf	×	ce	89.03 *(89.02,89.04)*	57.10 *(61.25,53.23)*	91.50 *(91.33,91.67)*	79.21 *(80.53,77.98)*

TW = FW⊙DW where TW = term weight, FW = frequency-based weight, DW = distribution-based weight.
⊙ is an element-wise multiplication.

vestigate the direction of the exponent of each term. The bold numbers in the table indicate the maximum geo-mean of accuracy and the f-measure value for the degree of exponent in each distribution term factor.

In the deviation-based model, '**acsd**', and '**sd**' are likely to be a demoter (/), that is they have better performance when they are assigned with a negative degree of exponent as shown in Table 2 (Panel I). On the other hand, '**icsd**' tends to act as a promoter (×) since its positive degree of exponent outperforms its negative version. Moreover, all entropy-based types ('**e**', '**cce**', '**ce**') work well as a demotor (/) since they perform well when they are assigned with a negative degree of exponent, as shown in Table 2 (Panel II). One more observation is that clustering quality is quite different between demoter (/) and promoter (×) for the deviation-based weighting. The most influential weight is '**acsd**', where the gap between its demotor version and its promotor version is quite large, i.e. 20.59% for Amazon, 52.35% for WebKB, and 43.94% for Thai-reform. The second most influential weight is '**sd**' with a big gap, i.e., 18.54% for Amazon, 55.89% for WebKB, and 39.12% for Thai-reform. The least influential weight is '**icsd**', which has a relatively smaller gap between its promotor version and its demotor version, i.e., 22.44% for Amazon, 4.83% for WebKB, and 7.43% for Thai-reform). The result implies that the deviation-based weighting affects clustering quality in the order of intra-class, collection, and inter-class factors. Similar to the deviation-based weighting, the entropy-based weighting can be used to improve clustering quality, the result shows that '**cce**' and '**ce**' should act as a demoter. Although '**e**' seems work as a demotor but the reversed affect is observed for the Thai-reform dataset. Recall Table 1, we can find that the number of class-overlapping terms in the Thai-reform dataset is lower, compared to the other two datasets, that is 76.16% (2,703 from 3,549 terms) while that of Amazon is 90.15% (6,864 in 7,614), and that of WebKB is 84.54% (5,518 in 6,527). Therefore, this different distribution may affect the role of the in-collection term entropy ('**e**'). Moreover, the Amazon has 2,722 two-class overlapping words and 3,647 three-class overlapping words (see the "Distribution Information" section in Table 1). That is, the Amazon includes a lot of overlapping words, compared to the other two. The result implies that the entropy-based weighting affects clustering quality in the order of intra-class, collection, and inter-class factors. In other words, due to the comparison between deviation-based and entropy-based weight, both intra-class factor and collection class factors have the best effect on deviation-based weight scheme. In contrast, the inter-class factor is preferred on entropy-based weight more than deviation-based weight. Although, some single term distribution factors can slightly be better than the base-line (ntf × idf), but on average unfortunately some factors can not be improved or may not show the obvious direction. Therefore, to overcome this issue, the cross multiple distribution factors in the same distribution are shown in Table 3 and Table 4.

6.2 Effect of Multiple Weighting Scheme

This second experiments are explored with 5 different exponents (−1.0, −0.5, 0, 0.5, 1.0) giving produce 125 patterns for each weighting based type. The results confirm the role of each factor (i.e., sd, acsd, icsd, e, cce, and ce) giving the same solution of analysis for single-factor effects. Moreover, the combination pattern gives better performance than using a single factor. There exist 24 patterns for standard deviation-based and 13 patterns for entropy-based models in seeded k-means clustering compared with the base-line. For the standard deviation-based experiment (best term distribution on average) the appropriate exponents are −0.5, −1, and 0.5 for '**sd**', '**acsd**', and '**icsd**', respectively. For the entropy-based model, the appropriate exponents are 1, −1, and −1 for '**e**', '**cce**', and '**ce**', respectively. The dataset with a short sentence (Thai-reform) tends to have a high variance between classes. Note that the short sentence has a high percentage between the number of terms in a single class and the number of distinct terms in a document, as shown in Table 1. The variance value between a class can be arranged from high to low as Thai-reform, Amazon, and WebKB, respectively. One more observation is that the best-10 of ranking used the high degree of the exponent to promote or demote term in order as '**acsd**', '**sd**', '**icsd**', respectively. It follows the result of a survey in Table 2 that the class information can be represented by a weight term preferred on acsd. It is clear that in our proposed scheme both deviation- and entropy- based term weighting give the effect that follows the quality of a class information. The suggested combinations (DTW1, ETW1) are shown with a detailed performance comparison with four combined weighting schemes in the next following experiment.

6.3 Combined Weighting Scheme

This third experiment investigates the effect of combined weighting that follows the same five-fold cross-validation of the previous experiment (Table 3 and Table 4). The two styles of distribution-based factors are attached to the frequency-based component (FW) that is the adaptive distribution term weighting on deviation-based (DTW) and entropy-based (ETW) models. To evaluate this combined weighting scheme two sub-experiments are performed. The first sub-experiment investigates the effect of both best distribution term weightings (i.e., DTW1 and ETW1) with frequency term weighting (ntf×idf) on both sides of promoter (for multiplier) and demoter (for divider). The second sub-experiment is performed to analyze the effect of combined six distribution weighting i.e., sd, acsd, icsd, e, cce, and ce on different exponent of term weighting.

6.3.1 Analysis on Paired Comparative Study (DTW vs. ETW)

The experiments explored various different exponents to

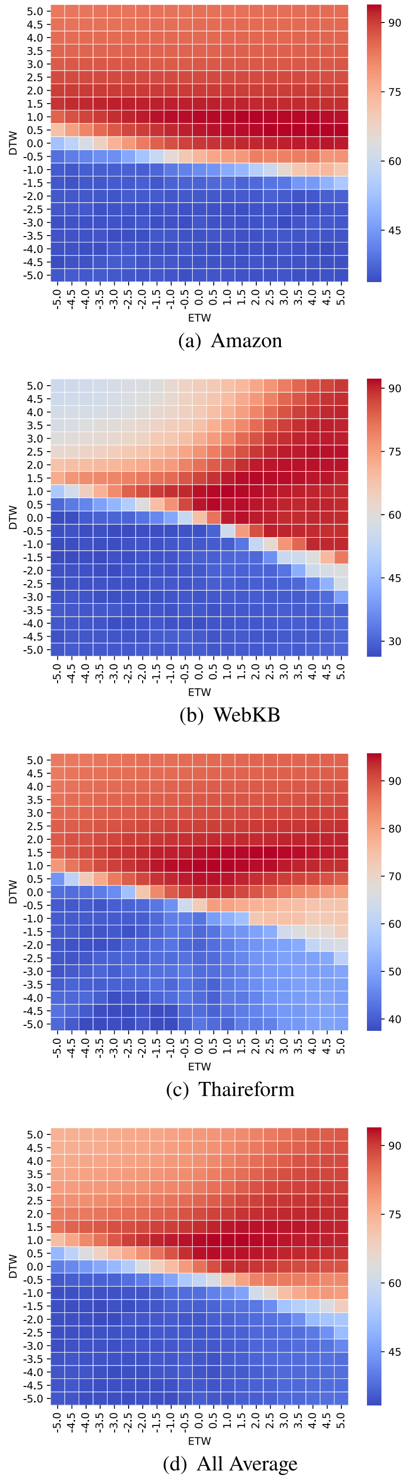
Table 3 Analysis of adaptive distribution term based on deviation (validation (panel I) best-10 and (panel II) worst-10), by seeded k-means with multiple distribution factors.

Method	Exponent of DW			Amazon	WebKB	Thai-reform	Avg.
	sd	acsd	icsd				
B-DTW24	0	0	0	90.35 *(90.35,90.35)*	68.51 *(71.27,65.84)*	93.01 *(92.90,93.12)*	83.96 *(84.84,83.10)*
Panel I : Best-10							
DTW1	-0.5	-1	0.5	**93.63 *(93.63,93.62)***	91.16 *(93.29,89.08)*	**95.56 *(95.53,95.59)***	**93.45 *(94.15,92.76)***
DTW2	0.5	-1	0	93.21 *(93.22,93.21)*	**91.49 *(93.58,89.46)***	95.23 *(95.20,95.27)*	93.31 *(94.00,92.65)*
DTW3	-1	-0.5	0.5	93.38 *(93.38,93.38)*	90.37 *(92.67,88.12)*	95.40 *(95.37,95.43)*	93.05 *(93.81,92.31)*
DTW4	0	-0.5	0	92.63 *(92.63,92.62)*	90.55 *(92.91,88.25)*	95.11 *(95.07,95.15)*	92.76 *(93.54,92.01)*
DTW5	0	-1	0.5	92.74 *(92.73,92.74)*	89.55 *(91.88,87.28)*	94.07 *(94.00,94.13)*	92.12 *(92.87,91.38)*
DTW6	-0.5	-0.5	0.5	92.28 *(92.28,92.28)*	88.98 *(91.47,86.55)*	93.87 *(93.80,93.95)*	91.71 *(92.52,90.93)*
DTW7	-1	0	0.5	91.88 *(91.88,91.88)*	88.68 *(91.25,86.18)*	94.00 *(93.93,94.07)*	91.52 *(92.35,90.71)*
DTW8	-1	-1	1	92.66 *(92.67,92.66)*	86.71 *(89.33,84.17)*	94.47 *(94.43,94.50)*	91.28 *(92.14,90.44)*
DTW9	-0.5	-0.5	0	91.51 *(91.52,91.50)*	87.96 *(90.77,85.24)*	93.58 *(93.53,93.64)*	91.02 *(91.94,90.13)*
DTW10	-0.5	0	0	92.06 *(92.07,92.05)*	85.71 *(88.30,83.21)*	94.55 *(94.50,94.60)*	90.77 *(91.62,89.95)*
Panel II : Worst-10							
DTW116	0.5	0.5	-1	36.47 *(36.57,36.37)*	35.68 *(43.06,29.57)*	63.07 *(62.07,64.08)*	45.07 *(47.23,43.34)*
DTW117	0.5	0.5	-0.5	39.49 *(40.58,38.43)*	27.48 *(31.72,23.81)*	67.26 *(67.70,66.83)*	44.74 *(46.67,43.02)*
DTW118	0	1	-1	36.31 *(35.60,37.04)*	33.07 *(39.52,27.67)*	61.85 *(60.80,62.92)*	43.74 *(45.31,42.54)*
DTW119	0	1	-0.5	38.50 *(39.58,37.45)*	26.89 *(31.09,23.25)*	62.12 *(62.70,61.54)*	42.50 *(44.46,40.75)*
DTW120	1	0.5	-1	35.95 *(34.82,37.12)*	26.66 *(31.49,22.57)*	50.69 *(51.47,49.92)*	37.77 *(39.26,36.54)*
DTW121	1	0.5	-0.5	38.19 *(38.40,37.97)*	29.70 *(36.91,23.90)*	45.07 *(45.80,44.36)*	37.65 *(40.37,35.41)*
DTW122	0.5	1	-1	35.26 *(34.93,35.58)*	27.35 *(31.63,23.65)*	50.03 *(50.83,49.25)*	37.55 *(39.13,36.16)*
DTW123	1	1	-0.5	38.35 *(38.25,38.46)*	28.40 *(33.69,23.94)*	42.91 *(43.80,42.03)*	36.55 *(38.58,34.81)*
DTW124	0.5	1	-0.5	38.10 *(38.18,38.02)*	28.40 *(35.17,22.94)*	42.94 *(43.83,42.07)*	36.48 *(39.06,34.34)*
DTW125	1	1	-1	34.89 *(34.48,35.29)*	27.55 *(32.70,23.21)*	42.18 *(41.13,43.25)*	34.87 *(36.10,33.92)*

Table 4 Analysis of adaptive distribution term based on entropy (validation (panel I) best-10 and (panel II) worst-10), by seeded k-means with multiple distribution factors.

Method	Exponent of DW			Amazon	WebKB	Thai-reform	Avg.
	e	cce	ce				
B-ETW13	0	0	0	90.35 *(90.35,90.35)*	68.51 *(71.27,65.84)*	93.01 *(92.90,93.12)*	83.96 *(84.84,83.10)*
Panel I : Best-10							
ETW1	1	-1	-1	**91.39 *(91.38,91.40)***	**90.84 *(93.37,88.38)***	**93.50 *(93.34,93.67)***	**91.91 *(92.70,91.15)***
ETW2	1	-1	-0.5	91.07 *(91.07,91.08)*	90.48 *(93.13,87.90)*	91.87 *(91.67,92.06)*	91.14 *(91.96,90.35)*
ETW3	0.5	-0.5	-1	91.35 *(91.35,91.35)*	88.17 *(91.01,85.42)*	93.15 *(93.03,93.27)*	90.89 *(91.80,90.01)*
ETW4	1	-1	0	90.70 *(90.68,90.71)*	90.26 *(92.98,87.61)*	91.45 *(91.23,91.68)*	90.80 *(91.63,90.00)*
ETW5	1	-1	0.5	90.15 *(90.13,90.17)*	87.28 *(89.75,84.89)*	90.82 *(90.57,91.07)*	89.42 *(90.15,88.71)*
ETW6	0.5	-0.5	-0.5	91.04 *(91.03,91.04)*	81.14 *(83.09,79.22)*	92.87 *(92.73,93.01)*	88.35 *(88.95,87.76)*
ETW7	0.5	-0.5	0	90.47 *(90.47,90.48)*	78.75 *(80.33,77.21)*	92.45 *(92.30,92.60)*	87.22 *(87.70,86.76)*
ETW8	1	-1	1	89.27 *(89.23,89.31)*	81.64 *(83.30,80.01)*	90.11 *(89.80,90.43)*	87.01 *(87.44,86.58)*
ETW9	0	0	-1	91.15 *(91.15,91.14)*	75.71 *(77.29,74.16)*	93.49 *(93.40,93.58)*	86.78 *(87.28,86.29)*
ETW10	0	0	-0.5	90.67 *(90.67,90.67)*	73.51 *(75.19,71.85)*	93.27 *(93.17,93.37)*	85.82 *(86.34,85.30)*
Panel II : Worst-10							
ETW116	0.5	1	0	57.05 *(56.95,57.14)*	28.68 *(32.11,25.62)*	67.90 *(66.33,69.51)*	51.21 *(51.80,50.76)*
ETW117	1	1	-1	53.90 *(53.85,53.95)*	28.37 *(30.95,26.00)*	70.67 *(69.13,72.24)*	50.98 *(51.31,50.73)*
ETW118	1	0.5	1	55.16 *(55.10,55.22)*	26.26 *(29.03,23.76)*	67.82 *(66.23,69.44)*	49.75 *(50.12,49.47)*
ETW119	0	1	1	59.37 *(59.23,59.50)*	29.12 *(32.20,26.34)*	60.64 *(59.73,61.57)*	49.71 *(50.39,49.14)*
ETW120	1	1	-0.5	52.96 *(52.95,52.98)*	26.44 *(29.34,23.84)*	65.59 *(64.13,67.07)*	48.33 *(48.81,47.96)*
ETW121	0.5	1	0.5	55.17 *(55.12,55.23)*	28.06 *(31.22,25.23)*	57.59 *(56.57,58.63)*	46.94 *(47.64,46.36)*
ETW122	1	1	0	51.83 *(51.80,51.86)*	27.04 *(29.58,24.71)*	58.46 *(57.37,59.57)*	45.78 *(46.25,45.38)*
ETW123	0.5	1	1	53.22 *(53.18,53.26)*	26.29 *(28.48,24.28)*	53.68 *(53.20,54.16)*	44.40 *(44.95,43.90)*
ETW124	1	1	0.5	50.56 *(50.53,50.58)*	27.15 *(29.92,24.64)*	53.73 *(53.23,54.23)*	43.81 *(44.56,43.15)*
ETW125	1	1	1	46.20 *(46.25,46.15)*	27.43 *(30.61,24.57)*	50.93 *(50.53,51.33)*	41.52 *(42.46,40.68)*

combine the two types of distribution term weighting schemes, i.e., distribution-based, and entropy-based with frequency-based term weighting. The best weighting of each distribution term weighting (refer to Table 3 and Table 4; DTW1 and ETW1) are investigated in place of the comparison combined between various exponent values. The exponent of each parameter (i.e., $\mathrm{dtw}_n=\alpha^1$, $\mathrm{etw}_n = \alpha^2$) varies between -5 to 5 with a step size of 0.5 and the total possible combinations (paired of weighting) are 441. In Fig. 3, each sub-figure shows the performance by average geo-mean of accuracy and f-measure when the positive number of exponent (for promoter) in both of distribution terms are affected for clustering quality. In the contrast, the negative number of exponent (for demoter) in both of distri-



(a) Amazon

(b) WebKB

(c) Thaireform

(d) All Average

Fig. 3 Effects of component size (α^1, α^2 by Eq. (14)) where paired comparative studied (best deviation-based (DTW1) vs. best entropy-based (ETW1) weighting), in term of average geo-mean of accuracy and f−measure: (a) amazon, (b) webkb, (c) Thai-reform, and (d) all average.

bution terms are indicated for low clustering quality. However, Amazon and Thai-reform have slightly high clustering performance, although ETW exponents are in negative side and DTW exponents are in positive side. This means the deviation-based distribution outperforms the entropy-based distribution. For all datasets, the most best exponent is the number one (α^1 = 1, α^2 =1) (i.e., 93.6% GM for all average, 93.65% GM for Amazon, 92.18% GM for WebKB, and 93.50% GM for Thai-reform). Specially for Amazon, there exist 8 patterns which have higher GM values more than the overview of best exponent (α^1 = 1, α^2 = 1), when DTW is in the range of 0.5-1 and the exponent of ETW is in the range of 2-5. Here, the highest clustering quality is 93.89% GM, when α^1 = 0.5 for DTW, and α^2 = 3 for ETW. Figure 3 (d) shows as the baseline to declare that both DTW and ETW factors have best effect on promoting side. Especially, DTW can enhance the performance of clustering process than ETW.

6.3.2 Analysis of All Combinations of Six Factors

While the results of the first and the second experiment survey the promoting/demoting role of six factors in DTW (refer to Table 3) and ETW (refer to Table 4), respectively. In the last experiment, the overall performance of a combination of six factors is examined to investigate the optimal combination of factors for discovering the threshold of combined distribution term weighting. The exponent of each parameter (i.e., '**sd**' = β^1), '**acsd**' = β^2, '**icsd**' = β^3, '**e**' = β^4, '**cce**' = β^5, and '**ce**' = β^6) are varied between -1 and 1 where step size of 0.5. The total combination is 15625. Table 5 shows the geo-mean of accuracy and f-measure values of quality seeded k-means on the combination of six factors of weighting; including, '**sd**', '**acsd**', and '**icsd**' for DTW, and '**e**', '**cce**', and '**ce**' for ETW. Based on the average GM on three datasets, Panel I shows the best-10 combination (weighting) and Panel II shows the worst-10 combination (weighting). The results implies that for six factors '**acsd**', '**cce**', and '**ce**' work well as demoter due to most of best-10 weighting have negative exponent for them. In the contrast, '**icsd**' and '**e**' act as demoter since most of the combinations have positive exponents and all of combination in the case of '**icsd**' have negative exponents. On the other hand, sd works rather well on not promoting/demoting. Further, we can conclude 2297 from 15625 combinations (weighting) are superior to the baseline for the seeded k-means algorithm. Finally, the best weighting to baseline is with a gap of 10.35% GM (vary on 3.48 % GM for Amazon, 25.30% GM for WebKB, and 2.20% GM for Thai-reform).

7. Discussion and Conclusion

This paper presented a method to improve seeded k-means clustering using deviation- and entropy-based schemes on term weightings. Both schemes utilized in-collection, intraclass, and inter-class distribution as constraints to guide clustering towards user intention. As the preliminary experiments, we have investigated our method on three text datasets, i.e., two English and one Thai texts. As the experimental result on the deviation-based factors, we find out that '**sd**' and '**acsd**' are negative factors in term weighting (i.e.,

Table 5 Analysis of adaptive distribution term based on deviation-based and entropy-based (validation (panel I) best-10 and (panel II) worst-10), by seeded k-means with multiple distribution factors.

Method	Exponent of DW						Amazon	WebKB	Thai-reform	Avg.
	sd	acsd	icsd	e	cce	ce				
B-TW2297	0	0	0	0	0	0	90.35 *(90.35,90.35)*	68.51 *(71.27,65.84)*	93.01 *(92.90,93.12)*	83.96 *(84.84,83.10)*
Panel I : Best-10										
TW1	0.5	-1	0	1	-1	-0.5	93.83 *(93.83,93.83)*	**93.81 *(95.51,92.14)***	95.30 *(95.27,95.34)*	94.31 *(94.87,93.77)*
TW2	0.5	-1	0	1	-1	-1	93.76 *(93.77,93.76)*	93.66 *(95.39,91.97)*	95.50 *(95.47,95.53)*	**94.31 *(94.88,93.75)***
TW3	0.5	-1	0	0.5	-0.5	-0.5	93.58 *(93.58,93.57)*	93.12 *(94.97,91.30)*	95.60 *(95.57,95.63)*	94.10 *(94.71,93.50)*
TW4	0.5	-1	0	1	-1	0	93.65 *(93.65,93.64)*	93.44 *(95.19,91.73)*	95.20 *(95.17,95.24)*	94.10 *(94.67,93.54)*
TW5	0	-1	0.5	0.5	-1	-1	93.30 *(93.30,93.29)*	92.91 *(94.83,91.02)*	**96.06 *(96.03,96.08)***	94.09 *(94.72,93.46)*
TW6	0	-1	0.5	0.5	-1	-0.5	93.46 *(93.47,93.46)*	92.88 *(94.81,90.99)*	95.92 *(95.90,95.95)*	94.09 *(94.73,93.47)*
TW7	0	-1	0.5	0.5	-1	0	93.53 *(93.53,93.53)*	92.84 *(94.76,90.95)*	95.89 *(95.87,95.92)*	94.09 *(94.72,93.47)*
TW8	0	-1	0.5	0.5	-1	0.5	93.68 *(93.68,93.68)*	92.61 *(94.57,90.69)*	95.83 *(95.80,95.85)*	94.04 *(94.68,93.41)*
TW9	0.5	-1	0	0.5	-0.5	-1	93.53 *(93.53,93.52)*	92.85 *(94.76,90.98)*	95.70 *(95.67,95.72)*	94.03 *(94.65,93.41)*
TW10	0	-1	0.5	0.5	-1	1	93.78 *(93.78,93.78)*	92.47 *(94.45,90.54)*	95.56 *(95.53,95.59)*	93.94 *(94.59,93.30)*
TW247	-0.5	-1	1	-0.5	0	-1	**94.08 *(94.08,94.08)***	87.8 *(90.31,85.37)*	95.26 *(95.23,95.29)*	92.38 *(93.21,91.58)*
Panel II : Worst-10										
TW15616	0.5	0.5	-1	1	1	-0.5	35.90 *(35.83,36.11)*	23.79 *(25.65,22.06)*	38.85 *(39.63,38.08)*	32.85 *(33.70,32.08)*
TW15617	0.5	0.5	-1	1	1	1	35.86 *(35.55,36.82)*	24.13 *(25.79,22.59)*	38.53 *(39.30,37.78)*	32.84 *(33.55,32.40)*
TW15618	0	1	-1	1	0	0	35.07 *(34.88,35.64)*	25.89 *(28.77,23.30)*	37.48 *(38.00,36.97)*	32.81 *(33.88,31.97)*
TW15619	1	0	-1	1	1	-0.5	35.44 *(35.35,35.73)*	23.94 *(25.69,22.32)*	39.06 *(39.83,38.30)*	32.81 *(33.62,32.12)*
TW15620	1	0.5	-1	0.5	0.5	1	35.05 *(34.93,35.39)*	25.02 *(27.97,22.38)*	38.33 *(39.23,37.46)*	32.80 *(34.04,31.74)*
TW15621	0.5	0.5	-1	1	0.5	1	35.42 *(35.35,35.65)*	24.08 *(26.27,22.07)*	38.87 *(39.67,38.09)*	32.79 *(33.76,31.94)*
TW15622	0	0.5	-1	1	1	-1	35.55 *(35.35,36.17)*	25.50 *(28.65,22.69)*	37.24 *(38.10,36.39)*	32.76 *(34.03,31.75)*
TW15623	-0.5	1	-1	1	1	-1	35.79 *(35.58,36.41)*	25.97 *(28.60,23.57)*	36.38 *(37.47,35.33)*	32.71 *(33.88,31.77)*
TW15624	0.5	0	-1	1	1	-1	35.76 *(35.55,36.39)*	24.81 *(26.78,22.98)*	37.31 *(38.23,36.42)*	32.63 *(33.52,31.93)*
TW15625	1	-0.5	-1	1	1	-1	35.32 *(35.13,35.88)*	24.53 *(26.42,22.78)*	37.72 *(38.60,36.87)*	32.52 *(33.38,31.84)*

demotor) but ‘**icsd**’ is a positive factor in term weighting (i.e., promoter). As an alternative for the entropy-based factors, we find out that ‘**e**’ is negative factors in term weighting (i.e., demotor) for using as single factor but ‘**e**’ is positive factor (i.e., promoter) when combined with others. Morover, ‘**cce**’, and ‘**ce**’ are all negative factors in term weighting (i.e., demotor). The result shows that the deviation-based distribution outperformed the entropy-based distribution, and a suitable combination of all distribution weights increases the clustering accuracy by 10% (2.20% to 25.30%), compared to the baseline, i.e., the pure frequency factor (FW: ntf × idf). With the deviation- and entropy-based term weighting, the reconstructed new vector has the potential to control/guide the document clustering process towards expected results with consideration of term distribution and ambiguity among classes. As our future work, we plan to examine the proposed method on various text datasets to check its validity in general context. Moreover, it is worth incorporating dimensionality reduction into the framework and investigating other clustering algorithms (such as hierarchical clustering, density-based clustering, grid-based clustering, etc.) to examine the effectiveness of the deviation- and entropy-based term weighting on clustering quality. We also aim to apply the proposed term weighting on the deep learning framework, which is a recent popular method.

Acknowledgments

This research work is partially supported by the Sirindhorn International Institute of Technology (SIIT), Thammasat University, Center of Excellence in Intelligent Informatics, Speech and Language Technology and Service Innovation (CILS), Intelligent Informatics and Service Innovation (IISI) Research Center, National Electronics and Computer Technology Center (NECTEC) of Thailand for the STEM Workforce, as well as the Thailand Research Fund (TRF) under grant number RTA-6080013, RTA6280015. The corresponding author is also thankful to Burapha University (Chanthaburi Campus) who provided financial assistance every month.

References

[1] G. Domeniconi, G. Moro, R. Pasolini, and C. Sartori, “A study on term weighting for text categorization: A novel supervised variant of tf. idf,” Proceedings of 4th International Conference on Data Management Technologies and Applications, pp.26–37, 2015.

[2] Y. Ko, “A new term-weighting scheme for text classification using the odds of positive and negative class probabilities,” Journal of the Association for Information Science and Technology, vol.66, no.12, pp.2553–2565, 2015.

[3] K. Sparck Jones, “A statistical interpretation of term specificity and its application in retrieval,” Journal of Documentation, vol.28, no.1, pp.11–21, 1972.

[4] R. Cummins and C. O’riordan, “Evolving general term-weighting schemes for information retrieval: Tests on larger collections,” Artificial Intelligence Review, vol.24, no.3-4, pp.277–299, 2005.

[5] K. Wagstaff, C. Cardie, S. Rogers, S. Schrödl, et al., “Constrained k-means clustering with background knowledge,” Proceedings of the 18th International Conference on Machine Learning, pp.577–584, 2001.

[6] M. Bilenko, S. Basu, and R.J. Mooney, “Integrating constraints and metric learning in semi-supervised clustering,” Proceedings of the Twenty-first International Conference on Machine Learning, p.11, ACM, 2004.

[7] H. Zhang, S. Basu, and I. Davidson, “Deep constrained clustering-

algorithms and advances," arXiv preprint arXiv:1901.10061, 2019.

[8] S. Basu, A. Banerjee, and R.J. Mooney, "Semi-supervised clustering by seeding," Proceedings of the 19th International Conference on Machine Learning, ICML '02, pp.27–34, 2002.

[9] 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," Proceedings of the 19th International Conference on Machine Learning, ICML '02, pp.307–314, 2002.

[10] S. Zhu, D. Wang, and T. Li, "Data clustering with size constraints," Knowledge-Based Systems, vol.23, no.8, pp.883–889, 2010.

[11] N. Ganganath, C.-T. Cheng, and C.K. Tse, "Data clustering with cluster size constraints using a modified k-means algorithm," 2014 International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery, pp.158–161, IEEE, 2014.

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

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

[14] R. Zhu and J.-H. Xue, "On the orthogonal distance to class subspaces for high-dimensional data classification," Information Sciences, vol.417, pp.262–273, 2017.

[15] C. Largeron, C. Moulin, and M. Géry, "Entropy based feature selection for text categorization," Proceedings The Symposium on Applied Computing, pp.924–928, ACM, 2011.

[16] J. Chai, Z. Chen, H. Chen, and X. Ding, "Designing bag-level multiple-instance feature-weighting algorithms based on the large margin principle," Information Sciences, vol.367, pp.783–808, 2016.

[17] N. Kittiphattanabawon, T. Theeramunkong, and E. Nantajeewarawat, "News relation discovery based on association rule mining with combining factors," IEICE Transactions on Information and Systems, vol.E94-D, no.3, pp.404–415, 2011.

[18] U. Buatoom, W. Kongprawechnon, and T. Theeramunkong, "Constrained clustering with seeds and term weighting scheme," IEEE Knowledge, Information and Creativity Support Systems, pp.116–121, 2018.

[19] F. Ren and M.G. Sohrab, "Class-indexing-based term weighting for automatic text classification," Information Sciences, vol.236, pp.109–125, 2013.

[20] Q. Hu, X. Che, L. Zhang, D. Zhang, M. Guo, and D. Yu, "Rank entropy-based decision trees for monotonic classification," IEEE Transactions on Knowledge and Data Engineering, vol.24, no.11, pp.2052–2064, 2012.

[21] A. Irpino, R. Verde, and F.d.A.T. de Carvalho, "Fuzzy clustering of distributional data with automatic weighting of variable components," Information Sciences, vol.406, pp.248–268, 2017.

[22] T.-P. Lo and S.-J. Guo, "Effective weighting model based on the maximum deviation with uncertain information," Expert Systems with Applications, vol.37, no.12, pp.8445–8449, 2010.

[23] M.A. Fattah, "New term weighting schemes with combination of multiple classifiers for sentiment analysis," Neurocomputing, vol.167, pp.434–442, 2015.

[24] K. Nigam, J. Lafferty, and A. McCallum, "Using maximum entropy for text classification," Workshop on Machine Learning for Information Filtering (IJCAI), vol.1, pp.61–67, 1999.

[25] W. Pan and Q. Hu, "An improved feature selection algorithm for ordinal classification," IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences, vol.E99-A, no.12, pp.2266–2274, 2016.

[26] K. Zheng and X. Wang, "Feature selection method with joint maximal information entropy between features and class," Pattern Recognition, vol.77, pp.20–29, 2018.

[27] K. Fragos, Y. Maistros, and C. Skourlas, "A weighted maximum entropy language model for text classification.," NLUCS, pp.55–67, 2005.

[28] H. Wu, X. Gu, and Y. Gu, "Balancing between over-weighting and under-weighting in supervised term weighting," Information Processing & Management, vol.53, no.2, pp.547–557, 2017.

[29] J.H. Lee, S.H. Jung, and J. Park, "The role of entropy of review text sentiments on online WOM and movie box office sales," Electronic Commerce Research and Applications, vol.22, pp.42–52, 2017.

[30] V. Lertnattee and T. Theeramunkong, "Effects of term distributions on binary classification," IEICE Transactions on Information and Systems, vol.E90-D, no.10, pp.1592–1600, 2007.

[31] V. Lertnattee and T. Theeramunkong, "Class normalization in centroid-based text categorization," Information Sciences, vol.176, no.12, pp.1712–1738, 2006.

[32] S. Basu, M. Bilenko, and R.J. Mooney, "A probabilistic framework for semi-supervised clustering," Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining, pp.59–68, ACM, 2004.

[33] A. Moreo Fernandez, A. Esuli, and F. Sebastiani, "Learning to Weight for Text Classification," IEEE Transactions on Knowledge and Data Engineering, vol.32, pp.302–316, 2018.

Uraiwan Buatoom received a B.Sc. in Computer Science and M.Sc. in Information Technology from Burapha University, in 2003 and 2008, respectively. Currently, she is a doctoral candidate in the Information Technology program at Sirindhorn International Institute of Technology (SIIT), Thammasat University, Thailand. Her research interests include data mining, clustering, and knowledge discovery.

Waree Kongprawechnon received a B.Eng. degree in Electrical Engineering from Chulalongkorn University, Thailand, in 1992; and an M.Eng. in Control Engineering from Osaka University, Japan, in 1995 and Ph.D. in Mathematical Engineering and Information Physics from the University of Tokyo, Japan, in 1998. She is an Associate Professor at SIIT, Thammasat University, Thailand. Her research interests include H^∞ control, control theory, robust control, system identification, modeling, adaptive control, learning control, neural networks and fuzzy control.

Thanaruk Theeramunkong received a B.Eng. in Electric and Electronics Engineering, and an M.Eng. and Ph.D. in Computer Science from the Tokyo Institute of Technology in 1990, 1992 and 1995, respectively. Now, he is serving as a Professor at SIIT, Thammasart University, and Associate Fellow at the Royal Society of Thailand. His research interests include data mining, machine learning, natural language processing, and knowledge engineering.