Coupled Term-Term Relation Analysis for Document Clustering

Xin Cheng, Duoqian Miao, Can Wang, Longbing Cao

***Abstract*— Traditional document clustering approaches are usually based on the Bag of Words model, which is limited by its assumption of the independence among terms. Recent strategies have been proposed to capture the relation between terms based on statistical analysis, and they estimate the relation between terms purely by their co-occurrence across the documents. However, the implicit interactions with other link terms are overlooked, which leads to the discovery of incomplete information. This paper proposes a coupled term-term relation model for document representation, which considers both the intra-relation (i.e. co-occurrence of terms) and inter-relation (i.e. dependency of terms via link terms) between a pair of terms. The coupled relation for each pair of terms is further used to map a document onto a new feature space, which includes more semantic information. Substantial experiments verify that the document clustering incorporated with our proposed relation achieves a significant performance improvement compared to the state-of-the-art techniques.**

I. Introduction

DOCUMENT clustering is an unsupervised technique to automatically organize documents into a list of meaningful categories based on their similarity. It aims at discovering the natural groupings of the document set to facilitate the document organization and navigation. In general, document clustering is divided into three consecutive steps: document representation, similarity calculation, and clustering analysis. Among them, building a high-quality representation for documents is a fundamental but critical step for the document clustering [1].

The classical representation of document, the Bag of Words (BOW) model, is to construct a feature space that contains all the distinct terms (words) in the document set. Each document is represented by a vector, whose component reflects the weight (usually tf-idf, term-frequencies and inverse document-frequencies) of every distinct term. The BOW model has been widely used in various approaches because of its simplicity and general applicability. However, it has the limitation that it assumes the terms are independent and totally ignores the semantic relation between them. Thus, it fails to assign the documents with a similar topic but described by different words into the same clusters. A simple example of the BOW document representation is shown in Table I, where each value is the tf-idf weight of every term in its corresponding document. There are three documents in the document set. The first document d_1 describes the concept of "*Data mining*", the second document d_2 discusses "*Machine learning*", and the third one d_3 talks about "*Database management*". As we all know, the similarity between documents d_1 and d_2 is much higher than between d_1 and d_3. However, with the BOW representation, the cosine similarity between d_1 and d_2 is 0.253, and 0.231 for d_1 and d_3. The similarity values are approximate, thus, it is unable to identify which two documents are more alike if the relation between terms is not captured.

TABLE I

An Example of Document Representation: "*DM*","*ML*","*DB*" and "*CS*" denote "*Data mining*", "*Machine learning*", "*Database*" and "*Computer science*", respectively.

	DM	ML	DB	CS
d_1	0.5	0.0	0.1	0.3
d_2	0.0	0.5	0.1	0.25
d_3	0.0	0.0	0.8	0.1

To address the absence of the relation between terms, various document representation models have been proposed to capture the relation between terms based on statistical analysis. Generalized Vector Space model (GVSM) [2] and Context Vector Model (CVM-VSM) [3] both consider the term-term relation by examining the co-occurrence information. Other approaches, such as Latent Semantic Indexing (LSI) [4], have been applied to estimate the similarity between documents by using the projected feature space that captures the semantic information in the original document. These approaches have consistently performed better than the BOW model for document clustering. For example, the cosine similarity between d_1 and d_2 based on the GVSM model is 0.522, d_1 and d_3 is 0.473. It is now easier to distinguish the similarity between d_1 and d_2 from that between d_1 and d_3, compared to the BOW model. However, those approaches estimate the similarity between terms (e.g. "*Data mining*" and "*Computer science*") by their co-occurrence in a simple way, while the implicit relation between terms (e.g. "*Data mining*" and "*Machine learning*") is overlooked. Therefore, the traditional measures based on the co-occurrence information do not take the underlying implicit relation into consideration, which means that the traditional measures fail to capture the complete semantic relation between terms.

In this work, we propose a novel approach to measure the relationship between terms by capturing both the *intra-relation (explicit)* and *inter-relation (implicit)* using the co-occurrence information between them. The *intra-relation*

X.Cheng and D.Miao are with the Department of Computer Science and Technology, Tongji University, Shanghai, China. C.Wang and L.Cao are with the Advanced Analytics Institute, University of Technology, Sydney, Australia. Email: {cx1227, canwang613, longbing.cao}@gmail.com, miaoduoqian@163.com

This work was supported by The National Natural Science Foundation of China (No. 61075056, 61273304) and The Fundamental Research Funds for the Central Universities. This work is also supported in part by the Australian Research Council Discovery grant (DP130102691) and Linkage grants (LP120100566 and LP100200774).

adapts the original co-occurrence based approaches, and the *inter-relation*, which has been overlooked, is also teased out to derive the complete description of the semantic relation between terms. The key contributions are as follows:

- We propose the intra-relation to describe the explicit co-occurrence of terms by adapting the Jaccard [5] measure.
- We reveal the inter-relation to capture the implicit dependency of terms via their link terms.
- We aggregate the intra-relation and inter-relation between terms together to compose the coupled relation. This provides a complete representation of the semantic information for the document set.
- We evaluate our proposed coupled relation based document representation by comparing with the existing techniques. The experimental results show that our proposed method outperforms the state-of-the-art approaches.

The rest of the paper is organized as follows. Section II describes the background knowledge and reviews the related work. Section III presents a new approach to capture the coupled relation between terms. The experiments and results are discussed in Section IV, and the conclusion and future work are described in Section V.

II. Background and Related Work

A. Background

The Bag of Words (BOW) model represents each document as a vector of the distinct terms that appear in the document set. Each component of the vector stands for the weight of each term in the document set, and the weight is calculated by using the tf-idf weighting scheme. Let *D* be a document set, $D = \{d_1, d_2, ..., d_m\}$. Each document is defined as follows:

$$\vec{d} : d \mapsto \vec{d} = (tfidf(t_1, d), tfidf(t_2, d), ..., tfidf(t_n, d)), \tag{1}$$

where n is the number of terms in *D*.

The document set *D* is then represented as a $m \times n$ matrix *W*. Each row in *W* corresponds to a document in *D*, and every column describes the distribution of each term across the entire document set.

The BOW model has been widely applied in the text mining owing to its simplicity and general applicability. But it assumes that the terms are independent and ignores the semantic relation between them. Thus, it cannot integrate the semantic content into document representation. To keep the original semantic information in the document representation, various extensions to the BOW model are proposed to project the document vector onto a new feature space [6], which enriches the semantic information into the document representation. Generally, the projection is defined as $\vec{d}' = \vec{d}S$, where *S* is the semantic matrix including all the relations between terms.

The semantic matrix S is to refine the document representation with the embedded semantic information. Various kinds of S lead to different extensions to the BOW model. They will be introduced in the following section.

B. Related Work

The problem of capturing the relation between terms has recently attracted considerable attention [2][3][5][6][7][8][9] [10]. Most of these works focus on estimating the relation using co-occurrence information. Examples of such works include the GVSM [2], CVM-VSM [3], and GTCV-VSM [6].

The generalized vector space model (GVSM) was proposed by Wong et al. [2], it captured the relation between terms by their co-occurrence information across the entire document set. It simply utilizes the document-term matrix W^T as *S*, and then each document vector is projected as $\vec{d}' = \vec{d}W^T$. The corresponding kernel between two document vector is expressed as

$$k'(d_i, d_j) = \vec{d_i} W^T W \vec{d_j}^T \tag{2}$$

The entry in matrix $W^T W$ reflects the similarity between terms which is measured by their frequency of co-occurrence across the document set, which means two terms are similar if they frequently co-occur in the same document.

The context vector model [3] is a different approach which is also based on the co-occurrence frequency of terms in the same document. It directly adds the co-occurrence influence into the BOW model, so it keeps the original dimension of the documents rather than mapping it to a low-dimension feature space. In some other models for document representation, like in [6], the local information is taken into consideration to build the global context information. In [11], the covariance matrix is used in the latent space to compose a hybrid vector mapping.

It is worth noting that most existing relation measures, which are based on statistical analysis, just use the co-occurrence frequency of terms in the same document to enhance the quality of clustering, but the underlying relation which could be estimated by their link with other terms cannot be handled. In this work, we combine the underlying implicitly semantic relation into the traditional measure based on the co-occurrence frequency. To the best of our knowledge, the incorporation of such relation for text clustering has not be researched until now.

III. Coupled Term-Term Relation Analysis

In this section, we present a coupled term-term relation model (CRM for short) for document representation. The coupled term-term relation is proposed to exploit both the intra-relation and inter-relation between terms. As shown in Fig. 1, the intra-relation captures the similarity of terms by their co-occurrence, while the inter-relation further explores the similarity of terms by examining other terms that co-occur with each of them. The difference between these two relations is that the former one only involves the terms to be considered, but the latter one also includes other relevant terms. Then we analyze the similarity between documents

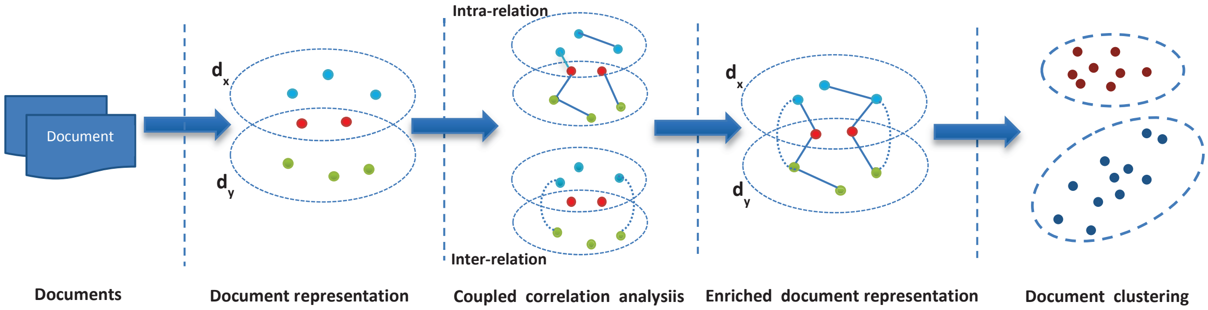


Fig. 1. An overview of term-term relation analysis based on CRM.

based on the relation between their terms. Finally, the documents are grouped into clusters according to their similarity. The whole procedure is presented in Fig. 1.

A. Intra-Relation Between Terms

As mentioned in Section I, terms are assumed to be independent in the traditional BOW model. However, in the real world, terms are always related to each other. To capture the relation between terms, some approaches are proposed to measure the similarity between terms by using statistical analysis [2][3][6]. They suppose that terms are relational if they co-occur in the same document. For instance, in Fig. $2(a)$, terms t_i and t_k co-occur in document d_x, while t_j is the co-occurrence term of t_k in document d_y. Then, term t_i is considered to be associated with t_k in document d_x, and term t_j is related with t_k in document d_y. The relation between terms is visually exhibited in Fig. $2(a)$. Accordingly, the relation between terms in the document set is estimated by calculating their co-occurrence frequency across all documents.

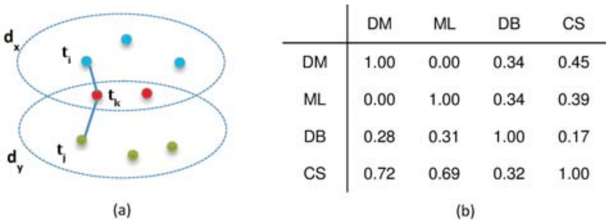
In most of the previous approaches, the relation between terms is simply estimated by the inner product of their distribution across the entire document set. Here, we adapt the popular co-occurrence measure *Jaccard* [5] to evaluate the relation rather than simply considering the inner product of them.

Definition 1: Terms t_i and t_j are said to be intra-related if they co-occur in at least one document d_x ($d_x \in D$). The co-occurrence relation between terms t_i and t_j across D is quantified as:

$$CoR(t_i, t_j) = \frac{1}{|H|} \cdot \sum_{x \in H} \frac{w_{xi} w_{xj}}{w_{xi} + w_{xj} - w_{xi} w_{xj}}, \tag{3}$$

where w_{xi} and w_{xj} represent the tf-idf weights of t_i and t_j in d_x, respectively; and $|H|$ denotes the number of elements in $H = \{x | (w_{xi} \neq 0) \vee (w_{xj} \neq 0)\}$. If $H = \emptyset$, we define $CoR(t_i, t_j) = 0$.

We further define the intra-relation as a conditional probability manner by normalizing the relation between t_i and t_j $CoR(t_i, t_j)$ to [0,1] with respect to the total amount of relation between term t_i and the other terms. The intra-relation reflects that when term t_i occurs in a document, the probability of term t_j that co-occurs with it together. Then the intra-relation between t_i and t_j is defined as follows.



(a)

	DM	ML	DB	CS
DM	1.00	0.00	0.34	0.45
ML	0.00	1.00	0.34	0.39
DB	0.28	0.31	1.00	0.17
CS	0.72	0.69	0.32	1.00

(b)

Fig. 2. An example of discovered Intra-relation between terms and the Intra-relation matrix.

Definition 2: The **intra-relation** between terms t_i and t_j is defined as:

$$IaR(t_i, t_j) = \begin{cases} 1 & i = j, \\ \frac{CoR(t_i, t_j)}{\sum\limits_{i=1, i \neq j}^{n} CoR(t_i, t_j)} & i \neq j, \end{cases} \tag{4}$$

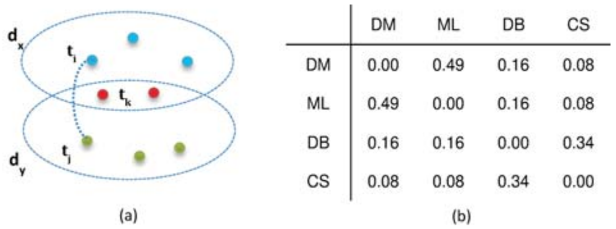
where $CoR(t_i, t_j)$ is the co-occurrence relation between terms t_i and t_j.

The intra-relation between t_i and t_j quantifies the possibility that t_i appears when t_j has already occurred. For all the terms t_i $(i \neq j)$, we have $IaR(t_i, t_j) \geq 0$ and $\sum_{i=1, i \neq j}^{n} IaR(t_i, t_j) = 1$. Note that $IaR(t_i, t_j) = IaR(t_j, t_i)$ usually does not hold, due to the fact that IaR essentially corresponds to the conditional probability. It also indicates that the intra-relation is not symmetrical. Then the intra-relation between terms in Table I can be captured using Equation (3). For example, the intra-relation $IaR(\text{“}\mathit{Machine\ learning}\text{”}, \text{“}\mathit{Computer\ science}\text{”})$ is 0.39, $IaR(\text{“}\mathit{Data\ mining}\text{”}, \text{“}\mathit{Computer\ science}\text{”})$ is 0.45, and IaR(“*Data mining*”, “*Database*”) is 0.34. The intra-relation matrix is shown in Fig. 2(b).

Exploiting the co-occurrence of terms helps to discover the explicit relation between terms. However, it lacks the ability to reveal the underlying relation other than co-occurrence frequency. In the following section, we will define the underlying relation and specify the inter-relation between them.

B. Inter-Relation Between Terms

The above intra-relation between terms only captures their co-occurrence frequency. Some terms co-relate with each



	DM	ML	DB	CS
DM	0.00	0.49	0.16	0.08
ML	0.49	0.00	0.16	0.08
DB	0.16	0.16	0.00	0.34
CS	0.08	0.08	0.34	0.00

Fig. 3. An example of discovered Inter-relation between terms and the Inter-relation matrix.

other closely though they do not happen to co-occur in the same document. For instance, "*Machine learning*" and "*Data mining*" are closely associated in the real world, but the relation between them has been not detected in Fig. 2(b). Hence, in this section, we introduce a novel approach to capture this kind of underlying relation (e.g. between "*Machine learning*" and "*Data mining*").

The inter-relation analysis is inspired by the fact that terms with the similar sense must appear in a similar context [3]. Therefore, we explore the relation between a pair of terms by their context, which is captured by their interaction with other terms across the entire document set. Inspired by the connected-triple proposed in [12] and the coupled similarity introduced in [13][14][15], we define the relative inter-relation between terms below.

Definition 3: Terms t_i and t_j are said to be **inter-related**, if there exists at least one term t_k such that both $IaR(t_i, t_k) > 0$ and $IaR(t_j, t_k) > 0$ hold. The term t_k is called the **link term** between them. The **relative inter-relation** between terms t_i and t_j linked by the term t_k is formalized as:

$$R\_IeR(t_i, t_j|t_k) = min(IaR(t_i, t_k), IaR(t_j, t_k)), \qquad (5)$$

where $IaR(t_i, t_k)$ and $IaR(t_j, t_k)$ are the intra-relations between t_i and t_k, t_k and t_j, respectively.

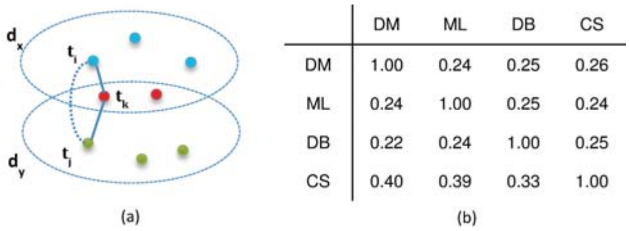
In other words, t_k is the link term if it is intra-related with both t_i and t_j. The relative inter-relation captures the smaller intra-relation to measure the difference between two terms with respect to a third term. For instance, in Fig. 2(a), terms t_i and t_j intra-related with t_k in respective documents d_x and d_y. Thus, t_i and t_j are said to be inter-related because they are linked by the term t_k. Fig. 3(a) shows the implicit relation (in a dotted line) discovered by the inter-relation analysis for the example in Fig. 2(a).

Definition 4: The **inter-relation** between two terms t_i and t_j is defined by their interaction with all the link terms, formalized as:

$$IeR(t_i, t_j) = \begin{cases} 0 & i = j \\ \frac{1}{|L|} \sum_{\forall t_k \in L} R\_IeR(t_i, t_j|t_k) & i \neq j \end{cases} \qquad (6)$$

where $|L|$ denotes the number of link terms in $L = \{t_k | (IaR(t_k, t_i) > 0) \wedge (IaR(t_k, t_j) > 0)\}$, and $R\_IeR(t_i, t_j|t_k)$ is the relative inter-relation between t_i and t_j linked by t_k. If $L = \emptyset$, we define $IeR(t_i, t_j) = 0$.

The value of $IeR(t_i, t_j)$ falls in [0,1]. When there is not a link term for t_i and t_j, we regard $IeR(t_i, t_j) = 0$. The definition indicates that the larger the average relative inter-relations with link terms, the closer a pair of terms are inter-related. For instance, the relation between "*Machine Learning*" and "*Data mining*" is captured by using Equation (5) and the value is 0.49, which is presented in Fig. 3(b). In this way, the relation between terms that have link terms is enhanced. Accordingly, the implicit relation revealed by the link terms makes the related documents more alike, which facilitates the clustering of documents.



	DM	ML	DB	CS
DM	1.00	0.24	0.25	0.26
ML	0.24	1.00	0.25	0.24
DB	0.22	0.24	1.00	0.25
CS	0.40	0.39	0.33	1.00

Fig. 4. An example of discovered coupled-relation between terms and the coupled-relation matrix.

C. Coupled Term-Term Relation Measure

To properly capture the underlying relationship between terms for accurate document clustering, the Intra-relation and Inter-relation are aggregated to capture the full relations between terms. Based on Equations (3) and (6), the coupled relation between terms t_i and t_j is defined as follows:

Definition 5: Given a pair of terms t_i and t_j in D, the **coupled relation (CR)** between t_i and t_j is defined as

$$CR(t_i, t_j) = \begin{cases} 1 & i = j \\ \alpha \cdot IaR(t_i, t_j) + (1 - \alpha) \cdot IeR(t_i, t_j) & i \neq j \end{cases} \qquad (7)$$

where $\alpha \in [0, 1]$ is the parameter that decides the weight of intra-relation, $IaR(t_i, t_j)$ and $IeR(t_i, t_j)$ are the respective intra-relation and inter-relation between terms t_i and t_j.

The similarity matrix which reflects the relation between terms is then presented as $S_{CR}(i, j) = CR(t_i, t_j)$. The value of $S_{CR}(i, j)$ falls within [0,1], in which 0 indicates that two terms are completely unrelated while 1 indicates that they are the same. The higher the coupled relation value, the more similar the two terms. By following the example in Fig 2, the coupled relation matrix S is shown in Fig. $4(b)$.

The coupled relation matrix is less sparse than the original intra-relation matrix because it further reveals the implicit relation in the original document. Therefore, the coupled relation matrix S_{CR} takes more semantic information in the document into consideration for clustering.

D. Document Similarity Measure

The coupled relation matrix S_{CR} contains not only the explicit but also implicit relations between each pair of terms across the entire document set. The mapping of each initial document vector $\vec{d'} = \vec{d}S_{CR}^T$ reflects the mutual influence of terms and reserves more semantic information from the original document. Then, the corresponding kernel [16] based on S_{CR} is written as:

TABLE II

AN EXAMPLE OF THE ENRICHED DOCUMENT REPRESENTATION WITH CR ($\alpha = 0.5$)

	DM	ML	DB	CS
d_1	0.64	0.26	0.32	0.46
d_2	0.24	0.62	0.31	0.39
d_3	0.22	0.23	0.83	0.30

$$k^{'}(d_i, d_j) = \vec{d_i}(S_{CR}^T * S_{CR})\vec{d_j}^T \tag{8}$$

The cosine similarity between documents is calculated as follows:

$$\cos(\vec{d_i}, \vec{d_j}) = \frac{k^{'}(d_i, d_j)}{\sqrt{k^{'}(d_i, d_i)}\sqrt{k^{'}(d_j, d_j)}} \tag{9}$$

Table II shows the projected vectors of the original document vectors shown in Table I. With the coupled relation based representation, the cosine similarity between d_1 and d_2 is 0.805, d_1 and d_3 is 0.724. We can observe that the similarity between d_1 and d_2 is much higher than that of d_1 and d_3, which is closer to the understanding in the real world. It is obvious that the new vectors more fully capture the semantic relations between terms, which enhances document clustering than other methods.

IV. EXPERIMENT AND EVALUATION

In this section, we empirically evaluate our coupled relation based representation in document clustering. The spherical k-means algorithm [17] is applied for clustering and the BOW model is used as the baseline for comparison. For the spherical k-means algorithm, a specific number of clusters k is required for clustering. In our experiments, we set k equal to the number of classes in the document set for comparing relative performance. In Equation 7, Intra- and Inter-relation are combined to measure the relation between two terms. As discussed in Definition 5, the parameter α controls the effect of intra- and inter-relation on document clustering, which is crucial in our experiment. We empirically assign α with 0.5. It means the inter-relation play the same important role as the intra-relation in document clustering. The effect of the inter-relation will be discussed in Section D.

A. Data Sets

We conduct experiments on four data sets. D1 is the subset of 20 Newsgroups [18] while D2 and D3 are the subsets of Reuters 21578 [19], and D4 is the WebKB benchmark document collection [20]. The characteristics of these data sets are summarized in Table III. *m*, *n* is the number of documents and terms respectively, and n_{avg} is the average number of terms per document.

Before conducting the document representation, all data sets are pre-processed to apply word stemming. We also discard the documents that are less than 10 words which means that there is less information for document clustering.

B. Evaluation Criteria

The quality of document clustering is evaluated by three criteria: Rand Index (RI), F_1-measure and Normalized mutual information (NMI).

The first measure is Purity, which is a simple and transparent way to measure the quality of clustering. The purity of cluster c_i is computed by the ratio between the size of the dominant class in the cluster ($max_j(|c_{ij}|)$) and the size of cluster ($|c_i|$): $purity(c_i) = \frac{1}{|c_i|}\max_j |c_{ij}|$. Then the overall purity can be expressed as the weighted sum of all individual cluster purity:

$$putity = \sum_{i=1}^{k} \frac{|c_i|}{N} purity(c_i), \tag{10}$$

where k is the number of clusters and N is the number of documents.

The second is Rand index. Rand index is used to measure the clustering quality by the percentage of the true positive and true negative decisions in all decisions during clustering:

$$RI = \frac{TP + TN}{TP + TN + FP + FN}, \tag{11}$$

where TP (true positive) denotes that two similar documents are assigned to the same cluster; TN (true negative) denotes that two dissimilar documents are assigned to different clusters; FP (false positive) denotes that two dissimilar documents are assigned to the same cluster, and FN (false negative) denotes that two similar documents are assigned to different clusters.

The third measure is F1-measure. It is a criterion considered both the precision and recall for clustering evaluation according to the following formula:

$$F1 = \frac{precision \times recall}{precision + recall}, \tag{12}$$

where $precision = TP/(TP + FP)$, $recall = TP/(TP + FN)$.

The last measure is Normalized mutual information (NMI), which is a popular information theoretic criterion for evaluating clustering quality. It's computed by dividing the Mutual Information between the clusters and the label of the dataset with the average of the clusters and the pre-exist classes entropy.

$$NMI(C, L) = \frac{I(C; L)}{(H(C) + H(L))/2}, \tag{13}$$

where C is a random variable for cluster assignments, L is a random variable for the pre-existing classes on the same data. $I(C; L)$ is the mutual information between the clusters and the label of the dataset:

$$I(C; L) = \sum_i \sum_j \frac{|c_i \cap l_j|}{N} \log \frac{N|c_i \cap l_j|}{|c_i||l_j|}, \tag{14}$$

and $H(C)$ and $H(L)$ is the entropy of C and L:

$$H(C) = -\sum_i \frac{|c_i|}{N} \log \frac{|c_i|}{N}, H(L) = -\sum_j \frac{|l_j|}{N} \log \frac{|l_j|}{N}, \tag{15}$$

TABLE III

CHARACTERISTICS OF DATA SETS

Data sets	Topics	Classes	m	n	n_{avg}
D1	20-NGs: atheism, graphics, windows.misc, pc.hardware, mac.hardware	5	1864	16516	76
D2	Reuters-21,578: acq, crude, earn, grain, interest, money-fx, ship, trade	8	2091	8674	33
D3	Reuters-21,578: acq, alum, bop, carcass, cocoa, coffee, copper, cotton, cpi, cpu, crude, dlr, earn, fuel, gas, gnp, gold, grain, heat, housing, income, instal-debt, interest, ipi, iron-steel, jet, jobs, lead, lei, livestock, lumber, meal-feed, money-fx, money-supply, nat-gas, nickel, orange, pet-chem, platinum, potato, reserves, retail, rubber, ship, strategic-metal, sugar,tea, tin, trade, veg-oil, wpi, zinc	52	2448	9728	36
D4	WebKB: course, faculty, project, student	4	4087	7769	32

where $|c_i|$, $|l_j|$ and $|c_i| \cap |l_j|$ is the number of documents in cluster c_i, pre-existing class l_j and in the number of the common documents in c_i and l_j, and N is the number of documents in the document set.

For these quality measures, a higher value in [0, 1] indicates a better clustering quality.

C. Performance Evaluation

In this paper, the performance of our approach (CRM) is compared with two other models: the classic BOW model (BOW) and the GVSM model (GVSM), and the classical BOW model is used as the baseline for comparison. Table IV illustrates the Purity, RI, F_1-measure and NMI scores computed from the clustering result by using the spherical k-means algorithm on four data sets.

From Table IV, we observe that the GVSM model augments the performance of BOW model on all data sets. The GVSM model achieves around 1%, 12%, 7% and 12% on the average Purity, RI, F_1-measure and NMI respectively. The GVSM achieves better performance over BOW model, demonstrates the benefit and incorporating the co-occurrence relation into document representation. It demonstrates the benefit of integrating the relation between terms into document representation.

Comparing our proposed CRM with the other two model, it further improves the performance over GVSM and achieves the best scores on all data sets. Compared with the BOW model, our approach achieves 8% improvement on the average Purity score of four data sets, and the improvement in RI is around 22%, 12% for the F_1-measure and 14% for the NMI. Compared with the GVSM model, our approach also achieves 5%, 8%, and 5.5% on the average Purity, RI and F_1-measure scores, respectively. These comparisons indicate that our approach efficiently improves the performance of document clustering. We believe that this is because that the coupled relation integrates more semantic information from the original documents into document representation.

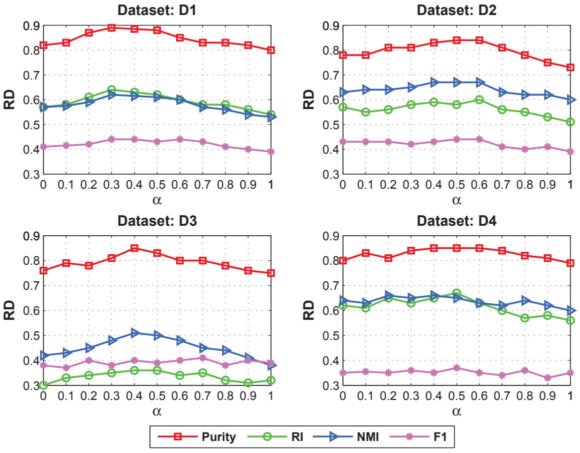


Fig. 5. The effect of varying the parameter α on the clustering performance for each data set.

D. Effect of Inter-relation

In order to illustrate the impact of inter-relation on the performance of clustering, we conduct experiments using our proposed model (CRM) with different values of parameter α on the four experimental data sets. In these experiments, we evaluate the clustering performance by setting α from 0 to 1 at increments of 0.1. The experimental results are presented in Fig. 5.

In Fig. 5, the curves of the purity, RI, F1-measure, and NMI scores depict the performance of our proposed approach, varying along with the value of parameter α for each data set. For the first data set, the purity, RI, F1-measure, and NMI scores reach a peak at $\alpha = 0.3$, which demonstrates that our approach achieves the best performance on the first data set when $\alpha = 0.3$. The clustering performance raises as the parameter α increases from 0 to 0.3. We believe that is because of the integration of the inter-relation. when the value of the parameter α grows since $\alpha = 0.3$, the

TABLE IV

DOCUMENT CLUSTERING RESULTS BY USING SPHERICAL K-MEANS

Data	Purity			RI			F1-measure			NMI		
Sets	*BOW*	*GVSM*	*CRM*	*BOW*	*GVSM*	*CRM*	*BOW*	*GVSM*	*CRM*	*BOW*	*GVSM*	*CRM*
D1	0.79	0.82	**0.88**	0.49	0.57	**0.62**	0.48	0.58	**0.61**	0.32	0.41	**0.44**
D2	0.80	0.80	**0.84**	0.49	0.58	**0.60**	0.62	0.66	**0.67**	0.44	0.48	**0.44**
D3	0.78	0.79	**0.83**	0.27	0.29	**0.36**	0.41	0.43	**0.50**	0.37	0.42	**0.39**
D4	0.82	0.81	**0.85**	0.63	0.65	**0.67**	0.66	0.65	**0.65**	0.32	0.35	**0.37**

performance of clustering declines, which means the inter-relation brings negative impact into the relation matrix. With the rest three data sets, the trend of the curve is similar to the first data set, and our approach achieves the best performance at $\alpha = 0.6, 0.4$ and 0.5 respectively.

In general, the result demonstrates that the inter-relation has great impact on the performance of document clustering, and it plays important role in document clustering as the intra-relation. Besides, we observe that the best performance with the different value of α on different data sets. It demonstrates that different data distributions influence the setting of parameter α which decides the importance of inter-relation for document clustering. Therefore, it's essential to optimize the setting of α when the application requires higher clustering accuracy.

E. Detail Analysis

To better understand the reason why our approach performs better than BOW and GVSM, we illustrate the discovered relations (i.e. the number of non-zero elements in the relation (or semantic) matrix) from the experimental data sets, as shown in Fig. 6. In this figure, we can observe that the GVSM incorporates the underlying relations between terms into document representation using their co-occurrence information, and it efficiently improve the performance of BOW. It demonstrates the benefit and potential of integrating term relation into the document representation.

In our approach, we consider both the intra-relation and inter-relation between terms to discover more underlying relations between terms, as shown in Fig. 6. Comparing with BOW and GVSM, our coupled approach achieves the best performance on all data sets. We conjecture the improvement stems from the coupled-relation which integrates the inter-relation into document representation with intra-relation to generate high quality document representation.

F. Scalability Analysis

In this section, we study the scalability of our approach on document clustering. The scalability of CRM is investigated on the affect of clustering performance with different sizes of data sets.We conduct a set of experiments on the data set WebKB by increasing the number of documents from $1,000$ to $4,000$ at increment of $1,000$.

The experimental result is shown in Fig. 7, which illustrates the purity, RI, F_1-measure, and NMI scores vary along with the size of document sets. We can observe that the four measure scores is not affected by the size of document set. The curves of these four measure scores indicate that the quality of clustering is stable with the increasing size of document set.

In summary, the experimental results verify that our proposed CRM approach outperforms both the traditional BOW approach and the GVSM approach in terms of the quality measures, and the inter-relation has the significant impact on the performance of clustering as the intra-relation. In addition, the quality of clustering is stable as the document size increases.

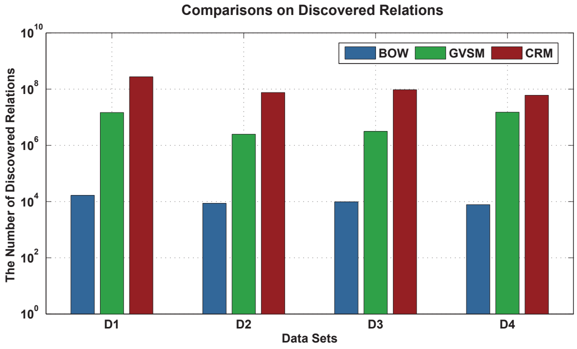


Fig. 6. Comparisons of the number of discovered relations on four data sets.

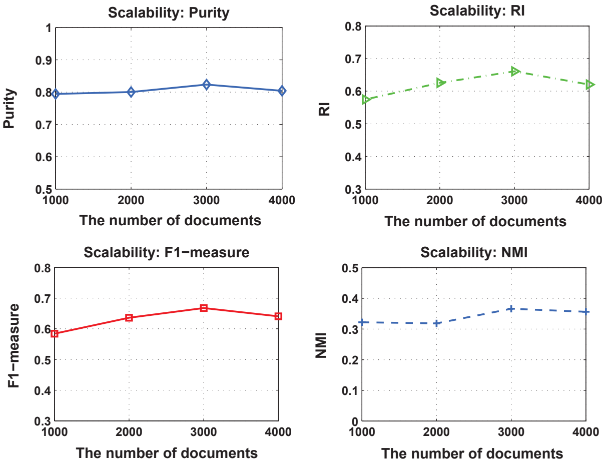


Fig. 7. The scalability of CRM on Purity, RI, F_1-measure, NMI scores.

V. Conclusion and Future Work

In this paper, we present a novel approach to capture the coupled relation between terms to improve the performance of document clustering. Based on the combination of intra-relation and inter-relation, our approach integrates more semantic information into document representation. Our approach operates in a sequence of four steps: (1) Capture the intra-relation between two terms from the original documents using statistical analysis. (2) Discover the underlying inter-relation between terms based on the related intra-relation. (3) Combine the intra- and inter-relation as the coupled-relation by an optimal parameter α in Equation 8 to capture the full relation between terms. (4) Project the original document into a new feature space using the coupled-relation matrix. In the experiment study, we compared our proposed representation model with the classical BOW and GVSM. Empirical evaluations demonstrates that our proposed approach significantly outperform the previous approaches.

In the future, we will conduct further research to improve our work. Firstly, we will study on optimizing the parameter α by analyzing the data distribution, as the clustering performance can be greatly improved by an optimal α. Secondly, we will study the independence test to determine whether two terms occur together more often than by chance. Finally, we aim to reduce the complexity of the proposed approach for document clustering.

References

[1] A. Huang, D. Milne, E. Frank, and I. Witten, "Clustering documents using a wikipedia-based concept representation," *Advances in Knowledge Discovery and Data Mining*, pp. 628–636, 2009.

[2] S. Wong, W. Ziarko, and P. Wong, "Generalized vector spaces model in information retrieval," in *SIGIR 1985*. ACM, 1985, pp. 18–25.

[3] H. Billhardt, D. Borrajo, and V. Maojo, "A context vector model for information retrieval," *Journal of the American Society for Information Science and Technology*, vol. 53, no. 3, pp. 236–249, 2002.

[4] S. Deerwester, S. Dumais, G. Furnas, T. Landauer, and R. Harshman, "Indexing by latent semantic analysis," *Journal of the American society for information science*, vol. 41, no. 6, pp. 391–407, 1990.

[5] D. Bollegala, Y. Matsuo, and M. Ishizuka, "Measuring semantic similarity between words using web search engines," in *WWW 2007*, vol. 7, 2007, pp. 757–786.

[6] A. Kalogeratos and A. Likas, "Text document clustering using global term context vectors," *Knowledge and information systems*, vol. 31, no. 3, pp. 455–474, 2012.

[7] J. A. Bullinaria and J. P. Levy, "Extracting semantic representations from word co-occurrence statistics: A computational study," *Behavior Research Methods*, vol. 39, no. 3, pp. 510–526, 2007.

[8] G. Tsatsaronis, I. Varlamis, and M. Vazirgiannis, "Text relatedness based on a word thesaurus," *Journal of Artificial Intelligence Research*, vol. 37, no. 1, pp. 1–40, 2010.

[9] J. Hu, L. Fang, Y. Cao, H. Zeng, H. Li, Q. Yang, and Z. Chen, "Enhancing text clustering by leveraging wikipedia semantics," in *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 2008, pp. 179–186.

[10] F. Figueiredo, L. Rocha, T. Couto, T. Salles, M. André Gonçalves, and W. Meira Jr, "Word co-occurrence features for text classification," *Information Systems*, 2011.

[11] A. Farahat and M. Kamel, "Statistical semantics for enhancing document clustering," *Knowledge and Information Systems*, vol. 28, no. 2, pp. 365–393, 2011.

[12] N. Iam-On, T. Boongoen, S. Garrett, and C. Price, "A link-based approach to the cluster ensemble problem," *IEEE TPAMI*, vol. 33, no. 12, pp. 2396–2409, 2011.

[13] C. Wang, L. Cao, M. Wang, J. Li, W. Wei, and Y. Ou, "Coupled nominal similarity in unsupervised learning," in *CIKM 2011*, 2011, pp. 973–978.

[14] L. Cao, Y. Ou, and P. S. Yu, "Coupled behavior analysis with applications," *IEEE Transactions on Knowledge and Data Engineering*, vol. 24, no. 8, pp. 1378–1392, 2012.

[15] Y. Song, L. Cao, X. Wu, G. Wei, W. Ye, and W. Ding, "Coupled behavior analysis for capturing coupling relationships in group-based market manipulations," in *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2012, pp. 976–984.

[16] L. AlSumait and C. Domeniconi, "Text clustering with local semantic kernels," *Survey of Text Mining II*, pp. 87–105, 2008.

[17] I. Dhillon, J. Fan, and Y. Guan, "Efficient clustering of very large document collections," *Data Mining for Scientific and Engineering Applications*, pp. 357–381, 2001.

[18] K. Lang, "Newsweeder: Learning to filter netnews," in *In Proceedings of the Twelfth International Conference on Machine Learning*. Citeseer, 1995.

[19] D. D. Lewis, "Reuters-21578 text categorization test collection, distribution 1.0," *http://www. research. att. com/˜ lewis/reuters21578. html*, 1997.

[20] M. Craven, D. DiPasquo, D. Freitag, A. McCallum, T. Mitchell, K. Nigam, and S. Slattery, "Learning to extract symbolic knowledge from the world wide web," in *Proceedings of the 15th National Conference on Artificial Intelligence*, 1998.