Neurocomputing 154 (2015) 127–138

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

A link-based approach to semantic relation analysis

Xin Cheng [a,*], Duoqian Miao [a], Can Wang [b]

[a] Department of Computer Science, Tongji University, Shanghai, China
[b] Commonwealth Scientific and Industrial Research Organisation, Australia

ARTICLE INFO

Article history:
Received 27 May 2014
Received in revised form
29 September 2014
Accepted 1 December 2014
Communicated by Y. Chang
Available online 15 December 2014

Keywords:
Semantic relation analysis
Co-occurrence statistics
Neighbor information
Link-based relation

ABSTRACT

The semantic relation analysis is an interesting issue in natural language processing. To capture the semantic relation between terms (words or phrases), various approaches have been proposed by using the co-occurrence statistics within corpus. However, it is still a challenging task to build a robust relation measure due to the complexity of the natural language. In this paper, we present a novel approach for the semantic relation analysis, which takes account of both the pairwise relation and the link-based relation within terms. The pairwise relation captures the relation between terms from the local view, which conveys the co-occurrence pattern between terms to measure their relation. The link-based relation involves the global information into the relation measure, which derives the relation between terms from the similarity of their context information. The combination of these two relations creates a model for robust and accurate semantic relation analysis. Experimental evaluation indicates that our proposed approach leads to much improved result in document clustering over the existed methods.

© 2014 Elsevier B.V. All rights reserved.

1. Introduction

The study of semantic relation between terms is an important issue in natural language processing. It is a big challenge to capture the complete and precise semantic relation between terms due to the complexity of natural language. A number of semantic relation measures have been proposed in the previous literatures, and different approaches have been proven useful in the specific areas of document processing, such as document retrieval [1,2], document summarization [3,4] and word sense disambiguation [5–7].

It is worth noting that the precise analysis of the semantic relation between terms is an integral part of natural language processing. As with the accurate semantic relation between terms, we can capture the real semantic meaning of each document, and the original documents can be mapped into a novel feature space by integrating the semantic meaning into document representation. In the novel feature space, it is much easier to distinguish whether the documents are semantically similar or not. According to the enhanced document representation, we can further improve the performance of natural language processing, such as document retrieval, clustering and classification.

To capture the semantic relation between terms, a number of approaches have been proposed by using the co-occurrence information within corpus, and the resultant models have become known as word co-occurrence models. The basic idea of these models is simple that words with similar meanings will tend to occur in similar contexts, and hence word co-occurrence statistics can provide a natural basis for semantic representation. Generally, these semantic relation analysis models can be categorized into two groups: pairwise relation analysis model and similarity-based relation analysis model. The pairwise relation analysis means if two terms co-occur frequently, they are considered to be relational. Explicit simulations show that the pairwise relation analysis can be used to perform remarkably well on various performance criteria, but it only considers the relation between terms by themselves but overlooks their interaction with other terms (the context information) and fails to discover the underlying relation between terms. The similarity-based relation analysis is that if two terms have the similar distribution of co-occurrence with other terms, they are considered to be similar. Therefore, the similarity-based relation analysis takes the interaction with other terms into consideration for the relation measure, but it is insensitive to the strong relation which can be captured from the co-occurrence information in the pairwise relation analysis.

The motivation behind the work in this paper is that we believe that the semantic relation analysis should be based on not only the pairwise relation analysis, but also the similarity-based relation analysis. In this paper, we propose an approach for the semantic relation analysis from two points of view. The first is the pairwise relation analysis based on the co-occurrence information in the document collection. The second is the link-based relation analysis by considering their interaction with other terms. Then the pair and link-based relation analysis are combined to achieve a more

[*] Corresponding author. Tel.: +86 186 0027 1086; fax: +86 21 6958 9979.
E-mail address: cx1227@gmail.com (X. Cheng).

http://dx.doi.org/10.1016/j.neucom.2014.12.011
0925-2312/© 2014 Elsevier B.V. All rights reserved.

accurate semantic relation analysis model. Besides, we propose an effective measure for the *pairwise relation* between terms, and a *link-based relation* measure to calculate relation based on the similarity of their context information. With the pairwise relation, which considers the local information, and the link-based relation, which considers the global information, the combination of them is further proposed to improve the accuracy of semantic relation analysis.

In our experiments, which are used for evaluating the performance of various semantic relation analysis strategies, our strategy achieves significant improvement on the experimental document collections with respect to clustering task. For Purity, which is used to measure the overall precision of the clusters, the average score is 0.791. For Rand Index, which is used to measure the quality by the percentage of the true positive and true negative decisions in all decisions during clustering, the average score is 0.794. For *F*-measure, which considers both the precision and recall for evaluation, the average score is 0.470. For Normalized Mutual Information, which is 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, the average score is 0.512. The document representation with the semantic relation captured by our strategy performs significantly better than the traditional bag-of-word approach on document clustering (the average scores of above four measures with the bag-of-word approach are 0.710, 0.748, 0.399 and 0.441, respectively).

The main contributions associated with this work are as follows:

- A novel pairwise relation analysis approach, assigns levels of significance to the semantic relation between terms according to their co-occurrence information.
- A novel link-based relation analysis approach that captures semantic relation between terms from the similarity of their context information, which is based on the theory that the context information can be seen as a semantic description of each term, and the similar context information indicates terms are semantically related.
- An integration of the pairwise and link-based relation to capture the precise and reliable semantic relation between terms from corpus.
- Detailed analysis of the impact of pairwise and link-based relation with respect to the clustering task on the real document collections.

The rest of this paper is organized as follows. Section 2 provides an overview of the related work. Section 3 presents our proposed relation measure, which starts with the detailed description of our pairwise and link-based relation measures, and then introduces the optimal combination of these two measures to capture the semantic relation between terms. Section 4 illustrates the experimental results to show the improvement of our strategy with the comparison of the traditional bag-of-word approach, GVSM, HAL and LSA strategies. Then further analysis and discussion are provided based on the experimental results. Finally, the conclusion and future work is described in Section 5.

2. Related work

The solution to the problem that captures the semantic relation between terms using their co-occurrence information has attracted much research interest. The two most prevalent works are Latent Semantic Analysis (LSA) [8,9] and the Hyperspace Analogue to Language (HAL) [10]. Recently, some new approaches have also been proposed, including the context vector model for information retrieval (CVM-VSM) [2], compound features for document classification [11] and the term-term similarity model based on covariance matrix [12], that seek to improve on the previous works.

LSA learns the relation between terms by examining the patterns of term co-occurrence across the document collection. It firstly constructs the document-term matrix, and then compresses the matrix using singular value decomposition (SVD), which extracts the most important underlying dimensions from document-term matrix. Some variants of LSI have also been proposed. The Probabilistic Latent Semantic Analysis (PLSA) proposed by Hofmann [13] has a similar approach with LSA. In PLSA, the documents are mapped to a reduced vector space too, the latent semantic space. As opposed to LSA, the model has a solid statistical foundation. It is based on a latent variable model for general co-occurrence data, which associates an unobserved latent class variable with each observation. The number of latent factors will be much smaller than the number of words and the factors act as prediction variables for words. The factors are obtained using a generalization of the Expectation Maximization algorithm. Given the fact that the LSA focuses on the co-occurrence information between terms within a general context (e.g., a document), LSA is particularly well suited to capture the general association and relation which exist between terms. However, LSA overlooks the importance of the context information in deriving the term relation, and thus fails to capture the potential relation that exists between terms.

The HAL model is the other prevalent work, which has been successfully used in various areas, including word sense disambiguation [14], verb morphology modeling [15], abstract words and concrete words representing [16], emotional connotations [17] and word meaning learning [18]. Representation vectors in the HAL model are built from information about the proximal co-occurrence of terms within a large body of document. Using the Euclidean distance between co-occurrence vectors building with weighted 10 term windows, they are able to predict the degree of priming of one term by another in a lexical decision task. Given that these word vectors represent the positional context within which each word occurs, HAL is well suited to capturing the positional relation between words. However, in capturing positional information, the HAL model is largely insensitive to the types of the information to which LSA is sensitive.

The other approaches, which measure the semantic relation between terms by using the similarity of their context information, are introduced in recent literatures [19–21]. For instance, the semantic relation between terms can be measured by the ratio of the number of common words divided by the larger of the absolute number of words in their context information [19]. The performance of three different functions: the Jensen–Shannon divergence (total divergence to the average), the L1 norm, and the confusion probability, to measure the context similarity, are compared in [20]. Recently, deep learning methods have been successfully applied in language processing [21–23], in which words are represented as dense real-valued vectors. Such representation is referred as distributed word representation, which is designed to capture the semantic relation between words.

The generalized vector space model (GVSM), which was proposed by Wong et al. [24], 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 $\overrightarrow{d}' = \overrightarrow{d} W^T$. The corresponding kernel between two document vectors is expressed as

$$k'(d_i, d_j) = \overrightarrow{d_i} W^T W \overrightarrow{d_j}^T \quad (1)$$

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 CVM-VSM [2] also evaluated term co-occurrence pattern to estimate term dependency. It integrated the semantic information into document representation for the calculation of the document similarity. The empirical results confirm that it improves the performance of document retrieval for particular document collections. The work of [11] is another model which extracted new discriminative feature from documents, which are composed by terms that co-occur in the documents, and the new feature is conjunction with the word feature to enhance document classification. The authors conclude that the new feature extraction strategy consistently improves the performance of document classification on different document collections. Farahat and Kamel [12] defined different representation model which captured the semantic information from the original document by estimating term-term correlation using their co-occurrence pattern. The experimental results showed that the representation model which is based on covariance matrix between terms achieved the best performance in the document clustering task.

Other models for document representation are based on using lexical ontologies, such as WordNet [25], to represent documents in the space of concepts and calculate their similarity. Similar representation models are based on exploiting knowledge from an encyclopedia (like Wikipedia). Explicit semantic analysis (ESA) [26] is such a model that represents terms as vectors in a space of concepts represented by articles from Wikipedia. Wikipedia-based representation models have recently been used to enhance the performance of different text-mining tasks [27,28]. However, ontology and Wikipedia-based techniques have the limitation that the external knowledge is built manually and its coverage is restricted, and they are computationally complex as they depend on mining term-term correlations from extremely large knowledge sources. In addition, general-purpose thesauri, like WordNet, suffer from the presence of noise and irrelevant information. On the other hand, building a domain-specific thesaurus is a difficult task.

As introduced above, a number of recent literatures have contributed to the development of effective schemes to measure the semantic relation between terms based on their co-occurrence patterns. A major problem with these semantic relation analysis approaches is that most efforts have been targeted toward the pairwise relation analysis, but the underlying relation which is also essential to semantic relation analysis has been overlooked. Besides, they mostly assume that higher frequency indicates a richer relation between terms, but the discriminative ability of terms have been overlooked which would cause an imprecise estimation of relation, for example, the common terms always co-occur with the other terms frequently, but these high co-occurrence frequency is not much meaningful.

Based on these insightful observations, this paper introduces a novel link-based approach to semantic relation analysis based on the co-occurrence information between terms. We firstly propose a novel pairwise relation analysis based on the significance of term co-occurrence pattern, which considers both co-occurrence frequency but also the discriminative ability of each term. Then a link-based relation analysis is further proposed to capture the underlying relation between terms by considering the similarity of their context information. Lastly, we present semantic relation analysis approach which is based on the combination of pairwise and link-based relation analysis. In addition, we also analyze the effect of pairwise and link-based relation on the semantic relation analysis, and the result illustrates that they are both essential for semantic relation analysis.

3. Semantic relation analysis

In this section we describe our proposed approach to measure the semantic relation between terms. In order to capture the relation between terms precisely and reliably, our approach is divided into three consecutive steps. The first step is to capture the pairwise relation between terms using their co-occurrence information across the whole document collection, which will be introduced in Section 3.1. The second step is further to calculate the link-based relation between terms using their interaction with other terms, which will be introduced in Section 3.2. It is important to notice that the pairwise relation is concerned with the local information between terms and the link-based relation takes the global information into account to capture the relation between terms. Finally, the pairwise and link-based relations are combined together to capture the complete semantic relation between terms, which will be introduced in Section 3.3.

3.1. The pairwise relation between terms

One way to capture the semantic relation between terms is based on the common co-occurrence of such terms in a single document. The major challenge in this approach is to build an efficient scheme to analyze the semantic relation using the co-occurrence information between terms. Large number of previous literatures have contributed in the development of the improved measures, and they assume that higher co-occurrence frequency indicates a higher semantic relation between terms. However, the discriminative ability of the terms has been overlooked which is critical for the relation measure. For example, a pair of terms, t_i and t_j co-occur frequently in the document set, while t_j also appears frequently in the other documents, so t_j is a common term and t_i happens to co-occur with it frequently. Meanwhile, term t_k also co-occurs with t_i, but it appears frequently in this document but seldom appears in the other documents, which means the co-occurrence between t_i and t_k is more meaningful. Hence, it does not make sense t_i and t_j is with the same level of semantic relation as t_i and t_k.

The following example illustrates the problem: in a document collection, “bird” co-occurs 100 times with “has” in a document, and also 100 times with “swing”. In the previous approaches which measure the relation only by their co-occurrence frequency, “bird” and “has”, “bird” and “swing” are with the same degree of semantic relation. However, the term “has” is so common in the document collection which will lead to incorrectly emphasize the relation between “has” and other terms which happen to co-occur with it frequently. On the contrary, term “swing” appears less frequently in the document collection, which is with better discriminative ability, and the co-occurrence with this term should be more meaningful and can reflect the real semantic relation between them.

Therefore, contrary to the previous works by only considering the co-occurrence frequency between terms, our approach provides a more precise evaluation approach, which estimates the pairwise relation using the *tfidf* (term frequency-inverse term frequency) value of terms, which is often used as the weighting scheme in document mining. It considers not only the frequency of term in a single document but also considers its discriminative ability in the whole document collection. The term frequency (*tf*) is the raw frequency of a term in the document, and the inverse term frequency (*idf*) is a measure of the term frequency across all documents which reflects the discriminative ability of this term. Thus, using the *tfidf* scheme to measure the relation between terms, we can more precisely capture the significance of term-term co-occurrence.

After choosing the weighting scheme, the pairwise relation between terms can be calculated using their co-occurrence pattern. Suppose that there are two terms t_i and t_j appearing in the same document d_x, when $i=j$, the pairwise relation between them is actually the semantic relation between the same term, so it is equal to 1. When $i \neq j$, the pairwise relation between t_i and t_j in the document d_x is calculated based on the significance of their co-occurrence pattern. Mathematically, it is defined as follows:

Definition 3.1. Let D be a document collection, a pair of terms t_i and t_j co-occur in the same document $d_x \in D$, then the pairwise relation between them related to document d_x is defined as

$$PR(t_i, t_j; d_x) = \frac{w_{xi} w_{xj}}{w_{xi} + w_{xj} - w_{xi} w_{xj}}, \tag{2}$$

where w_{xi}, w_{xj} are the *tfidf* values of term t_i and t_j in the document d_x, respectively.

The intuition behind this definition is, if two terms co-occur in the same document with similar frequency, and also with the similar discriminative power, they are with a higher semantic relation. According to the semantic terms between terms in the document level, the semantic relation in the corpus level can be calculated by summarizing the pairwise relation across the whole document collection, and it is defined as follows:

Definition 3.2. Let D be a document collection, a pair of terms t_i and t_j co-occur in the same documents, then the pairwise relation between t_i and t_i related to document collection D is defined as

$$PR(t_i, t_j; D) = \sum_{d_x \in H} PR(t_i, t_j; d_x)/|H|, \tag{3}$$

where $PR(t_i, t_j; d_x)$ is the relation between t_i and t_j in document d_x, and H denotes the documents in $H = \{d_x | (w_{xi} \neq 0) \vee (w_{xj} \neq 0), d_x \in D\}$. If $H = \varnothing$, we define $PR(t_i, t_j; D) = 0$.

In the previous approaches, the semantic relation between terms is considered as being symmetric, which cannot express the semantic relation between terms accurately. For example, hyponymy is a relation between two terms in which the meaning of one of the terms includes the meaning of the other term. E.g. Blue, Green are kinds of color. They are specific colors and color is a general term for them. Therefore, color is called the super ordinate term, and blue, red, green, yellow, etc are called hyponyms. Intuitively, "blue" is highly related to "color", but "color" is that high related to "blue". Following that, we define the pairwise relation between t_i and t_j in a probabilistic manner to reflect the possibility that when t_i occurs in a document and t_j co-occur with it together.

Definition 3.3. Let D be a document collection, a pair of terms t_i and t_j that appear in the document collection, the *probabilistic pairwise relation* between terms t_i and t_j is defined as

$$PPR(t_i, t_j) = \begin{cases} 1 & i = j \\ \dfrac{PR(t_i, t_j; D)}{\sum_{k \neq i} PR(t_i, t_k; D)} & i \neq j \end{cases} \tag{4}$$

The value of $PPR(t_i, t_j)$ falls into the interval of [0, 1]. The larger value of $PPR(t_i, t_j)$ indicates the higher possibility that t_j co-occur with t_i in the same document. Note that the relation defined here is not symmetric, generally $PPR(t_i, t_j) \neq PPR(t_j, t_i)$, due to the denominator is with respect to the amount of the relation t_i and other terms in the document set.

The pairwise relation captured by the co-occurrence information can augment the quality of document representation and achieve a certain improvement in related tasks. However, it lacks the ability to reveal the underlying relation which is essential to capture more semantic information from the original documents. In the following section, we will introduce a link-based relation measure which is used to analyze the underlying semantic relation between terms.

3.2. The link-based relation between terms

It is noted that the pairwise relation analysis conveys the co-occurrence information between terms to measure their semantic relation, which is essential in capturing the precise semantic relation between terms. However, it solely considers the explicit relation between terms, but overlooks the implicit relation between them which can be captured from the global view. In this section, we will introduce a novel semantic relation analysis approach, the link-based relation analysis, which aims to reveal the implicit semantic relation between terms by considering the similarity of their context information. The link-based relation analysis is inspired by the fact that terms appearing in a similar context should be with the similar sense.

For example, there are two documents d_1 and d_2, d_1 describes the nature of "bird" and d_2 talks about "fish". In these two documents, term "bird" only appears in the document d_1 while "fish" only appears in the document d_2, so they do not co-occur in the same document d_1 or d_2, and they are not related to each other based on the pairwise relation analysis. But they have the similar co-occurrence terms (e.g. "animal", "skin", "eyes") as shown in Fig. 1. In fact, the terms that co-occur together can be seen as a semantic description of each other, which reflect the influence of terms in the conceptual description of each other. Therefore, if two terms have the similar co-occurrence terms, they should be semantically related. Here, the same co-occurrence terms, "animal", "skin" and "eyes", means the "bird" and "fish" are both animals, and they both have skin and eyes. Except for the same co-occurrence terms, the other co-occurrence terms (like "swing" and "fin"), which are semantically similar, also can be used to connect the relation between "bird" and "fish". Hence, terms "bird" and "fish" can be linked by these terms and they are related to each other, which is same as the judgement in the real world.

In this work, we capture the underlying relation between terms using the link-based relation measure, which is based on the similarity of the neighbor information of each term. At first, we introduce the definition of neighbor.

Definition 3.4. Let D be a document collection, a pair of terms t_i and t_j appear in the document collection, t_j is defined as the neighbor of t_i if

$$PPR(t_i, t_j) > \theta, \tag{5}$$

where $PPR(t_i, t_j)$ is the probabilistic pairwise relation between t_i and t_j, and $\theta \in [0,1]$ is a user-defined threshold which determines whether t_j should be the neighbors of t_i.

Thus, higher value of θ is corresponding to a higher threshold for a pair of terms to be considered as neighbor. A value of 1 for θ indicates that two terms are the neighbor of each other only if they are identical, on the other hand, a value of 0 for θ permits any pair of terms to be neighbors. It is noted that because the probabilistic

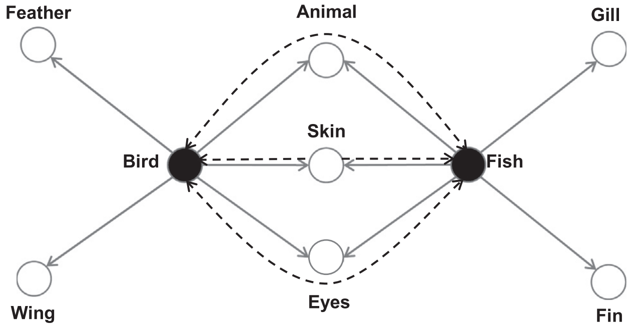


Fig. 1. A sample using the neighbor information to capture the relation between terms.

pairwise relation is not symmetric, when t_j is the neighbor of t_i, t_i is not definitely the neighbor of t_j.

Based on the above definition, we can get the neighbor list of each term which contains all the neighbors of that term, and the neighbor lists of t_i and t_j can be expressed as $\Gamma(t_i)$ and $\Gamma(t_j)$. The neighbor list has prune the terms that are not closely related with t_i and t_j, because their contributions to the context information of t_i and t_j are tiny, so it is reasonable to increase the efficiency of link-based relation analysis by pruning these terms.

As follows, we can measure the relation between terms using their neighbor list. Shared neighbors have been widely recognized as the basic evidence to justify the similarity among vertices in a link network [29,30] which are based on the idea that two nodes are similar if their neighbors have large overlap, so they consider each neighbor equally important and also ignores that some neighbors may represent very similar features. In this work, the relation between terms is defined in such a way that the common neighbors do not contribute the same weight but depend on their similarity with the linked terms. Besides, the neighbors only present in one of the neighbor list have a contribution that depends on the similarity of the concepts they represent and the most similar concept in the other neighbor list. For example, “animal”, “skin” and “eyes” are the same neighbors between “bird” and “fish”, so they can build the underlying relation between “bird” and “fish”. Because they are not with the same similarity with “bird” and “fish”, they have the different contribution on the semantic relation between “bird” and “fish”. Besides, “wing” and “fin” are not the same neighbors of “bird” and “fish”, but they are semantically similar and this similarity also contribute to the semantic relation between “bird” and “fish”.

In the following, we present the proposed link-based relation between terms t_i and t_j, which is judged as the similarity of their neighbor information. Assuming term t_k is the neighbor of t_i and t_j at the same time, the relation linked by term t_k is defined as the product of the probabilistic pairwise relation $PPR(t_i, t_k)$ and $PPR(t_k, t_j)$. Besides, if term t_m is the neighbor of t_i and t_n is the neighbor of t_j, and t_m, t_n are semantically similar, the relation between t_i and t_j, which is linked by t_m and t_n, can be calculated by multiplying $PPR(t_i, t_m)$, $PPR(t_j, t_n)$ and the similarity of t_m and t_n. It is defined as follows:

Definition 3.5. Let D be a document collection, a pair of terms t_i and t_j appear in the document collection, then the link-based relation between them according to document collection D is defined as

$$LR(t_i, t_j; D) = \frac{1}{|\Gamma(t_i) \cup \Gamma(t_j)|}\left(\sum_{t_m \in \Gamma(t_i)} PPR(t_i, t_m)\left(\max_{t_n \in \Gamma(t_j)}\{PPR(t_m, t_n)\}\right)PPR(t_n, t_j)\right.$$
$$\left. + \sum_{t_n \in (\Gamma(t_j) - \Gamma(t_i))} PPR(t_i, t_m)\left(\max_{t_m \in \Gamma(t_j)}\{PPR(t_m, t_n)\}\right)PPR(t_n, t_j)\right), \quad (6)$$

where $\Gamma(t_i)$ and $\Gamma(t_j)$ are the neighbors of term t_i and t_j.

The basic assumption in this strategy is the relation between two terms is based on the similarity of their neighbors, as the neighbors can be seen as a semantic description of each term, and more similar neighbors means they are more semantically similar.

Here we also define the link-based relation between t_i and t_j as a probability value that divides the link-based relation between t_i and t_j by the amount of the link-based relation between t_i and other terms.

Definition 3.6. Let D be a document collection, a pair of terms t_i and t_j appear in the document collection, the *probabilistic link-based relation* between them is defined as

$$PLR(t_i, t_j) = \begin{cases} 1 & i = j \\ \dfrac{LR(t_i, t_j; D)}{\sum_{k \neq i} LR(t_i, t_k; D)} & i \neq j \end{cases} \quad (7)$$

With the probabilistic link-based relation measure, $\forall t_i, t_j \in T$, $PLR(t_i, t_j) \in [0, 1]$ and $PLR(t_i, t_i) = 1$. It is also not symmetric such that $PLR(t_i, t_j)$ is not equal to $PLR(t_j, t_i)$.

It is noted that the pairwise relation analysis only captures the relation between terms from their co-occurrence information, so it measures the relation between terms only from the local view. Meanwhile, the link-based relation analysis takes the knowledge of link information into consideration for evaluating the relation between terms, so it measures the relation between terms from the global view. Hence, the link-based relation analysis is also a good candidate to capture the semantic relation between terms.

3.3. The semantic relation between terms

The link-based relation analysis provides the global information into the measure of semantic relation between terms, but only link-based relation is not sufficient, while terms cannot be judged relational if they have little interaction with other terms even they are high pairwise related. To avoid this problem and produce high quality semantic relation, the pairwise and link-based relation are integrated together to analyze the semantic relation between terms in this work.

As the pairwise and link-based relation analysis discover the relation between terms from local and global information respectively, we conclude that they are complementary to each other and the integration of pairwise and link-based relation is defined as the linear combination of them. In the following, the semantic relation between terms is calculated as a weighted average of the two quantities from Eqs. (4) to (7):

Definition 3.7. Let D be a document collection, a pair of terms appear in the document collection, the *semantic relation* between them is defined as

$$SR(t_i, t_j) = (1 - \alpha) \cdot PPR(t_i, t_j) + \alpha \cdot PLR(t_i, t_j), \quad (8)$$

where parameter α controls the weight of pairwise and link-based relation in the calculation of the semantic relation between terms.

As both $PPR(t_i, t_j)$ and $PLR(t_i, t_j)$ fall in the range of [0,1], with $0 \leq \alpha \leq 1$, the value of $SR(t_i, t_j)$ is also between 0 and 1 for all cases. When α is set to 0, the relation measure becomes the pairwise relation, and it becomes the link-based relation when α is 1.

Algorithm 1. Term–term semantic relation measure.

```
Input: Document set $D$, Document dictionary $T$, Term $t_i$, $t_j$
Output: Relation between $t_i$ and $t_j$, $SR(t_i, t_j)$
Procedure PAIRWISE-BASED RELATION MEASURE
    for all document $d_x \in D$ do
        $PR(t_i, t_j; d_x) \leftarrow$ Compute the pairwise relation between $t_i$ and $t_j$ in document $d_x$
    end for
    $PR(t_i, t_j; D) \leftarrow$ Compute the pairwise relation between $t_i$ and $t_j$ in the document set $D$
    $PPR(t_i, t_j) \leftarrow$ Define the pairwise relation in a probabilistic manner
end Procedure

Procedure LINK-BASED RELATION MEASURE
    for all term $t_i \in T$ do
        $\Gamma(t_i) \leftarrow$ Compute the neighbors of $t_i$
    end for
    $LR(t_i, t_j; D) \leftarrow$ Compute the similarity of the neighbors of $t_i$ and $t_j$
$PLR(t_i, t_j) \leftarrow$ Define the link-based relation in a probabilistic manner
```

```
end Procedure

Procedure Combined Relation Measure
    $SR(t_i, t_j) \leftarrow$ Combine the pairwise and link-based relation with an optimal value of $\alpha$
end Procedure
```

3.4. The semantic analysis of documents

In our approach, the semantic analysis of the documents consists of four steps: (1) the pairwise relation between each pair of terms is captured from their co-occurrence information; (2) the link-based relation between terms is captured from their interaction with other terms, proposed in Section 3; (3) and the semantic relation matrix *SR* is then constituted by all the combined relation between each pair of terms in the document set. Algorithm 1 summarizes the computation of the semantic relation between t_i and t_j which entails a series of steps; (4) the original bag-of-word representation is mapped into a new feature space using the captured semantic relation matrix *SR*, which consists of the semantic relation between each pair of terms.

$$d : \overrightarrow{d} \mapsto \overrightarrow{d'} = \overrightarrow{d} * SR \tag{9}$$

In this way, the new feature space takes more reliable features from the original documents into representation than the previous approaches which reveal limited relationship between terms in an explicit way.

Since the pairwise and link-based relation analysis evaluate the relation between terms in different aspects, our proposed approach is more comprehensive than the previous methods. By integrating the semantic relation captured by our approach into the document representation, the similar documents are much easier to identify and can produce better experimental performance. More detailed information will be described in Section 4.

4. Experiment and evaluation

In this section, we present the experimental results of our proposed approach with respect to the clustering task. We compare our approach with BOW, GVSM and the other two prevalent strategies: LSA and HAL, which also capture the relation between terms using their co-occurrence information. We also implement the contextual approach to semantic similarity based on the contextual similarity ratio [19] and Jensen–Shannon distance [20]. The spectral clustering algorithm [31] is used to cluster the enhanced document representation with different semantic relation analysis strategies. In the following, we first describe the document collections and clustering algorithm used in the experiment.

4.1. Experimental data and clustering method

To validate our strategy, we conduct experiments on four document collections. *D*1 is the subset of 20 Newsgroups while *D*2 and *D*3 are the subsets of Reuters 21578, and *D*4 is the WebKB document collection. The detailed information of these document collections is described as follows:

1. The 20 Newsgroups (20NG) document collection [32] is a widely used data set for document clustering, which consists 20,000 newsgroup documents across 20 classes, and we use a subset of 20NG with 1864 documents across 5 classes in our experiment.
2. *D*2 contains 2091 documents with 8 classes. It is a subset of Reuters-21578 document collection [33] which is the most widely used document collection.
3. *D*3 is also a subset derived from the document collection Reuters-21578. *D*3 includes 6245 documents belonging to 52 classes.
4. *D*4 consists of 4087 web pages classified into 4 categories, and it is the WebKB document collection which is collected by the WebKB project of the CMU text learning group [34].

The characteristics of these experimental document collections are summarized in Table 1. *m*, *n* is the number of documents and terms respectively, and n_{avg} is the average number of terms per document.

In order to demonstrate the quality of document representation based on different strategies, we employ the spectral clustering algorithm for document clustering.

Spectral clustering is one of the most popular modern clustering algorithms and usually outperforms traditional clustering algorithms such as the *k*-means algorithm. Spectral clustering makes use of the eigenvalues of the similarity matrix of the data to perform dimensionality reduction, and it consists of a few basic steps for clustering. It begins with the construction of the similarity matrix *S* whose elements are the similarities between each pair of points in data set. In our approach, the elements of *S* are the cosine similarities between each pair of documents. Then *S* is transformed into a Laplacian matrix by a multiplication with a diagonal matrix *D*, which is a diagonal matrix consisting of the sums of the rows of *S*. The transformation of *S* is as following:

$$L = D^{-1/2} S D^{1/2} \tag{10}$$

Then the columns of *S* correspond to the *k* largest eigenvectors of *L* construct the new feature space *X* whose *i*th row is the vector representation of the *i*th document. After normalizing the rows of *X* to the unit length, the standard *k*-means algorithm is finally used to cluster the rows of matrix *X*.

4.2. Evaluation criteria

The quality of document clustering is evaluated by four criteria: Purity, 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

Table 1
Characteristics of data sets.

Data sets	Topics	Classes	*m*	*n*	n_{avg}
*D*1	20-NGs: atheism, graphics, windows.misc, pc.hardware, mac.hardware	5	1864	16516	76
*D*2	Reuters-21,578: acq, crude, earn, grain, interest, money-fx, ship, trade	8	2091	8674	33
*D*3	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	6245	16,142	36
*D*4	WebKB: course, faculty, project, student	4	4087	7769	32

the cluster ($\max_j(|c_{ij}|)$) and the size of cluster ($|c_i|$): $purity(c_i) = 1/|c_i| \max_j |c_{ij}|$. Then the overall purity can be expressed as the weighted sum of all individual cluster purity:

$$purity = \sum_{i=1}^{k} \frac{|c_i|}{N} purity(c_i), \tag{11}$$

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:

$$RI = \frac{TP+TN}{TP+TN+FP+FN}, \tag{12}$$

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 F_1-measure. It is a criterion considering both the precision and recall for clustering evaluation according to the following formula:

$$F_1 = 2 \times \frac{precision \times recall}{precision + recall}, \tag{13}$$

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 is 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.

$$NMI(C, L) = \frac{I(C;L)}{(H(C)+H(L))/2}, \tag{14}$$

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:

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

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

$$H(C) = -\sum_i \frac{|c_i|}{N} \log \frac{|c_i|}{N}, H(L) = -\sum_j \frac{|l_j|}{N} \log \frac{|l_j|}{N}, \tag{16}$$

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

4.3. Experimental results

In this section, we compare the document representation based on our strategy (CPL) with GVSM, LSA, HAL, and the contextual similarity approaches based on the ratio of context similarity (CSR) and the Jensen–Shannon distance (JSD). The bag-of-words (BOW) approach is used as the baseline for comparison. The experimental results in Tables 2–5 illustrate the Purity, RI, F_1-measure and NMI scores, respectively, which are computed from the clustering results with respect to different representation strategies on the four document collections. In order to obtain a precise comparison, all methods are initialized using the same random document seeds, and all the experiments are repeated 10 times. In these tables, the columns with heading "avg" and "best" present the average and best scores of 10 runs, respectively, while the column with heading "imp" presents values for the relative improvement compared with the baseline, and the column with heading "t" presents the significance of the performance difference between our method and the compared methods.

Notice that, for our strategy, we use the values of parameter α which achieve the best performance for each document collection, are 0.3, 0.7, 0.6 and 0.5 for $D1$, $D2$, $D3$ and $D4$, respectively. Besides, in order to find the right θ which determines whether two terms are neighbors, we tried the average value of the pairwise relation on each document collections as suggested in [35]. In fact, we also tried the value from 0 to 1 with increment of 0.1, and the experimental results show that when θ is the average value, the clustering performance can achieve nearly the best performance on all document collections. In these experiments, the parameters of other approaches are also well-tuned. Based on the results of the parameter exploration, we use the optimal 300-factor space for LSA and 10 words window size for HAL, which achieves the best performance across the four experimental document collections.

In comparison with the results of baseline, our strategy results in significant improvement in all four experimental document collections. It achieves 8.1% improvement on the average Purity score of the four document collections, and 5.6% improvement on the average RI score, 7.1% improvement on the average $F1$-measure score and 7.1% improvement on the average NMI score. Comparing with GVSM, LSA, HAL, CSR and JSD, our strategy also performs better in all experimental collections. It also achieves 3.6%, 3.6%, 3.0%, 2.2% improvement over LSA, 3.9%, 3.5%, 3.3%, 3.4% improvement over JSD, 4.8%, 4.0%, 3.8%, 3.0% improvement over CSR, 4.7%, 3.9%, 3.5%, 4.7% improvement over HAL and 5.0%, 4.7%, 5.1%, 5.3% over GVSM on the average Purity, RI, F_1-measure and NMI scores of the four document collections. The experimental results demonstrate the benefit of combining the pairwise and link-based relation to capture the semantic relation between terms.

It is interesting to notice that for the document collection $D1$, the bag-of-word approach gets rather poor score on this collection (Purity: 0.563, RI: 0.784, F_1-measure: 0.482, NMI: 0.415), whereas our strategy obtains a significant improvement (Purity: 0.725, RI: 0.837, F_1-measure: 0.595, NMI: 0.530). Besides, on the

Table 2
The purity scores for different document representation model using spectral clustering.

Method	$D1$				$D2$				$D3$				$D4$			
	avg	*best*	*imp* (%)	*t*	*avg*	*best*	*imp* (%)	*t*	*avg*	*best*	*imp* (%)	*t*	*avg*	*best*	*imp* (%)	*t*
BOW	0.563	0.566	–	89.3	0.811	0.832	–	4.03	0.798	0.808	–	4.39	0.669	0.686	–	3.05
GVSM	0.669	0.682	10.6	2.43	0.821	0.840	1.0	4.86	0.801	0.811	0.3	6.83	0.672	0.673	0.3	3.21
HAL	0.678	0.690	11.5	2.54	0.823	0.833	1.2	2.90	0.802	0.809	0.4	5.13	0.673	0.689	0.4	2.83
CSR	0.671	0.683	11.2	2.38	0.819	0.836	0.8	3.12	0.809	0.817	1.1	4.87	0.671	0.683	0.2	3.09
JSD	0.681	0.690	11.8	2.02	0.838	0.859	2.6	2.73	0.812	0.819	1.4	4.39	0.679	0.689	1.0	1.98
LSA	0.684	0.698	12.1	1.93	0.850	0.868	3.9	3.14	0.809	0.827	1.1	3.85	0.676	0.687	0.7	2.13
CPL	0.746	0.746	**18.3**	0.00	0.869	0.886	**5.8**	0.00	0.845	0.850	**4.7**	0.00	0.704	0.726	**3.5**	0.00

Table 3
The RI scores for different document representation model using spectral clustering.

Method	*D*1				*D*2				*D*3				*D*4			
	avg	*best*	*imp* (%)	*t*	*avg*	*best*	*imp* (%)	*t*	*avg*	*best*	*imp* (%)	*t*	*avg*	*best*	*imp* (%)	*t*
BOW	0.784	0.785	–	93.1	0.722	0.735	–	4.35	0.753	0.755	–	4.12	0.733	0.734	–	3.89
GVSM	0.815	0.819	3.1	2.09	0.719	0.739	–	3.06	0.752	0.755	–	3.89	0.738	0.740	0.5	2.32
HAL	0.821	0.823	3.7	2.36	0.741	0.771	1.9	3.25	0.758	0.759	0.3	3.35	0.742	0.745	0.9	2.53
CSR	0.818	0.824	3.4	2.49	0.733	0.768	1.1	3.43	0.756	0.757	0.1	3.39	0.740	0.749	0.7	2.47
JSD	0.825	0.828	4.1	2.03	0.747	0.780	2.5	3.36	0.755	0.761	0.2	3.78	0.743	0.747	1.0	2.29
LSA	0.824	0.829	4.0	1.91	0.749	0.763	2.7	2.79	0.759	0.785	0.4	2.98	0.744	0.744	1.1	1.78
CPL	0.837	0.837	**9.3**	0.00	0.782	0.790	**6.0**	0.00	0.797	0.804	**4.4**	0.00	0.761	0.767	**2.8**	0.00

Table 4
The F_1-measure scores for different document representation model using spectral clustering.

Method	*D*1				*D*2				*D*3				*D*4			
	avg	*best*	*imp* (%)	*t*	*avg*	*best*	*imp* (%)	*t*	*avg*	*best*	*imp* (%)	*t*	*avg*	*best*	*imp* (%)	*t*
BOW	0.482	0.488	–	24.0	0.468	0.509	–	6.32	0.128	0.135	–	4.09	0.518	0.560	–	4.23
GVSM	0.552	0.559	7.0	2.55	0.446	0.504	–	5.87	0.127	0.138	–	3.08	0.526	0.526	0.8	2.82
HAL	0.573	0.576	9.1	2.30	0.502	0.586	3.4	2.98	0.132	0.139	0.4	3.04	0.531	0.562	1.3	2.82
CSR	0.568	0.573	8.6	2.69	0.499	0.543	3.1	3.14	0.131	0.135	0.3	3.32	0.528	0.549	1.0	2.78
JSD	0.572	0.581	9.0	2.46	0.505	0.546	3.7	2.31	0.136	0.148	0.8	1.97	0.539	0.565	2.1	2.45
LSA	0.578	0.587	9.6	1.36	0.507	0.548	3.9	2.27	0.135	0.149	0.7	2.11	0.541	0.570	2.3	1.78
CPL	0.595	0.595	**11.3**	0.00	0.571	0.627	**10.3**	0.00	0.151	0.184	**2.3**	0.00	0.562	0.575	**4.4**	0.00

Table 5
The NMI scores for different document representation model using spectral clustering.

Method	*D*1				*D*2				*D*3				*D*4			
	avg	*best*	*imp* (%)	*t*	*avg*	*best*	*imp* (%)	*t*	*avg*	*best*	*imp* (%)	*t*	*avg*	*best*	*imp* (%)	*t*
BOW	0.415	0.416	–	81.8	0.489	0.511	–	5.22	0.480	0.489	–	4.89	0.379	0.405	–	4.32
GVSM	0.481	0.489	6.6	2.49	0.481	0.509	–	4.87	0.481	0.492	0.1	4.29	0.385	0.390	0.6	3.86
HAL	0.472	0.520	5.7	2.98	0.515	0.550	2.6	3.02	0.486	0.490	0.6	3.21	0.388	0.401	0.9	3.18
CSR	0.478	0.523	6.3	2.37	0.523	0.548	3.4	2.83	0.483	0.472	0.3	3.04	0.391	0.413	1.2	3.01
JSD	0.495	0.519	8.0	1.89	0.537	0.552	4.8	2.48	0.491	0.503	1.1	2.32	0.390	0.408	1.1	2.89
LSA	0.511	0.521	9.6	1.28	0.566	0.579	7.7	2.13	0.489	0.497	0.9	2.87	0.394	0.404	1.5	2.01
CPL	0.530	0.530	**11.5**	0.00	0.568	0.609	**7.9**	0.00	0.518	0.520	**3.8**	0.00	0.431	0.453	**5.2**	0.00

collections which has achieved good performance with the bag-of-words approach, we also achieve significant gains. Like for *D*2, the bag-of-words approach achieves rather good score (Purity: 0.811, RI: 0.722, F_1-measure: 0.468, NMI: 0.489), we also achieved the gains of 5.8%, 6.0%, 10.3% and 7.9% on the average Purity, RI, F_1-measure and NMI scores, respectively.

The results of LSA also discover a potential improvement which obtain fairly high average scores on Purity, RI, F_1-measure and NMI (Purity: 0.755, RI: 0.769, F_1-measure: 0.440, NMI: 0.490). However, there exists some collections which achieve lower improvement by using LSA. We conclude that the major reason affects the performance is the lack of an effective measure to evaluate the underlying relation between terms. In fact, the LSA computes the relation between terms by only considering their explicit co-occurrence information, and it does not concern whether two terms are semantically related even though they have the similar context with other terms. In other word, the semantic relation in the global view has been overlooked. Thus, the LSA cannot capture the robust and accurate semantic relation between terms from corpus. Besides, the HAL strategy, which evaluates the relation between terms by their interaction with other terms but overlooks their inherent relation, also performs better than the baseline. It achieves a performance improvement of 3.4% on the average Purity score of the four document collections, and an improvement of 1.5% on RI score, an improvement of 3.3% on the F_1-measure score and an improvement of 2.3% on the NMI score. HAL is well suited to capture the positional similarities between words. However, in capturing positional information, the HAL model is largely insensitive to the types of information to which LSA is sensitive. The other contextual similarity-based approaches, like CSR and JSD, have the same limitation as HAL. Therefore, an optimal combination of the pairwise and link-based relation in our work is a comprehensive approach to capture the precise semantic relation between terms, which considers the relation between term not only in the local view but also the global view, and that is why it can perform much better than the previous approaches, like HAL and LSA.

In order to illustrate the statistical significance of the obtained results, the well-known *t*-test was applied for each data set to determine the significance of the performance difference between our method and the compared methods. Within a confidence interval of 95% and for the value of degrees of freedom equal to 18 (for two sets of 10 experiments each), the critical value for *t* is $t_c = l.734$. This means that if the computed $t \geq t_c$, then the null hypothesis is rejected, i.e., our method is superior, otherwise the null hypothesis is accepted. As it can be observed from the results of the statistical tests for spectral clustering presented in Tables 2–5, the performance superiority of our method is clearly significant in most cases with respect to all other methods, which demonstrates that our method achieves significantly better results than the compared representations.

4.4. The effect of pairwise and link-based relation

In order to better understand the effect of pairwise and link-based relation, we conduct experiments with different values of parameter α, which controls the weight of pairwise and link-based relation in the semantic relation analysis. In this experiment, we evaluate the clustering performance by setting the parameter α from 0 to 1 at increment of 0.1. The Purity, RI, F_1-measure and NMI scores computing from clustering results with different parameter values on the four collections are presented in Fig. 2. The BOW line depicts the performance of bag-of-word approach, while the CPL curve shows that the performance of our strategy varies along with different values of α.

From Fig. 2, we can observe that the pairwise and link-based measure solely can perform better than the baseline approach in most cases, which illustrates that the measure scheme for the pairwise and link-based relation are effective to capture the relation between terms. With the varying of the value of α, we observe that the linear combination of pairwise and link-based relation measure performs superior than using either pairwise or link-based relation measure alone in most cases, and it can achieves better performance with an optimal value of α. For example, the performance of our approach in Fig. 2(a) always performs better than the baseline, and it achieves the best performance at $\alpha = 0.3$. The Purity, RI, F_1-measure and NMI score at this point is 0.746, 0.837, 0.595 and 0.530, respectively. This is a relative improvement of 7.9%, 2.6%, 4.5% and 3.4% over the pairwise measure solely, and of 13.4%, 4.7%, 10.6% and 10.3% over the link-based measure solely. The similar phenomena have been observed in Fig. 2(b)–(d), in which the trend of curves is similar to that in Fig. 2(a). The only difference between them is they achieve the best performance on different certain points.

In general, the results demonstrate that both the pairwise and link-based relation have great impact on the evaluation of relation between terms, and an optimal combination of them is essential to capture more reliable semantic relation between terms.

4.5. Parameter tuning using cross-validation

To choose the optimal value of α automatically, in this work, the 10-fold cross validation is exploited which has been proved an efficient way for parameter tuning [36]. The general idea of this method is to randomly partition the overall sample into 10 equal size folds. Of the 10 folds, a single fold is leaved out for testing while the remaining 9 folds (training samples) are used to perform clustering analysis, and the results of the analyses are applied to the testing sample (the sample that was not used to estimate the parameters) to compute some index of predictive validity. The results for the 10 replications are aggregated (averaged) to yield a single measure of the stability of the respective model, i.e., the validity of the model for predicting new observations.

Cluster analysis is an unsupervised learning technique, and we cannot observe the real label of data. However, it is reasonable to replace the usual notion of accuracy with that of Davies-Bouldin index (DBI) [37], which is a metric for evaluating clustering performance. The lower the DBI value, the better the separation of the clusters and the tightness inside the clusters. In general, we can apply the 10-fold cross-validation method to each value of the parameter α in clustering, which is from 0 to 1 with increment of 0.1, and observe the DBI value of the testing samples where a smaller DBI value indicates a better validation performance. The value of α achieving the best validation performance is then selected as the estimate of α.

The results of 10-fold cross-validation on different parameter values for all document collections are shown in Fig. 3. We can see that there are considerable difference between the validation performance across the tuning parameter considered. Hence, the parameter tuning is important for choosing the best parameter value to improve clustering performance. With the four document collections, the best validation performance is achieved at $\alpha = 0.4, 0.7, 0.6$ and 0.5 respectively. Meanwhile, the parameter values chosen by cross-validation on all document collections are nearly same as the best score in the above experiment, which proves that the cross-validation is valid for the choice of the best parameter value.

4.6. Analyzing the best performance

To achieve the best performance, an optimal combination of pairwise and link-based relation is necessary. It is interesting to notice that the best performance on different document collections is achieved with the different values of α, which determines the weight of pairwise and link-based relation in the calculation of the semantic relation between terms. In our approach, 10-fold cross validation is used to optimize the setting of parameter α for different document collections.

Besides, from the experimental results, we observe that our strategy achieves quite different gains on the four collections. It works well on the document collections $D1$ and $D2$, and has less advantage on $D3$ and $D4$. We conclude that it is partially due to the difference of the collection characteristics. To the collections with better gains, $D1$ and $D2$, the documents are probably more semantically alike than in $D3$ and $D4$. That means the different words in these collections are used in more or less the same context. In such a situation, the bag-of-word is too general and will not recognize the small difference between documents. In this sense, our approach has advantage by considering the semantic relation into document representation.

In summary, to achieve the best performance, an optimal value of α is required to combine the pairwise and link-based relation, which is influenced by the distribution of the document collection. Meanwhile, our approach performs quite differently on different collections. It works much better in the collections which are more semantically alike, in which the difference between documents cannot be recognized using the word matching approach.

4.7. Complexity and scalability analysis

Besides the previous quality assessments, computational time requirements of the proposed approach is discussed here. For the proposed semantic relation analysis, the time complexity of calculating the pairwise relation is $O(nm^2)$, while the time complexity of calculating the link-based relation is $O(m^3)$, where n is the number of documents and m denotes the number of terms. Hence, the time complexity of calculating the semantic matrix S is $O(nm^2 + m^3)$. It is noted that the proposed semantic relation analysis is with more computational complexity than the previous approaches as it considers the relation between terms by themselves and also their interaction with other terms. But we think that it is worthy as the clustering performance has been greatly improved. Besides, our model can be easily and effectively parallelized, which can lead to faster execution times on large document collections given an available parallel system.

In addition, we also study the effect of the document collection size on the capturing of the semantic relation with our strategy. In order to show the performance varying with different document collection sizes, we conduct a set of experiments on the document collection $D4$ by increasing the number of documents from 1000

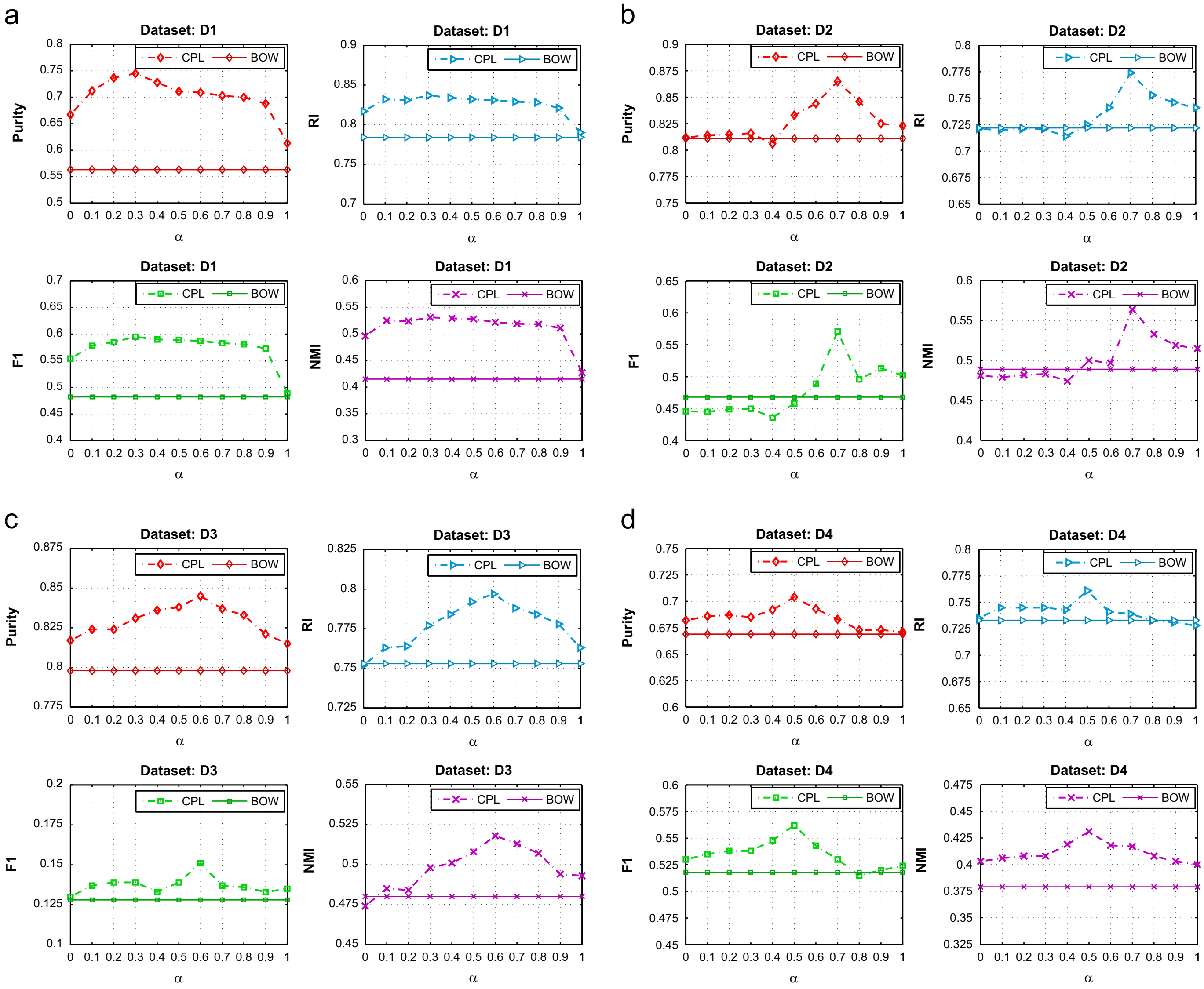


Fig. 2. The effect of varying the parameter α on the clustering performance of document collection $D1$, $D2$, $D3$, $D4$.

to 4000 at increment of 1000 documents. The experimental results are shown in Fig. 4, which illustrate the Purity, RI, F_1-measure, and NMI scores vary along with the size of document collections. It is interesting to notice that the clustering performance achieves significant gains comparing with the baseline on the small collection with 1000 documents. Meanwhile, with the increasing of the document collection size, the performance of document clustering is with a little higher improvement. We conclude that it is because the semantic information is more reliable to be captured with the larger size of document collection. In summary, the curve of the scores is basically stable with the increasing of the document collection size, which means that our strategy can capture the semantic relation between terms precisely on different sizes of document collection, even on the small document collection.

4.8. Summary of the analysis

In summary, our analysis on the experimental results shows that, with the integration of pairwise and link-based relation analysis, our strategy captures more precise relation between terms than the prevalent strategy GVSM, LSA and HAL. By integrating the captured semantic relation into the document representation, we achieve different gains on the performance of document clustering on the different document collections. On the other hand, the parameter α, which determines the weight of pairwise and link-based relation, needs to be optimized to achieve the best performance, and the optimal values of the parameter α are not identical on the different document collections. We conclude that the optimal value depends on the inherent structure of each document collection. Finally, the quality of our strategy is stable on the performance of document clustering with the increasing of the document size, and it can also achieve significant improvement even on the small document collections.

5. Conclusion and future work

In this paper, we present a novel strategy to capture the semantic relation between terms, which is based on the integration of pairwise and link-based relation. Our approach operates in a sequence of three steps: (1) capture the pairwise relation between terms from the original documents using their co-occurrence information. (2) Capture the link-based relation between terms based on their neighbor information. (3) Integrate the pairwise and link-based relation by an optimal value of parameter α to capture the precise relation between terms.

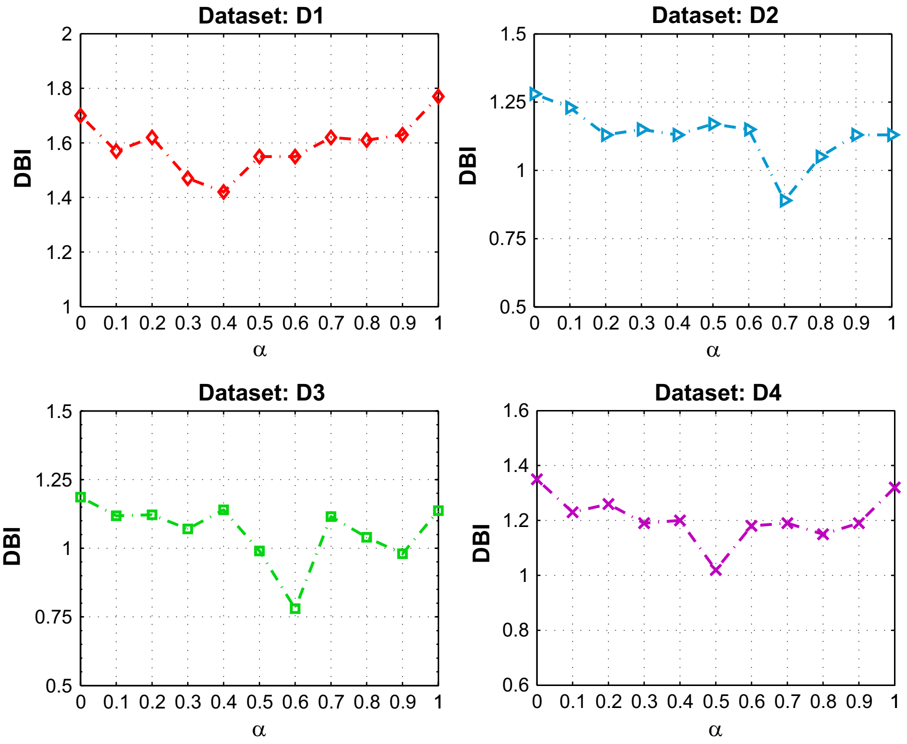


Fig. 3. The tuning of parameter α using cross-validation.

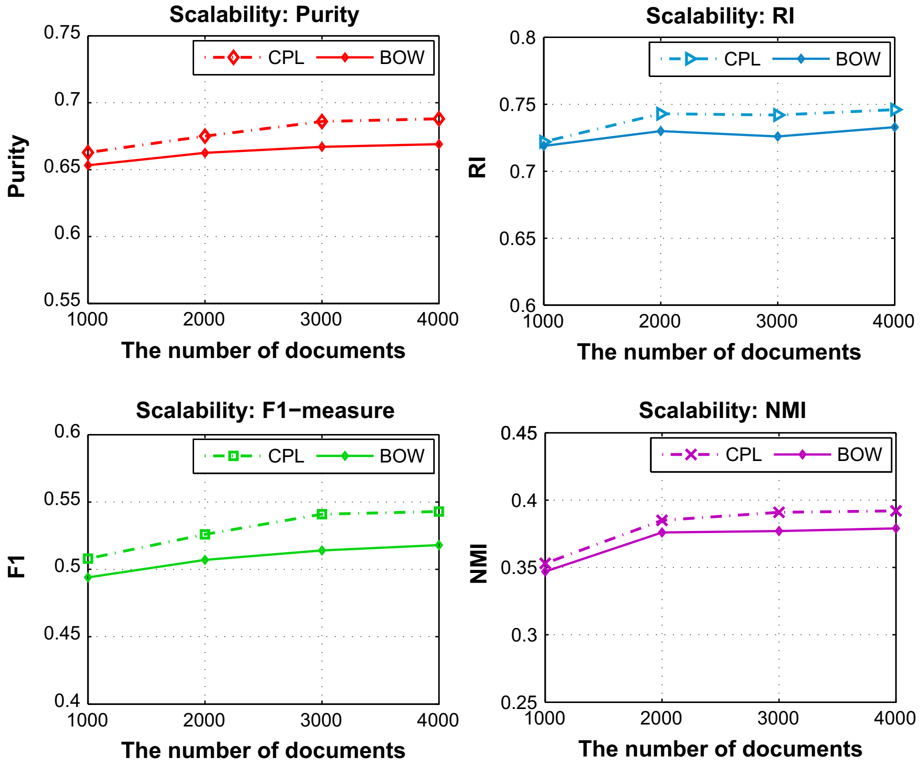


Fig. 4. The effect of varying the document collection size on the clustering performance of document collection *D*4.

In the experimental section, we conduct experiment over the collection of *D*1, *D*2, *D*3 and *D*4 using different relation measure strategies. The experimental results show that our proposed strategy significantly outperforms than the prevalent strategies GVSM, LSA and HAL. It demonstrates that the strategy proposed in our work can capture more precise semantic relation between terms than the previous approaches, and we conclude that it is because we take both the pairwise and link-based relation into account to measure term–term semantic relation.

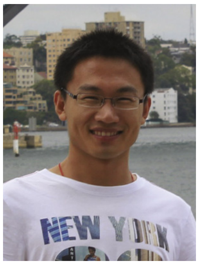
There are a number of future research directions to extend and improve this work. One direction that this work might continue on is to improve on the accuracy of relation calculation between term by employing different term calculation strategies, although the current scheme proved more accurate than traditional methods, there is still room for improvement. Besides, the linear combination of the pairwise and link-based relation in our paper was intuitively and empirically derived, and we believe that our approach may be improved further if a better combination function is found.

Acknowledgment

This work is supported by The National Natural Science Foundation of China (No. 61075056, 61273304), and this work is also supported in part by The Fundamental Research Funds for the Central Universities.

References

[1] R.K. Srihari, Z. Zhang, A. Rao, Intelligent indexing and semantic retrieval of multimodal documents, Inf. Retr. 2 (2–3) (2000) 245–275.
[2] H. Billhardt, D. Borrajo, V. Maojo, A context vector model for information retrieval, J. Am. Soc. Inf. Sci. Technol. 53 (3) (2002) 236–249.
[3] A. Budanitsky, G. Hirst, Evaluating wordnet-based measures of lexical semantic relatedness, Comput. Linguist. 32 (1) (2006) 13–47.
[4] L. Wenyin, X. Quan, M. Feng, B. Qiu, A short text modeling method combining semantic and statistical information, Inf. Sci. 180 (20) (2010) 4031–4041.
[5] P. Resnik, Semantic similarity in a taxonomy: an information-based measure and its application to problems of ambiguity in natural language, J. Artif. Intell. Res. 11 (1999) 95–130.
[6] P.D. Turney, Mining the web for synonyms: PMI-IR versus LSA on TOEFL, in: Proceedings of the 12th European Conference on Machine Learning, Springer-Verlag, 2001, pp. 491–502.
[7] H. Schütze, Automatic word sense discrimination, Comput. Linguist. 24 (1) (1998) 97–123.
[8] S.T. Dumais, Latent semantic analysis, Ann. Rev. Inf. Sci. Technol. 38 (1) (2004) 188–230.
[9] S.C. Deerwester, S.T. Dumais, T.K. Landauer, G.W. Furnas, R.A. Harshman, Indexing by latent semantic analysis, J. Am. Soc. Inf. Sci. 41 (6) (1990) 391–407.
[10] K. Lund, C. Burgess, Producing high-dimensional semantic spaces from lexical co-occurrence, Behav. Res. Methods Instrum. Comput. 28 (2) (1996) 203–208.
[11] F. Figueiredo, L. Rocha, T. Couto, T. Salles, M.A. Gonçalves, W. Meira Jr, Word co-occurrence features for text classification, Inf. Syst. 36 (5) (2011) 843–858.
[12] A.K. Farahat, M.S. Kamel, Statistical semantics for enhancing document clustering, Knowl. Inf. Syst. 28 (2) (2011) 365–393.
[13] T. Hofmann, Probabilistic latent semantic indexing, in: Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM, 1999, pp. 50–57.
[14] C. Burgess, Representing and resolving semantic ambiguity: a contribution from high-dimensional memory modeling., in: On the Consequences of Meaning Selection: Perspectives on Resolving Lexical Ambiguity, American Psychological Association, 2001.
[15] C. Burgess, K. Lund, Representing abstract words and emotional connotation in a high-dimensional memory space, in: Proceedings of the Cognitive Science Society, 1997, pp. 61–66.
[16] C. Audet, C. Burgess, et al., Using a high-dimensional memory model to evaluate the properties of abstract and concrete words, in: Proceedings of the Cognitive Science Society, Erlbaum, Mahwah, NJ, 1999, pp. 37–42.
[17] Curt Burgess, Kevin Lund, Modelling parsing constraints with high-dimensional context space, Lang. Cogn. Process. 12 (2–3) (1997) 177–210.
[18] P. Li, C. Burgess, K. Lund, The acquisition of word meaning through global lexical co-occurrences, in: Proceedings of the Thirtieth Annual Child Language Research Forum, 2000, pp. 166–178.
[19] G.A. Miller, W.G. Charles, Contextual correlates of semantic similarity, Lang. Cogn. Process. 6 (1) (1991) 1–28.
[20] I. Dagan, L. Lee, F.C. Pereira, Similarity-based models of word cooccurrence probabilities, Mach. Learn. 34 (1–3) (1999) 43–69.
[21] T. Mikolov, I. Sutskever, K. Chen, G.S. Corrado, J. Dean, Distributed representations of words and phrases and their compositionality, in: Advances in Neural Information Processing Systems, 2013, pp. 3111–3119.
[22] R. Collobert, J. Weston, A unified architecture for natural language processing: deep neural networks with multitask learning, in: Proceedings of the 25th International Conference on Machine Learning, ACM, 2008, pp. 160–167.
[23] A. Mnih, G.E. Hinton, A scalable hierarchical distributed language model, in: Advances in Neural Information Processing Systems, 2009, pp. 1081–1088.
[24] S. Wong, W. Ziarko, P. Wong, Generalized vector spaces model in information retrieval, in: SIGIR 1985, ACM, 1985, pp. 18–25.
[25] G.A. Miller, Wordnet: a lexical database for English, Commun. ACM 38 (11) (1995) 39–41.
[26] E. Gabrilovich, S. Markovitch, Computing semantic relatedness using Wikipedia-based explicit semantic analysis., in: IJCAI, vol. 7, 2007, pp. 1606–1611.
[27] X. Hu, X. Zhang, C. Lu, E.K. Park, X. Zhou, Exploiting Wikipedia as external knowledge for document clustering, in: Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, 2009, pp. 389–396.
[28] P. Wang, J. Hu, H.-J. Zeng, Z. Chen, Using Wikipedia knowledge to improve text classification, Knowl. Inf. Syst. 19 (3) (2009) 265–281.
[29] L. Getoor, C.P. Diehl, Link mining: a survey, ACM SIGKDD Explor. Newsl. 7 (2) (2005) 3–12.
[30] D. Liben-Nowell, J. Kleinberg, The link-prediction problem for social networks, J. Am. Soc. Inf. Sci. Technol. 58 (7) (2007) 1019–1031.
[31] A.Y. Ng, M.I. Jordan, Y. Weiss, On spectral clustering: analysis and an algorithm, Adv. Neural Inf. Process. Syst. 2 (2002) 849–856.
[32] K. Lang, Newsweeder: learning to filter netnews, in: Proceedings of the Twelfth International Conference on Machine Learning, Citeseer, 1995.
[33] D.D. Lewis, Reuters-21578 Text Categorization Test Collection, Distribution 1.0, ⟨http://www.research.att.com/∼lewis/reuters21578.html⟩.
[34] M. Craven, D. DiPasquo, D. Freitag, A. McCallum, T. Mitchell, K. Nigam, S. Slattery, Learning to extract symbolic knowledge from the world wide web, in: Proceedings of the 15th National Conference on Artificial Intelligence, 1998.
[35] C. Luo, Y. Li, S.M. Chung, Text document clustering based on neighbors, Data Knowl. Eng. 68 (11) (2009) 1271–1288.
[36] R. Kohavi, A study of cross-validation and bootstrap for accuracy estimation and model selection, in: International Joint Conference on Artificial Intelligence, vol. 14, 1995, pp. 1137–1145.
[37] D.L. Davies, D.W. Bouldin, A cluster separation measure, in: IEEE Transactions on Pattern Analysis and Machine Intelligence (2) (1979) 224–227.



Xin Cheng received a Bachelor degree in Computer Science from Southern Yangtze University, China. He is at present a Ph.D. candidate in the Department of Computer Science and Technology, Tongji University, China. His research interest is in knowledge discovery from document data. His current research interests focus on semantic relation analysis for document clustering based on statistic-based analysis.

Duoqian Miao received his Ph.D. in Pattern Recognition and Intelligent System at the Institute of Automation, Chinese Academy of Sciences. He is a Professor and Vice Dean of the School of Electronics and Information Engineering at Tongji University. He has published over 180 scientific articles in international journals, books, and conferences. His present research interests include rough sets, granular computing, principal curve, web intelligence, and data mining.

Can Wang received the B.Sc. and M.Sc. degrees from the Faculty of Mathematics and Statistics, Wuhan University, China, in 2007 and 2009, respectively, and the Ph.D. degree in computing sciences from the Advanced Analytics Institute, University of Technology, Australia, in 2013. She is currently a Post-Doctoral Fellow with the Commonwealth Scientific and Industrial Research Organization, Australia. Her current research interests include behavior analytics, data mining, machine learning, and knowledge representation.