Short Text Feature Extension Based on Improved Frequent Term Sets

Huifang Ma(✉), Lei Di, Xiantao Zeng, Li Yan, and Yuyi Ma

College of Computer Science, Northwest Normal University,
Lanzhou, Gansu, China
mahuifang@yeah.net

Abstract. A short text feature extension algorithm based on improved frequent word set is proposed. By calculating support and confidence, the same category tendencies of frequent term sets are extracted. Correlations based frequent term sets are defined to further extend the term set. Meanwhile, information gain is introduced to traditional TF-IDF, better expressing the category distribution information and the weight of word for each category is enhanced. All term pairs with external relations are extracted and the frequent term set is expanded. Finally, the word similarity matrix is constructed via the frequent word set, and the symmetric non-negative matrix factorization technique is applied to extend the feature space. Experiments show that the constructed short text model can improve the performance of short text clustering.

Keywords: Term weighing · Information gain · Frequent term set · Correlation Non-negative matrix factorization

1 Introduction

In recent years, with the development of technology in Web 2.0, short texts, such as short messages, microblogs, and news comments, increase in a geometrical ratio. Unlike traditional texts, some inherent characteristics of short texts, such as extremely feature sparsity and highly unbalancing samples, hinder the traditional approaches for long texts being easily applied.

To extend short text feature, the most recent popular researches have been mainly focused on three aspects. Firstly, some researchers try to use language models, grammar and syntax analysis method to obtain more specific semantic information [1–3]. Secondly, some researchers take advantage of statistical approaches, such as global term context vectors, coupled term-term relations for short text extension [4, 5]. Lastly, both semantic information obtained from a hierarchical lexical database and statistical information contained in the corpus are involved for short text extension [6].

This paper proposes a short text feature extension strategy based on improved frequent term sets. We mainly focus on news titles and take news content as background knowledge. The feature extension algorithm is designed to extract frequent term sets and build the word similarity matrix, based on which to extend feature space of short text. The overview of our framework is shown in Fig. 1. Firstly, by calculating support and confidence, double term sets with co-occurring relation and identical class

Z. Shi et al. (Eds.): IIP 2016, IFIP AICT 486, pp. 169–178, 2016.
DOI: 10.1007/978-3-319-48390-0_18

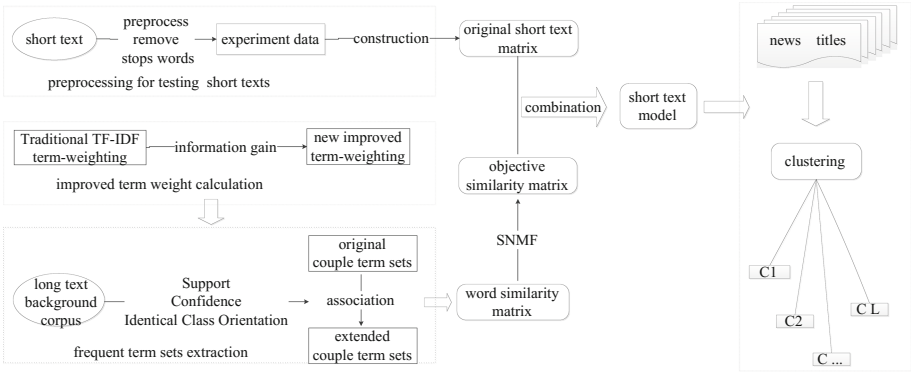

Fig. 1. Algorithm framework

orientation are extracted from long text corpus. Then we extend these frequent term sets with external association. Meanwhile, the information gain is introduced to traditional TF-IDF, better expressing the category distribution information and the weight of words for each category is enhanced. Finally, the word similarity matrix is constructed via the frequent word sets, and symmetric non-negative matrix factorization (SNMF) technique is used to extend the feature space.

2 Frequent Term Sets Extraction

We briefly introduce some concepts and notations employed in frequent term sets extraction in Table 1. Support is defined as the number of documents which contain term set T dividing the total number of documents in data set while confidence is defined as the number of documents containing t in class c dividing the number of all documents involving t [7].

To select frequent term sets efficiently, several other concepts are needed:

Definition 1 (Co-occurring relation). If the support of term set T surpasses the threshold α ($0 < \alpha < 1$), T is considered as a frequent term set and all terms in T have a Co-occurring relation.

Definition 2 (Class orientation). For term t and class $c \in C$, if conf(t,c) surpasses the threshold $\beta(0.5 \leq \beta < 1)$, term t has a class orientation to c, formulated as Tendency (t) = c.

Definition 3 (Identical Class Orientation). For two terms t1 and t2, if there is a class c, Tendency(t1) = c and Tendency(t2) = c, then t1 and t2 have an Identical Class Orientation.

In order to obtain more semantic information, we extract frequent term sets with both co-occurring relation and identical class orientation.

Table 1. Notation definition

Notation	Meaning
$\boldsymbol{D} = \{d_1,d_2,\ldots,d_M\}$	$\boldsymbol{D}$: the collection of training set d_i :the ith document in D M : total number of documents in D
N	the number of words appearing one and only one time in D
$\boldsymbol{W}$, $\boldsymbol{W}_e$	$\boldsymbol{W}$: the expression matrix of $\boldsymbol{D}$ $\boldsymbol{W}_e$: documents matrix after extension
$\boldsymbol{T} = \{t_1,t_2,\ldots,t_n\}$	$\boldsymbol{T}$: a term set in background knowledge-base t_i : a feature term in $\boldsymbol{T}$
$\boldsymbol{S}$, $\boldsymbol{P}$	$\boldsymbol{S}$: words similarity matrix $\boldsymbol{P}$: matrix factorized by $\boldsymbol{S}$
$\boldsymbol{C} = \{c_1,c_2,\ldots,c_L\}$	$\boldsymbol{C}$: the set of categories C_i : the ith category in C L : total number of categories
R, R_e	R :the original frequent couple term set R_e :frequent couple term sets after extension
K	the number of words appearing one and only one time in R

Table 2. Double frequent term sets extraction algorithm

Algorithm1. Double frequent term sets extraction algorithm

Input: Feature set F、Support α、Confidence β、 and the collection C of class.

Output: Frequent couple term set R、total number K of words in R

1: Initialize R as empty, K = 0

2: For every term t_i in F, if Support(t_i) *and* $conf(t_i,c_j)$ is no less than α *and* β respectively, then add t_i into class c_j

3: For every class in the category collection,calculate the Support between any two terms in this class,if Support $\geqq \alpha$, then put this pair of terms into R,and K should increase corresponding number.

4: Return R,K.

As co-occurring relation can perfectly represent semantic association between terms, while terms with identical class orientation can be potentially from the same or close topic, feature extension of these terms is expected to have better discriminative ability. Considering that there are often 2 words for most of Chinese phases, this paper focuses on extracting double frequent term sets. The algorithm is described in Table 2 [7], and the feature set F represents the collection of words in background knowledge.

3 Word Similarity Matrix Construction

3.1 Improved Term Weighing Scheme

The training set is taken as a whole in the calculation of IDF, which ignores the distribution information of feature terms among categories. Thus, an improved term-weighting scheme is applied in our work.

Assuming that X = {x1:p1, x2:p2,…,xn:pn} is the information probability space, the information entropy of X is formulated as [8]:

$$H(X) = H(p_1, p_2, \ldots, p_n) = -\sum_{i=1}^{n} p_i \log_2 p_i \tag{1}$$

Information gain is the difference between information entropy, represented as:

$$I(X, y) = H(X) - H(X|y) \tag{2}$$

where H(X) is the entropy without information of y, H(X|y) is the conditional entropy, representing the incertitude degree of X with the information of y obtained.

From the aspect of information theory, the essential part of the improved term weighting scheme is: for a training set with certain probability distribution, word categorical information mostly depends on information gain. With this consideration, the improved term-weighting is define as:

$$Q_{ij} = TF(t_j) \times \log\left(\frac{k}{n_i} + 0.01\right) \times IG \tag{3}$$

Here Q_{ij} is the weight of t_j in c_i, n_i is the total number of words in c_i.

3.2 Word Similarity Calculation

With the improved term-weighting scheme, we can construct the word similarity matrix. For two terms t_i and t_j, in frequent term sets, the semantic similarity between them is derived according to Jaccard similarity [9] as follows:

$$CoR\left(t_i, t_j = \frac{1}{|\bar{C}|} \times \sum_{x \in |\bar{C}|} \frac{Q_{xi} Q_{xj}}{Q_{xi} + Q_{xj} - Q_{xi} Q_{xj}}\right) \tag{7}$$

Since the frequent term sets extraction is based on categories, x here stands for a category and Q_{xi} is the improved term weight of t_i in c_x. $\bar{C}$ is a subset of collection C, satisfying $\bar{C} = \{x|(Q_{xi} \neq 0) \cup (Q_{xj} \neq 0)\}$. And when $\bar{C}$ is empty, $CoR(t_i, t_j) = 0$.

We then normalize the semantic similarity as:

$$IaR(t_i, t_j) = \begin{cases} 1 & i = j \\ \frac{CoR(t_i, t_j)}{\sum_{i=1, i \neq j}^{N} CoR(t_i, t_j)} & i \neq j \end{cases} \tag{8}$$

However, in the whole term set there may be the case that co-occurring relations and identical class orientations might exist between one term and other terms. For example,{(computer, mouse), (computer, keyboard), (mouse, keyboard), (cellphone, computer), (cellphone, Internet)} is a frequent term set, in which 'keyboard' and 'mouse' are not only co-occurred but also linked by 'computer'. Similarly, we can also relate 'computer' to 'Internet' by 'cellphone', though they are not co-occurred directly. Thus, we define another relation to strengthen semantic association of these words.

Definition 4(inter-relation). Terms t_i and t_j are defined to be inter-related, if there exists at least one term t_k, which is the linking term inter-related with both t_i and t_j.

As is shown in Fig. 2, t_i and t_k are co-occurred as well as t_j and t_k. Therefore, we believe that there is semantic relation between t_i and t_j, though they are not co-occurred directly.

All term pairs with external relations are extracted and added into R, the extended frequent couple term sets R_e is formed. Moreover, words in R_e are strongly related with each other, which provides solid foundation to word similarity matrix construction.

It is necessary to take measures to quantify for external relations before word similarity matrix construction.

$$\mathrm{R\_IeR}(t_i, t_j|t_k) = \min\left(\mathrm{IaR}(t_i, t_k), \mathrm{IaR}(t_j, t_k)\right) \tag{9}$$

where $\mathrm{IaR}(t_i, t_k)$ and $\mathrm{IaR}(t_j, t_k)$ represents the semantic similarity between t_i and t_k, t_j and t_k respectively. In the quantification for external relations, we assume that the semantic similarity between t_i and t_j is at least valued as the minimization in all semantic similarities, which is feasible in fact.

The final external relation between t_i and t_j is calculated with all the linking terms of t_i and t_j. After normalization, the inter-relation is formalized as:

$$\mathrm{IeR}(t_i, t_j) = \begin{cases} 0 & i = j \\ \frac{1}{|L|}\sum_{\forall t_k \in L} \mathrm{R\_IeR}(t_i, t_j|t_k) & i \neq j \end{cases} \tag{10}$$

Here $L = \{t_k|((\mathrm{IaR}(t_i, t_j) > 0) \cap (\mathrm{IaR}(t_k, t_j) > 0))\}$, $|L|$ is the total number of terms. If the set of L is empty, the inter-relation between t_i and t_j, formulated as $\mathrm{IeR}(t_i, t_j)$ is zero. And if t_i and t_j indicates the same word, We regard $\mathrm{IeR}(t_i, t_j)$ as zero, too. Besides, when t_i is different from t_j, there may be one or more linking terms

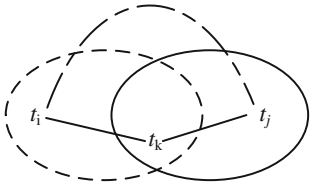

Fig. 2. External relation

relating them together, which are the elements in set L, taking the influence of all linking terms into consideration.

Word similarity matrix S is constructed based on frequent term sets, in which Sij denotes the semantic similarity and is defined as follows:

$$S_{ij} = \begin{cases} 1 & i = j \\ (1-\gamma)\cdot \mathrm{IaR}\left(t_i, t_j\right) + \gamma \cdot \mathrm{IeR}\left(t_i, t_j\right) & i \neq j \end{cases} \tag{11}$$

Where $\gamma \in [0, 1]$ is an important parameter deciding the weight of inter-relations. In our work, we set γ as 0.5.

At this point, word similarity matrix S is constructed successfully, where semantic similarity of word represents not only co-occurring relation but also extended inter-relation, therefore, semantic associations are further enhanced.

4 Short Text Feature Extension Based on Semantic Similarity Matrix

The non-negative factorization method was first proposed by Lee in Nature in 1999 [10]. Different from original non-negative factorization algorithm, the symmetric non-negative factorization (SNMF) is pretty special, whose duty is to factor a non-negative matrix into a product of a non-negative matrix and its transposed matrix. More specifically, for a given non-negative matrix $Z_{n\times n}$, a non-negative matrix factor $Y_{n\times k}$, satisfying:

$$Z \approx YY^{T}, Y \geq 0 \tag{12}$$

Since the semantic similarity matrix S is obviously symmetric, SNMF aims to factor S into P and the transposed matrix of P. Each element in P is calculated iteratively as follows:

$$P_{i,j} \leftarrow \frac{1}{2}\left[P_{i,j}\left(1 + \frac{(SP)_{ij}}{(PP^{T}P)_{ij}}\right)\right] \tag{13}$$

We build the original feature space W with TF-IDF and factor S into $S_{K\times K} = P_{K\times N} \times P^{T}_{N\times K}$. Then, the matrix W_e extended is obtained.

$$W_e = WP^{T} \tag{14}$$

As the transposed matrix P^T is factored by S, each one in P^T is certainly not equals to 0. Meanwhile, W is the original matrix where each row represents a document, it is impossible to be 0 in rows, too. Therefore, the new extended feature space W_e is no more sparse than before, which is vital to the construction of short text model. Furthermore, word similarity and categorical information as well as word semantic

information W are assimilated into the new feature space W_e, which is favor of the similarity calculation of short text.

5 Experiments

In this section, we conduct a series of experiments to evaluate the performance of our algorithm and analyze these experiments and the results.

5.1 Data Set

Experiments are conducted on two datasets: 20-Newsgroups [11] and Sougou corpus [12]. 20-Newsgroups is composed of 20 different news groups and 20000 short text snippets. Sougou corpus is a data set of news pages from Sohu news provided by Sougou lab, including 18 categories such as International, Sports, Society, Entertainment, etc. Each page has its page URL, page ID, page title and body content.

Short text refers to the title, the news contents and the description of short text are used for background knowledge extraction. All Chinese documents were pre-processed by word segmentation using ICTCLAS. After pre-processing, we select 10 categories from 20-Newsgroups and each category contains 200 documents. We also choose 9 categories from Sougou corpus, 2000 pages in total. The traditional K-means algorithm is employed to verify our experiment performance.

5.2 Experiment Results

As α and β are the most important parameters in our algorithm, we first vary their values to testify the performance of double term set extraction. Then, the performance of the original frequent term set and improved frequent term set with external relations are compared. Finally, we make a comparison of five short text representation method for clustering using Purity and F-measure as evaluation criteria.

The extraction of double term sets is significant to word similarity matrix construction, which greatly relies on the number of support and confidence restraints. The support guarantees the co-occurring relation of terms while the confidence determines whether the terms have identical class orientation. We extract the couple term sets using different parameter settings: α = 1.0 %, 1.5 %, 2.0 %, 2.5 %, 3.0 % and β = 0.5, 0.6, 0.7, 0.8, 0.9 respectively, and the results are listed in Table 3.

As is shown in Table 3, the number of extracted double term sets decreases dramatically with the increase of support and confidence. It is understandable that higher support means more co-occurring relations and the constrains will be more strict when confidence increases. This shows that when support and confidence are set to be high, the information for background knowledge is too rare to have essential influence on the original feature space. Therefore, we choose α = 1.0 %, β = 0.5 to construct our background knowledge in the following experiment.

What's more, we define external relation based on traditional frequent term sets to further extend term sets, obtaining more semantic information. With the parameter of

Table 3(a). Double frequent term sets distribution with different support and confidence

α/%	β				
	0.5	0.6	0.7	0.8	0.9
1.0	9117	7648	5315	2754	1231
1.5	5383	2114	973	518	374
2.0	2854	811	574	430	292
2.5	659	443	207	161	144
3.0	530	336	132	25	9

Table 3(b). Double frequent term sets distribution with external relation

α/%	β				
	0.5	0.6	0.7	0.8	0.9
1.0	10942	9301	6378	2904	1377
1.5	6459	2737	1167	569	392
2.0	3467	962	638	473	306
2.5	790	511	228	177	151
3.0	609	389	145	25	9

support fixed and external relations taken into account, the number of extracted frequent term sets is demonstrated in Table 3b.

From Table 3b, frequent term sets with external relation and original algorithm are different in quantity, though the distributions of them are the same. It is obvious that the number of extracted term sets with external relation is much more than that of original algorithm. As we can observe, when $\alpha = 1.0$ %, $\beta = 0.5$, the original frequent term sets is 9117 while the number of term sets with external relation is 10942 which increases about 20 %. The growth ratio decreases with the increase of parameter. Besides, the double frequent term sets with external relation is in favor of more semantic information.

Finally, we conduct experiments to compare clustering performances of five methods on two different data sets. The experiment results on two evaluation index — Purity and F-measure are presented in Fig. 3.

In Fig. 3, it is clear that different performance potentially depends on different data set. And the proposed method performs much better than any other methods. Furthermore, the result of the Sougou corpus is a little superior than that of 20-newsgroups in our experiment. As is depicted in Fig. 3, results of these five methods can be roughly divided into 3 levels. The traditional TF-IDF performs the worst, obviously blaming on ignoring semantics in model construction. The performances of coupled term-term relations method and the improved term-weighting method are similar, which are still worse than that of method based on frequent term sets. This phenomenon can be explained: the coupled term-term relations method extract inner and inter relation of word with co-occurring relation to enhance semantic information, while the improved

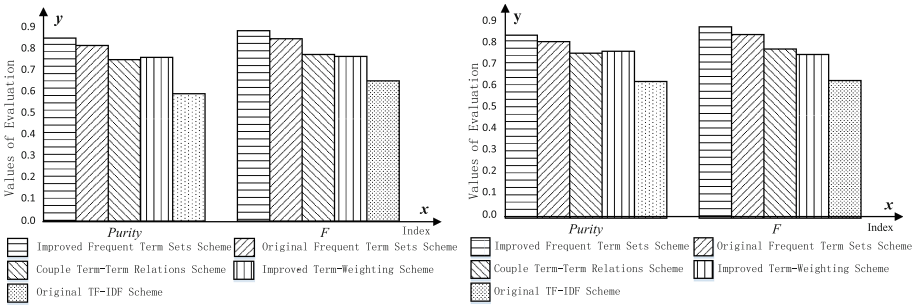

Fig. 3. Clustering results of different methods on 20-Newsgroups (left) and Sougou corpus (right)

term-weighting method considers information gain and statistical information. They are both unfortunately one-sided considered.

Extending the extracted semantic information to short text space, the original frequent term sets method to some extent releases the problem of high dimension sparsity in short texts. However, as is shown in the above Fig. 3, the best scheme for short text representation is the improved frequent term sets method. The superior of this method are summarized as follows: additional semantic information is first revealed with external relation. Then the word similarity matrix is built via improved term-weighting scheme and word categorical information. Finally, the symmetric non-negative matrix factorization technique is used to extend the feature space, which alleviates the problem of high-dimensional sparsity.

6 Conclusion

This paper discusses a short text feature extension algorithm based on improved frequent word sets. The external relation is proposed based on frequent term sets, further enhancing associations of words. Considering the distribution information of categories, an improved term-weighting scheme using information gain is presented, which efficiently remained the categorical information. What is more, all term pairs with external relations are extracted and the frequent term set is expanded. Finally, the word similarity matrix is constructed via the frequent term set, and the symmetric non-negative matrix factorization technique is used to extend the feature space. Experiments show that the constructed short text model can significantly improve the performance of clustering and effectiveness.

Acknowledgement. This work is supported by the National Natural Science Foundation of China (No.61363058), Youth Science and technology support program of Gansu Province (145RJZA232, 145RJYA259, 1606RJYA269), 2016 Provincial College Students Innovation and entrepreneurship training program (No.201610736040, 201610736041) and 2016 annual public record open space Fund Project (No.1505JTCA007).

References

1. Alexander, P., Patrick, P.: Twitter as a corpus for sentiment analysis and opinion mining. In: Proceeding of the Seventh International Conference on Language Resources and Evaluation, Valletta, Malta, pp. 19–21 (2010)
2. Zhang, W., Yoshida, T., Tang, X.: Text classification based on multi-word with support vector machine. Knowl.-Based Syst. **21**(8), 879–886 (2008)
3. Sun, A.: Short text classification using very few words. In: Proceedings of the 35th international ACM SIGIR conference on Research and Development in Information Retrieval, New York, pp. 1145–1146 (2012)
4. Kalogeratos, A., Likas, A.: Text document clustering using global term context vectors. Knowl. Inf. Syst. **31**(3), 455–474 (2012)
5. Cheng, X., Miao, D.Q., Wang, C., et al.: Coupled term-term relation analysis for document clustering. In: Proceedings of Neural Networks International Joint Conference on Artificial Intelligence (IJCNN), Dallas, pp. 1–8. IEEE (2013)
6. Liu, W., Quan, X., Feng, M., et al.: A short text modeling method combining semantic and statistical information. Inf. Sci. **180**(20), 4031–4041 (2010)
7. Yuan, M.: Feature extension for short text categorization using frequent term sets. Procedia Computer Science **31**, 663–670 (2014)
8. Qinghua, H.U., Guo, M.Z., DaRen, Y.U.: Information entropy for ordinal classification. Science China **53**(6), 1188–1200 (2010)
9. Bollegala, D., Matsuo, Y.: Measuring semantic similarity between words using web search engines. In: Proceedings of the Workshop on Social and Collaborative Construction of Structured Knowledge at the 16th International World Wide Conference (WWW 2007), pp. 757–786. ACM, New York (2007)
10. Lee, D.D., Seung, H.S.: Learning the parts of objects by non-negative matrix factorization. Nature **401**(6755), 788–791 (1999)
11. Lang, K.: Newsweeder. Learning to filter netnews. In: Proceedings of the Twelfth International Conference on Machine Learning, pp. 331–339. Morgan Kaufmann, Tahoe City (1995)
12. Sougou lib. http://www.Sogou.com/labs/dl/c.html. Accessed 30 Apr 2012