Zhang Y, Liu M, Xia HX. Clustering context-dependent opinion target words in Chinese product reviews. JOURNAL OF COMPUTER SCIENCE AND TECHNOLOGY 30(5): 1109–1119 Sept. 2015. DOI 10.1007/s11390-015-1586-2

Clustering Context-Dependent Opinion Target Words in Chinese Product Reviews

Yu Zhang* (张 宇), *Member, CCF*, Miao Liu (刘 妙), and Hai-Xia Xia (夏海霞)

School of Information Science and Technology, Zhejiang Sci-Tech University, Hangzhou 310018, China

E-mail: yzh@zstu.edu.cn; miaolium@163.com; xiahx@zstu.edu.cn

Received November 15, 2014; revised June 8, 2015.

Abstract In opinion mining of product reviews, an important task is to provide a summary of customers' opinions based on different opinion targets. Due to various knowledge backgrounds or linguistic habits, customers use a variety of terms to describe the same opinion target. These terms are called as context-dependent synonyms. In order to provide a comprehensive summary, the first step is to classify these opinion target words into groups. In this article, we mainly focus on clustering context-dependent opinion target words in Chinese product reviews. We utilize three clustering methods based on distributional similarity and use four different co-occurrence matrices for experiments. According to the experimental results on a large number of reviews, we find that our proposed heuristic k-means clustering method using opinion target words co-occurrence matrix achieves the best clustering result with lower time complexity and less memory space. In addition, the accuracy is more stable when choosing different combinations of centroids. For some kinds of co-occurrence matrices, we also find that using small-size (low-dimensional) matrices achieves higher average clustering accuracy than using large-size (high-dimensional) matrices. Our findings provide a time-efficient and space-efficient way to cluster opinion targets with high accuracy.

Keywords clustering, context-dependent, opinion target word, product review, opinion mining

1 Introduction

With the rapid development of e-commerce in recent years, product reviews have become increasingly important. They are customer-driven responses to convey personal experiences and subsequent product use. This constitutes a new and measurable source for business intelligence. Recently, the number of product reviews has grown very quickly. We take China's largest C2C (Customer to Customer) e-commerce website — Taobao.com as an example. It is quite common that each best-selling product has more than one hundred thousand reviews. Therefore, it is very time-consuming, and sometimes even impossible, for customers to read all the reviews. Meanwhile, it is also very hard for manufacturers or e-commerce platforms to fully understand customers' needs. In this situation, opinion mining turns out to be a quite effective method to solve the problem[1]. In recent years, there has been extensive literature on opinion mining[2-6]. Through mining and summarizing massive amounts of explicit or implicit information from product reviews, we manage to provide support for a variety of applications, such as product comparison, purchase decision, marketing strategy, and product promotion. How to cluster opinion targets turns out to be a very important step of opinion mining. Therefore, in this article, we mainly focus on clustering context-dependent opinion targets in Chinese product reviews.

Clustering is a common technique for statistical data analysis, which is to classify similar objects into different groups, or to partition an object into subsets.

Regular Paper

Special Section on Social Media Processing

This work was supported by the Commonweal Technical Project of Zhejiang Province of China under Grant No. 2013C33063, the National Natural Science Foundation of China under Grant Nos. 61100183, 61402417, the Natural Science Foundation of Zhejiang Province of China under Grant No. LQ13F020014, and the 521 Talents Project of Zhejiang Sci-Tech University.

*Corresponding Author

Clustering can provide unique insights into the behaviors of customers and also make the organization of business more efficient[7]. To begin, let us look at the following definitions.

Opinion Target. An opinion target is the object being commented on in an online review by customers.

It may be an entity, such as cell phone or service. It may also be the components, attributes or functions of an entity, such as screen (component), price (attribute), and photographing (function).

Usually, customer reviews on Taobao.com focus on the following three aspects:

• *product*: including quality, attributes, function, etc.;

• *seller*: including service, attitude, reputation, etc.;

• *logistics*: including speed, package, delivery, etc.

We assume that all opinion targets fit into one of the above three categories. In Taobao reviews, a "cell phone" is usually called a "机器" (machine) in Chinese. Meanwhile, the terms "机子" and "机" are also employed by Chinese customers to denote a cell phone. All three terms have the same meaning — machine in English. As can be seen, the first Chinese character of three terms is the same. They can be regarded as *general synonyms*.

At the same time, customers also use other terms to denote cell phone in their corresponding reviews, such as "treasure" (宝贝), "stuff" (东东), and "goods" (货). For example, the literal translation for "淘宝" (Taobao) is "hunt for treasure", and therefore products sold on Taobao.com are habitually called "treasure" by customers. It is quite difficult to cluster these terms because they are not synonyms at all according to lexical semantics. These kinds of terms can be regarded as synonyms only if they are used in a certain scenario. Therefore, we denote them as *context-dependent synonyms*.

Existing research has paid little attention to the context-dependent characteristic of the objects to be clustered. Due to the complexity of Chinese language, this issue becomes more prominent. In this article, we conduct a thorough study and propose an effective solution to the problem.

The organization of this article is as follows. Section 2 presents the related work. Section 3 introduces our proposed method of clustering opinion targets. Section 4 presents the detailed experimental results using real product reviews from Taobao.com. Section 5 concludes this article and outlines future work.

2 Related Work

There have been some researches on text document clustering. Most researches focus on clustering product features. These researches can be divided into two main categories according to similarity measures[8]: the first kind of methods is based on pre-existing knowledge, while the second kind of methods relies on the distributional similarity of words in the corpus.

Knowledge-based methods calculate lexical similarity according to existing knowledge and term taxonomy, such as Thesaurus①, WordNet②, and HowNet③. Based on user-specific prior knowledge, Carenini *et al.*[9] mapped learned features into a user-defined taxonomy to obtain their corresponding term similarity.

Wagstaff *et al.*[10] incorporated domain knowledge in the form of instance-level constraints into a k-means clustering algorithm. Their experimental results show that there are impressive gains in accuracy; however, the assignment of instances to clusters sometimes is order-sensitive, thus the algorithm should be able to backtrack.

Zhai *et al.*[11] proposed a semi-supervised learning method with two soft constraints based on sharing of words and lexical similarity to cluster product features. A must-link constraint specifies that two data instances must be in the same cluster, while a cannot-link constraint specifies that two data instances cannot be in the same cluster.

Knowledge-based lexical similarity is widely used in the NLP (natural language processing) area to measure the similarity of two phrases[11]. However, there are several weaknesses of knowledge-based similarity measure:

1) Many neologisms or colloquial phrases that are widely used in reviews cannot be found in the knowledge base, especially the terms creatively used by Taobao customers. If we cannot identify a term in the knowledge base, we are not able to determine its similarity to others.

①http://www.thesaurus.com/, May 2015.

②http://wordnet.princeton.edu/, May 2015.

③http://keenage.com/, May 2015.

2) Even if the terms can be identified in the knowledge base, the meaning of some phrases may have already changed. As we mentioned in Section 1, "宝贝 (treasure)" does not always mean valuable things. On Taobao.com, it usually refers to an ordinary product sold online, such as a cell phone. It is impossible to tell that "treasure" and "cell phone" are synonyms according to WordNet-like knowledge base.

3) Term similarity is context-dependent; therefore knowledge-based dictionaries such as WordNet or Thesaurus are not able to suit different categories.

The second kind of methods relies on distributional similarity of words. Distributional similarity is based on the assumption that the meaning of words is related to their contexts, namely, words with similar meaning tend to appear in similar contexts[12]. Therefore, term similarities can be judged by their context.

In literature, there are many definitions for contexts, such as correlation between terms and documents[13], web search results for short texts[14-15], pointwise mutual information[12]. The information about each word and its corresponding context is stored in a matrix. In the end, the similarity of two words is calculated by measuring the similarity between two context vectors in the matrix[16].

Recent research has also applied topic modeling to solve the clustering problem. Andrzejewski *et al.*[17] proposed an LDA (Latent Dirichlet Allocation) framework, which incorporates domain knowledge in the form of must-link and cannot-link constraints. Zhai *et al.*[4] combined topic modeling method LDA with some pre-existing knowledge in the form of automatically extracted constraints to group product features. Experimental results show that the proposed constrained-LDA outperforms the typical LDA and mLSA (multilevel latent semantic association) by a large margin. Zhao *et al.*[18] proposed a topical document clustering method which exploits linguistic features of the document.

In named entity clustering field, there is also prior work on similar tasks. Bu *et al.*[15] proposed a knowledge-free, training-free and language-independent multiword expression distance to recognize named entities and terminologies, which measures the distance from an n-gram to its semantics. Elsner *et al.*[19] built a fully unsupervised generative model which makes use of entity feature, syntactic context, and coreference information for entity clustering. Andrews *et al.*[20] proposed a model for cross-document co-reference resolution by learning similarity from unlabeled data. Green *et al.*[21] developed new methods to cluster text mentions across documents and languages based on cross-lingual similarity and context similarity.

Compared with previous research, the contribution of this article is multi-fold. First, in this article, we not only focus on product feature clustering that has been widely studied before, but also pay attention to other aspects of clustering opinion targets such as seller and logistics. Second, we employ several clustering methods using various contexts and conduct a thorough study on distributional similarity measure. Third, we propose a simple and practical method for coarse-grained opinion targets clustering which is time-efficient and space-efficient. The method achieves high accuracy and can be applied in large-scale applications.

3 Methodology

As discussed in Section 2, knowledge-based clustering methods are not sufficient to solve the problem of context-dependent opinion targets clustering. Therefore, in this article, we make use of distributional similarity of words for clustering. The knowledge that we use is the co-occurrence of words in a review sentence. First, we briefly illustrate the settings of our proposed method.

We let $P = \{p_1, p_2, \cdots, p_{n_\mathrm{p}}\}$ be a group of products sold online, where n_p denotes the number of the products. Each product $p_i (1 \leqslant i \leqslant n_\mathrm{p})$ has a set of reviews $R_i = \{r_1, r_2, \cdots, r_{n_\mathrm{r}}\}$, where n_r denotes the number of reviews written by customers for product p_i. Before we illustrate our method, we first present several definitions.

Opinion Target Word. An opinion target word in a review r_j $(1 \leqslant j \leqslant n_\mathrm{r})$ is a word or compound word that refers to a specific target that has been commented on.

For example, "cell phone" is the opinion target word in the following review sentence: "cell phone great".

Note that we have translated the Chinese review to English literally. The translation may not be grammatically correct in English④. All the example reviews in this article follow this principle.

We let $B = \{b_1, b_2, \cdots, b_{n_\mathrm{b}}\}$ be the opinion target words set, where n_b denotes the number of words in B.

④The correct English translation should be as follows: the cell phone is great. However, there are neither definite articles nor "be" verbs in Chinese. Therefore, to illustrate how our proposed method works for Chinese reviews, we just translate it literally with neither definite articles nor "be" verbs.

Opinion Word. An opinion word is a word that expresses a positive, negative or neutral opinion on an opinion target b_i $(1 \leqslant i \leqslant n_{\mathrm{b}})$.

For example, "like" is an opinion word which shows a positive attitude towards the product in the following review sentence: "I like this cell phone."

We let $O = \{o_1, o_2, \cdots, o_{n_{\mathrm{o}}}\}$ be the opinion words set, where n_{o} denotes the number of opinion words in O.

Content Word. A content word is a word that refers to some object, action, or characteristic.

In contrast to content words, function words are used to depict the grammatical relationships between other words in a sentence. In Chinese, content words include nouns, verbs, adjectives, adverbs, idioms, numerals, quantifiers, and pronouns. Take the following review as an example: "I think cell phone beautiful and practical."

In the above sentence, "I" (pronoun), "think" (verb), "cell phone" (noun), "beautiful" (adjective), and "practical" (adjective) are content words while "and" is a function word.

We let $T = \{t_1, t_2, \cdots, t_{n_{\mathrm{t}}}\}$ be the content words set, where n_{t} denotes the number of content words in T.

Context Word. For a given opinion target word b_i $(1 \leqslant i \leqslant n_{\mathrm{b}})$, a word that meets the following two requirements can be regarded as the context word of b_i: 1) the word belongs to one of the ten lexical categories: nouns, verbs, adjectives, idioms, distinguishing words, conjunctions, pronoun, adverbs, numerals, and classifier; 2) the context word is immediately adjacent to b_i. We first filter out the words that do not belong to the above mentioned ten categories, and then we determine the context word of b_i according to adjacency. The context word can be in front of b_i or behind it.

Take the previous mentioned review as an example: "I think cell phone very good."

For opinion target word "cell phone", "think" (verb) is the context word in front of it while "very" (adverb) is the context word behind it.

We let $X = \{x_1, x_2, \cdots, x_{n_{\mathrm{x}}}\}$ be the context words set, where n_{x} denotes the number of context words in X.

3.1 Different Kinds of Co-Occurrence Matrices

3.1.1 Definition of Co-Occurrence Matrices

Usually, a product review contains multiple clauses that are separated by punctuations. Therefore, we consider the words co-occur if they appear in the same clause. In this article, we consider four kinds of co-occurrence matrices:

1) *Opinion target words co-occurrence matrix*: constructed by aggregating the co-occurrence frequency of an opinion target word b_i and another opinion target word b_j in one clause, denoted as $\boldsymbol{M}^B$. $\boldsymbol{M}^B$ is an $n_{\mathrm{b}} \times n_{\mathrm{b}}$ symmetric matrix and both its rows and columns represent the opinion target words, where $b_i \in B, b_j \in B, 1 \leqslant i \leqslant n_{\mathrm{b}}, 1 \leqslant j \leqslant n_{\mathrm{b}}$.

2) *Opinion target words and opinion words co-occurrence matrix*: constructed by aggregating the co-occurrence frequency of an opinion target word b_i and its corresponding modifier — an opinion word o_j — in one clause, denoted as $\boldsymbol{M}^O$. $\boldsymbol{M}^O$ is an $n_{\mathrm{b}} \times n_{\mathrm{o}}$ matrix. Its rows and columns represent the opinion target words and the opinion words respectively, where $b_i \in B, o_j \in O, 1 \leqslant i \leqslant n_{\mathrm{b}}, 1 \leqslant j \leqslant n_{\mathrm{o}}$.

3) *Opinion target words and content words co-occurrence matrix*: constructed by aggregating the co-occurrence frequency of an opinion target word b_i and a content word t_j in one clause, denoted as $\boldsymbol{M}^T$. $\boldsymbol{M}^T$ is an $n_{\mathrm{b}} \times n_{\mathrm{t}}$ matrix. Its rows and columns represent the opinion target words and the content words respectively, where $b_i \in B, t_j \in T, 1 \leqslant i \leqslant n_{\mathrm{b}}, 1 \leqslant j \leqslant n_{\mathrm{t}}$.

4) *Opinion target words and context words co-occurrence matrix*: constructed by aggregating the co-occurrence frequency of an opinion target word b_i and a context word x_j in one clause, denoted as $\boldsymbol{M}^X$. $\boldsymbol{M}^X$ is an $n_{\mathrm{b}} \times n_{\mathrm{x}}$ matrix. Its rows and columns represent the opinion target words and the context words respectively, where $b_i \in B, x_j \in X, 1 \leqslant i \leqslant n_{\mathrm{b}}, 1 \leqslant j \leqslant n_{\mathrm{x}}$.

3.1.2 Example for Co-Occurrence Matrices

Now we offer an example to explain how the four types of matrices are obtained. Given the following review:

Chinese: "爸爸说屏幕和按键都很大",

English: "Father says screen and keypad both very big".

After words segmentation and part-of-speech tagging, the review sentence turns out to be: Father/n says/v screen/n and/c keypad/n both/d very/d big/a.

In the above, /n denotes noun, /v denotes verb, /c denotes conjunction, /d denotes adverb, and /a denotes adjective.

Here, we define a rule: if there is a conjunction, such as "and", which connects two opinion target words, then we determine that the corresponding opinion word

modifies both opinion target words. Therefore, we can obtain the co-occurrence matrices as follows.

• In $\boldsymbol{M}^B$, two opinion target words in the review will be considered: "screen" and "keypad". The corresponding programming statements are as follows:

$$m^B_{(\text{screen, keypad})} = m^B_{(\text{screen, keypad})} + 1,$$
$$m^B_{(\text{keypad, screen})} = m^B_{(\text{keypad, screen})} + 1.$$

• In $\boldsymbol{M}^O$, two opinion target words (screen and keypad) and one opinion word (big) in the review will be considered. The corresponding programming statements are as follows:

$$m^O_{(\text{screen, big})} = m^O_{(\text{screen, big})} + 1,$$
$$m^O_{(\text{keypad, big})} = m^O_{(\text{keypad, big})} + 1.$$

• In $\boldsymbol{M}^T$, two opinion target words (screen and keypad) and five content words (father, say, screen, keypad, big) will be considered. Therefore, for opinion target word "screen", the corresponding programming statements are as follows:

$$m^T_{(\text{screen, father})} = m^T_{(\text{screen, father})} + 1,$$
$$m^T_{(\text{screen, say})} = m^T_{(\text{screen, say})} + 1,$$
$$m^T_{(\text{screen, keypad})} = m^T_{(\text{screen, keypad})} + 1,$$
$$m^T_{(\text{screen, both})} = m^T_{(\text{screen, both})} + 1,$$
$$m^T_{(\text{screen, very})} = m^T_{(\text{screen, very})} + 1,$$
$$m^T_{(\text{screen, big})} = m^T_{(\text{screen, big})} + 1.$$

For opinion target word "keypad", the corresponding programming statements are as follows:

$$m^T_{(\text{keypad, father})} = m^T_{(\text{keypad, father})} + 1,$$
$$m^T_{(\text{keypad, say})} = m^T_{(\text{keypad, say})} + 1,$$
$$m^T_{(\text{keypad, screen})} = m^T_{(\text{keypad, screen})} + 1,$$
$$m^T_{(\text{keypad, both})} = m^T_{(\text{keypad, both})} + 1,$$
$$m^T_{(\text{keypad, very})} = m^T_{(\text{keypad, very})} + 1,$$
$$m^T_{(\text{keypad, big})} = m^T_{(\text{keypad, big})} + 1.$$

• In $\boldsymbol{M}^X$, two opinion target words (screen and keypad) will be considered. For opinion target word "screen", its context words are "say" and "and". The corresponding programming statements are as follows:

$$m^X_{(\text{screen, say})} = m^X_{(\text{screen, say})} + 1,$$
$$m^X_{(\text{screen, and})} = m^X_{(\text{screen, and})} + 1.$$

For opinion target word "keypad", its context words are "and" and "both". Therefore, we have the following formulas:

$$m^X_{(\text{keypad, and})} = m^X_{(\text{keypad, and})} + 1,$$
$$m^X_{(\text{keypad, both})} = m^X_{(\text{keypad, both})} + 1.$$

3.2 Similarity Measure

Clustering is the task of grouping a set of objects in such a way that objects in the same cluster are more similar to each other than to those in other clusters. That is, the clusters have high intra-cluster similarity and low inter-cluster similarity.

In this article, we exploit distributional similarity for clustering. Distributional similarity assumes that words with similar meaning tend to appear in similar scenarios[11]. If two opinion target words are similar, then their co-occurrent words in clause tend to be the same. The more similar the two groups of co-occurrent words, the more similar the two opinion targets. The two groups of words correspond to the two vectors in the co-occurrence matrix.

We make use of Cosine distance formula to measure the similarity of two vectors, which is the similarity of two opinion target words b_y and b_z:

$$\begin{aligned} Sim(b_y, b_z) &= \frac{\boldsymbol{V}_{b_y} \cdot \boldsymbol{V}_{b_z}}{\|\boldsymbol{V}_{b_y}\| \|\boldsymbol{V}_{b_z}\|} \\ &= \frac{\sum_{i=1}^{n} l_{yi} \times l_{zi}}{\sqrt{\sum_{i=1}^{n} (l_{yi})^2} \times \sqrt{\sum_{i=1}^{n} (l_{zi})^2}}, \end{aligned}$$

where, $\boldsymbol{V}_{b_y}$ and $\boldsymbol{V}_{b_z}$ are co-occurrence vectors of opinion target words b_y and b_z respectively. l_{yi} denotes the i-th element of vector $\boldsymbol{V}_{b_y}$, while l_{zi} denotes the i-th element of vector $\boldsymbol{V}_{b_z}$. There are also several other distance metrics for clustering, such as Euclidean distance. A good distance metric can be obtained using some learning algorithms[22].

3.3 Clustering Algorithms

We utilize three different algorithms for clustering opinion target words: standard k-means clustering, heuristic k-means clustering, and hierarchical clustering. We will illustrate these algorithms one by one.

3.3.1 Standard k-Means Clustering

In fact, standard k-means clustering is the so-called k-means clustering. We add "standard" in the front

to distinguish it from heuristic k-means clustering. k-means clustering aims to partition n observations into k sets in which each observation belongs to the cluster with the nearest mean. As mentioned in Section 1, opinion targets can be classified into three main categories: product, seller, and logistics. Therefore, we set $k = 3$ in this article.

The standard k-means clustering algorithm consists of the following five steps:

Step 1: initialization: set $k = 3$ and randomly generate three cluster centroids from B.

Step 2: calculate pairwise similarities between each remaining opinion target word and respective centroids using Cosine similarity measure.

Step 3: assign each opinion target word to the cluster that has the highest similarity.

Step 4: calculate the new means to be the centroids of the opinion target words in the new clusters.

Step 5: steps 2∼4 are repeated until convergence is reached. The algorithm terminates when the assignments no longer change.

3.3.2 Heuristic k-Means Clustering

Heuristic k-means clustering is almost the same as the standard k-means clustering except for bringing in some prior knowledge to guide clustering algorithm to work out meaningful clusters for humans.

As mentioned before, the opinion target words can be classified into three main categories. From each category, we select one typical word as the initial centroid. Then we make use of the chosen centroids for clustering. We expect this algorithm achieves better performance by integrating human knowledge.

3.3.3 Hierarchical Clustering Algorithm

In this article, we exploit a bottom-up clustering method — the agglomerative hierarchical clustering.

The hierarchical clustering algorithm includes the following five steps.

Step 1: calculate the pairwise distances between opinion target words using Cosine similarity measure and construct a distance matrix.

Step 2: each opinion target word is regarded as a cluster of its own. The number of clusters is denoted as n_{h}.

Step 3: the nearest two clusters are combined into a higher-level cluster sequentially.

Step 4: calculate the pairwise distance between the new cluster and the remaining clusters, and then remove the just combined cluster and update the distance matrix.

Step 5: repeat step 3 and step 4 until n_{h} reaches the predefined value or the pairwise distance is smaller than the threshold θ_{d}.

4 Experimental Study

In this section, we first introduce the experimental setup. Then we present the following experimental results: 1) comparisons of different clustering methods using different co-occurrence matrices; 2) the sensitivity of selected centroids; 3) the influence of matrix size on clustering accuracy.

4.1 Experimental Setup

In this article, the corpus we use was crawled from Taobao.com — China's largest C2C e-commerce website. There are 106 950 reviews in the corpus S and all these reviews are from the "cell phone" category.

We first preprocess the data in S:

• *Fraud Reviews Deletion.* Fraud reviews include advertisements, reputation manipulation review, false transaction review, etc.

• *Long Sentences Filter.* A few customers write very long reviews without any punctuation, which will greatly affect the accuracy for constructing co-occurrence matrix. Therefore, we delete those reviews when at least one of its clauses contains more than 120 bytes without any punctuation.

• *Chinese Word Segmentation and Part-of-Speech Tagging.* We utilize ICTCLAS5.0[⑤] (Institute of Computing Technology, Chinese Lexical Analysis System), which was developed by Chinese Academy of Sciences for Chinese word segmentation and part-of-speech tagging.

• *Word Segmentation Error Correction.* Some of the word segmentation results obtained from ICTCLAS5.0 are wrong; therefore we wrote a program to detect and correct some universal errors. A few small probability errors are ignored since they will not have a big impact on clustering accuracy.

We assume that the opinion targets have already been correctly extracted from the text using some algorithms[23-24], and our task is only to classify them into the right clusters. In this article, we have 85 opinion targets to cluster, that is, $n_{\mathrm{b}} = 85$.

⑤http://ictclas.nlpir.org/, Aug. 2015.

Afterwards, we construct the four co-occurrence matrices $\boldsymbol{M}^B$, $\boldsymbol{M}^O$, $\boldsymbol{M}^T$, and $\boldsymbol{M}^X$ based on corpus S. The dimensions of the four matrices are as follows.

• $\boldsymbol{M}^B$: a matrix with 85 rows and 85 columns. Both the rows and the columns of $\boldsymbol{M}^B$ denote the opinion target words in B; therefore $\boldsymbol{M}^B$ is a symmetric matrix.

• $\boldsymbol{M}^O$: a matrix with 85 rows and 1 156 columns. The rows of $\boldsymbol{M}^O$ denote the opinion target words in B, while the columns denote the opinion words in O.

• $\boldsymbol{M}^T$: a matrix with 85 rows and 1 689 columns. The rows of $\boldsymbol{M}^T$ denote the opinion target words in B, while the columns denote the content words in T.

• $\boldsymbol{M}^X$: a matrix with 85 rows and 1 105 columns. The rows of $\boldsymbol{M}^X$ denote the opinion target words in B, while the columns denote the context words in X.

For matrix $\boldsymbol{M}^T$, the number of extracted content words from corpus S is more than 1 689. However, we do not use all of them because low frequency words will cause matrix sparsity, which leads to low clustering accuracy. Therefore, we select words whose word frequencies are greater than 50 to construct matrix $\boldsymbol{M}^T$. The reason why we choose 50 as a threshold is that the word frequency declines sharply under 50. When constructing matrix $\boldsymbol{M}^X$, we also filter out those low-frequency words, the same as constructing matrix $\boldsymbol{M}^T$. The threshold of selecting words for $\boldsymbol{M}^X$ is 20.

Using the above four matrices, we can calculate pairwise similarity of opinion target words using Cosine similarity measure as shown in Subsection 3.2. Then we run the three algorithms on four different matrices. The results and comparisons are shown below.

4.2 Experimental Results

We evaluate the clustering methods proposed in this article and make comparisons among them. For accuracy comparison, three postgraduate students assign the 85 opinion target words to three categories (product, seller, and logistics). We use majority voting when they disagree with each other. Note that the three students are quite familiar with online shopping and they are not members of our research team. We take their annotations as the gold standard for our evaluation.

4.2.1 Results of Standard k-Means Algorithm

In this experiment, we let $k = 3$. The standard k-means algorithm randomly selects three words from 85 opinion target words as centroids. Each group of the chosen words is regarded as a combination; therefore there are 98 770 combinations. We enumerate all the above combinations and run the standard k-means clustering algorithm based on the four matrices: $\boldsymbol{M}^B$, $\boldsymbol{M}^O$, $\boldsymbol{M}^T$, and $\boldsymbol{M}^X$. Table 1 shows the average accuracy of all combinations and the accuracy is obtained by making comparisons with gold standard.

Table 1. Average Accuracy of Standard k-Means Clustering

Matrix	Average Accuracy (%)
$\boldsymbol{M}^B$	85.32
$\boldsymbol{M}^O$	62.55
$\boldsymbol{M}^T$	72.53
$\boldsymbol{M}^X$	67.68

Table 1 shows the average accuracy of standard k-means clustering method. Matrix $\boldsymbol{M}^B$ achieves the best result among the four matrices, and the average accuracy is 85.32%. Using matrices $\boldsymbol{M}^T$ and $\boldsymbol{M}^X$, the average accuracies are 72.53% and 67.68% respectively. Using matrix $\boldsymbol{M}^O$ produces the lowest accuracy — 62.55%.

4.2.2 Results of Heuristic k-Means Clustering

The clustering accuracies vary greatly when the three centroids are randomly selected. Therefore, we propose the heuristic k-means clustering algorithm to introduce some knowledge by choosing three centroids manually.

Taking opinion target words co-occurrence matrix $\boldsymbol{M}^B$ for example, we choose "cell phone", "service" and "logistics" as the three centroids for clustering. (They are emphasized by underlines in the following result.) The experimental results are as follows:

• cell phone, thing, machine (机子), goods, battery, quality, machine (机器), screen, machine (机), price, function, performance price ratio, voice, system, software, accessory, appearance, earphone, package, signal, hand feeling, telephone, response, rear cover, communicate by telephone, pixel, screen protector film, character, make telephone call, standby, performance, charger, stuff, value, tone quality, keypad, battery charge, memory card, purity, operation, start up, shell, photographing, take pictures, ring, font, key, resolution, keyboard, style, electric plate, data line, sound volume, radio, cell phone shell, design, GPS (Global Position System), electric torch, screensaver, antenna, thermal dissipation;

• service, seller, attitude, boss, shopkeeper, people, **treasure**, support staff, shop owner, after-sales, shop

manager, cooperation, merchant, transaction, reputation;

• logistics, goods delivery, speed, express, S.F. express, **running**, goods shipment, EMS (Express Mail Service), postal delivery.

From the above results, we can see that the accuracy of this word selection combination is 97.65%. Except for two words ("treasure" and "running", shown in bold) that are wrongly assigned, all the other words are correctly classified to the right cluster. The "cell phone" cluster is 100% correct, while the assignments to "service" and "logistics" clusters are wrong. This is the highest accuracy that can be obtained by using $\boldsymbol{M}^B$ for heuristic k-means clustering when selecting one centroid from each category.

Using matrix $\boldsymbol{M}^B$, there are also other combinations that achieve the same accuracy (97.65%). The two wrongly assigned words may not be the same for different centriods combinations. We also notice that for different matrices, the optimal combinations of selected centroids may not be the same. In other words, the selected three centroids which manage to obtain the highest accuracy for matrix $\boldsymbol{M}^B$ may produce a very low accuracy using matrix $\boldsymbol{M}^T$.

The average accuracy of heuristic k-means clustering by using different matrices is shown in Table 2.

Table 2. Average Accuracy of Heuristic k-Means Clustering

Matrix	Average Accuracy (%)
$\boldsymbol{M}^B$	90.81
$\boldsymbol{M}^O$	64.25
$\boldsymbol{M}^T$	79.41
$\boldsymbol{M}^X$	73.70

As can be seen from Table 2, matrix $\boldsymbol{M}^B$ achieves the highest clustering accuracy (90.81%), followed by matrix $\boldsymbol{M}^T$ (79.41%). Using matrix $\boldsymbol{M}^X$, the average accuracy is 73.70%. Matrix $\boldsymbol{M}^O$ produces the lowest accuracy (64.25%).

It is surprising that matrix $\boldsymbol{M}^B$, which has the smallest column dimension (85), achieves the best results on average by using the heuristic k-means clustering method. The column sizes of matrix $\boldsymbol{M}^O$, $\boldsymbol{M}^T$, and $\boldsymbol{M}^X$ are 1 156, 1 689, and 1 105 respectively. These three matrices contain much more information than $\boldsymbol{M}^B$. From this result, we can see that the choice of words to construct co-occurrence matrix plays a very important role in clustering accuracy.

The highest accuracy of heuristic k-means clustering (denoted as h-accuracy) by using different matrices is shown in Table 3.

Table 3. Highest Accuracy of Heuristic k-Means Clustering

Matrix	Highest Accuracy (%)
$\boldsymbol{M}^B$	97.65
$\boldsymbol{M}^O$	81.18
$\boldsymbol{M}^T$	97.65
$\boldsymbol{M}^X$	96.47

For heuristic k-means clustering, using matrixes $\boldsymbol{M}^B$ and $\boldsymbol{M}^T$ achieves the highest accuracy, followed by using matrix $\boldsymbol{M}^X$. Matrix $\boldsymbol{M}^O$ produces the worst result and the gap between the highest accuracy and the lowest accuracy is large. The main reason is that the same opinion word is able to modify opinion target words from different categories. For example, the following two review sentences use the same opinion word "good" to modify opinion target words from two different categories (product category and seller category):

"The cell phone is good."

"The shop manager is good."

The similarity of the modification vectors causes low accuracy in words clustering.

We also observe that all the opinion target words from the logistics category can be correctly classified into the right cluster, no matter which kind of matrix we use. The words in the logistic category are distinctive from the words of the other two categories.

We also notice that a small number of words are prone to be wrongly classified. For example, "running" is one of such words. It always co-occurs with word "fast" or "slow" in the same clause. These opinion words also always co-occur with the opinion target words from the logistics category, such as "Logistics fast" or "Express very slow". That is why "running" is wrongly clustered using matrix $\boldsymbol{M}^O$.

We also notice that "running" always co-occurs with the opinion target word "speed". Similarly, "speed" often co-occurs with the opinion target words from the logistics category, which leads to the fact that "running" is easily confused with the words from the logistics category. That is why "running" is wrongly clustered using matrix $\boldsymbol{M}^B$.

The above two reasons together explain why "running" also cannot be correctly grouped using matrices $\boldsymbol{M}^T$ or $\boldsymbol{M}^X$. Content words include opinion words, opinion target words and some other words, and meanwhile opinion words and opinion target words may constitute the context of "running". This example shows that there is a small portion of words that cannot be correctly classified using distributional similarity.

Methods which consider word semantics may solve this problem.

4.2.3 *Results of Hierarchical Clustering*

For the hierarchical clustering method, no matter which matrix we use or which linkage criterion we utilize, the clustering results are far from satisfactory. Among all the linkage criteria, the algorithm using complete-linkage outperforms that using the other two criteria. However, the opinion target words from different categories are mixed together in the same cluster. There is no reasonable explanation for the whole clustering results. Therefore, we consider that the hierarchical clustering method is not suitable for coarse-grained opinion target words clustering under this scenario.

4.2.4 *Sensitivity of Centroids Selection*

From the above results, we can tell that heuristic k-means clustering outperforms standard k-means clustering and hierarchical clustering. Therefore, in this subsection, we focus on heuristic k-means clustering and discuss how centroids selection impacts clustering accuracy.

For all 7 056 combinations of centroids selection as introduced in Subsection 4.2.1, we sort their clustering accuracy in descending order for the four matrices: $\boldsymbol{M}^B$, $\boldsymbol{M}^O$, $\boldsymbol{M}^T$, and $\boldsymbol{M}^X$. Fig.1 shows the result.

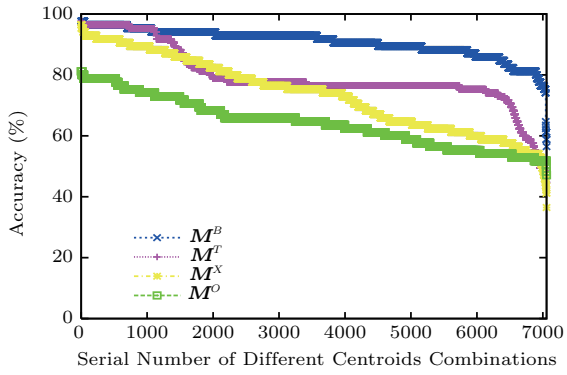


Fig.1. Sensitivity of centroids selection.

From the above figure, we can tell that $\boldsymbol{M}^B$ is the least sensitive among the four matrices when selecting centroids for clustering. In other words, as long as an expert selects three typical centroids according to his/her domain knowledge, it is very likely to produce a good clustering result. According to statistics, 48 groups of centroids combinations achieve the highest accuracy (97.65%) and 646 groups of combinations obtain the second highest accuracy (96.47%). The accuracy curve of matrix $\boldsymbol{M}^B$ declines very slowly except for the last dozens of centroids combinations.

On the contrary, the accuracy curves of matrices $\boldsymbol{M}^X$ and $\boldsymbol{M}^O$ decline much faster, which means if you cannot choose the right centroids, the clustering accuracy will be very low. The method using matrix $\boldsymbol{M}^T$ is more sensitive than the one using matrix $\boldsymbol{M}^B$. Overall, matrix $\boldsymbol{M}^B$ is more suitable and stable for coarse-grained opinion targets clustering. In addition, it is easy for domain experts to select centroids.

4.2.5 *Matrix Size*

As mentioned previously, $\boldsymbol{M}^B$ has only 85 rows and 85 columns, $\boldsymbol{M}^O$ has 85 rows and 1 156 columns, $\boldsymbol{M}^T$ has 85 rows and 1 689 columns, and $\boldsymbol{M}^X$ has 85 rows and 1 105 columns. Although $\boldsymbol{M}^O$, $\boldsymbol{M}^T$ and $\boldsymbol{M}^X$ contain much more information than $\boldsymbol{M}^B$, their performances are worse than that of using $\boldsymbol{M}^B$. Therefore, in this subsection, we discuss whether matrix size has any impact on clustering accuracy.

For $\boldsymbol{M}^O$, $\boldsymbol{M}^T$, and $\boldsymbol{M}^X$, we reduce their column dimension to 85 and obtain three 85×85 matrices. The original rows of the three matrices remain the same while their columns are composed of the top 85 high frequency opinion words, content words, and context words in corpus S, respectively. The three small-size matrices are denoted as $\boldsymbol{M}^{O\text{-}85}$, $\boldsymbol{M}^{T\text{-}85}$, and $\boldsymbol{M}^{X\text{-}85}$. We conduct the same experiment as illustrated in Subsection 4.2.1 by enumerating all the 7 056 combinations of centroids selection and compare the results with those obtained using the large matrices.

From Fig.2, we are surprised to find that for matrixes $\boldsymbol{M}^T$ and $\boldsymbol{M}^X$, the average accuracy of the heuristic k-means method using the small-size matrix outperforms the same method using the large-size matrix.

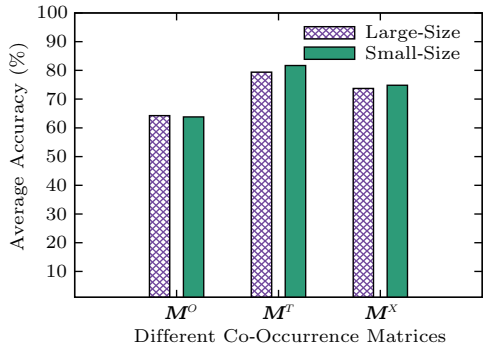


Fig.2. Average accuracy of large-size and small-size matrices.

From Fig.3, we can see that for matrixes $\boldsymbol{M}^T$ and $\boldsymbol{M}^X$, using a large-size matrix achieves a better result than using a small-size matrix. However, there are no significant differences in the highest accuracies of both size matrices. For matrix $\boldsymbol{M}^O$, the highest accuracy is the same for the large-size matrix and the small-size matrix. In this experiment, the size of the big matrix is more than ten times larger than that of the small matrix; however, larger storage has limited impact on the overall performance. Therefore, we think that there is no need to build a very large co-occurrence matrix for opinion targets clustering, because the marginal benefits are limited.

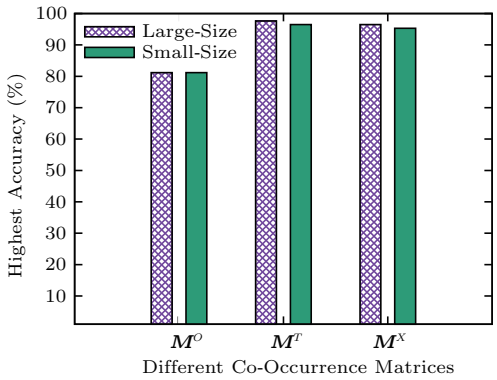


Fig.3. Highest accuracy of large-size and small-size matrices.

5 Conclusions

In this article, we aimed to cluster opinion target words based on distributional similarity in the review text. We utilized three clustering methods: standard k-means clustering, heuristic k-means clustering, and hierarchical clustering. We also introduced four different types of co-occurrence matrices: opinion target words co-occurrence matrix $\boldsymbol{M}^B$, opinion target words and opinion words co-occurrence matrix $\boldsymbol{M}^O$, opinion target words and content words co-occurrence matrix $\boldsymbol{M}^T$, opinion target words and context words co-occurrence matrix $\boldsymbol{M}^X$. Then we conducted a thorough experimental study on the performance of different methods and various matrices using real review data from Taobao.com. We evaluated their clustering accuracy, sensitivity of centroids selection, and the impact of co-occurrence matrix size.

According to our experimental results, we found that using opinion target words co-occurrence matrix achieves the best clustering result with lower time complexity and less memory space. The accuracy is stable when choosing different combinations of centroids. We also found that using small-size matrices achieves higher average clustering accuracy than using large-size matrices. Our findings provide a time-efficient and space-efficient way to cluster opinion targets with high accuracy.

We also noticed that few opinion target words cannot be correctly classified using co-occurrence matrices. In the future, we will try to integrate some knowledge or semantics into the distributional similarity method and further improve the clustering accuracy.

References

[1] Li D, Shuai X, Sun G, Tang J, Ding Y, Luo Z. Mining topic-level opinion influence in microblog. In *Proc. the 21st ACM International Conference on Information and Knowledge Management*, Oct. 29-Nov. 2, 2012, pp.1562-1566.

[2] Socher R, Perelygin A, Wu J Y, Chuang J, Manning C D, Ng A Y, Potts C. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proc. the 2013 Conference on Empirical Methods in Natural Language Processing*, Oct. 2013, pp.1631-1642.

[3] Poria S, Cambria E, Winterstein G, Huang G B. Sentic patterns: Dependency-based rules for concept-level sentiment analysis. *Knowledge-Based Systems*, 2014, 69: 45-63.

[4] Zhai Z, Liu B, Xu H, Jia P. Constrained LDA for grouping product features in opinion mining. In *Proc. the 15th Pacific-Asia Conference on Knowledge Discovery and Data Mining, Part 1*, May 2011, pp.448-459.

[5] Cambria E, Mazzocco T, Hussain A, Eckl C. Sentic medoids: Organizing affective common sense knowledge in a multi-dimensional vector space. In *Proc. the 8th International Symposium on Neural Networks, Part 3*, May 29-Jun. 1, 2011, pp.601-610.

[6] Cambria E, Hussain A, Havasi C, Eckl C, Munro J. Towards crowd validation of the UK national health service. In *Proc. the Web Science Conference 2010*, Apr. 2010.

[7] Deshpande B. How to use clustering for product categorization or segmentation. Feb. 2013. http://www.simafore.com-/blog/bid/113689/How-to-use-clustering-for-product -categorization-or-segmentation, Aug. 2015.

[8] Agirre E, Alfonseca E, Hall K, Kravalova J, Pasca M, Soroa A. A study on similarity and relatedness using distributional and WordNet-based approaches. In *Proc. the 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, May 31-Jun. 5, 2009, pp.19-27.

[9] Carenini G, Ng R T, Zwart E. Extracting knowledge from evaluative text. In *Proc. the 3rd International Conference on Knowledge Capture*, Oct. 2005, pp.11-18.

[10] Wagstaff K, Cardie C, Rogers S, Schrödl S. Constrained k-means clustering with background knowledge. In *Proc. the 18th International Conference on Machine Learning*, Jun. 28-Jul. 1, 2001, pp.577-584.

[11] Zhai Z, Liu B, Xu H, Jia P. Clustering product features for opinion mining. In *Proc. the 4th International Conference on Web Search and Data Mining*, Feb. 2011, pp.347-354.

[12] Lin D, Wu X. Phrase clustering for discriminative learning. In *Proc. the 47th Annual Meeting of the Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, Aug. 2009, pp.1030-1038.

[13] Deerwester S, Dumais S T, Furnas G W, Landauer T K, Harshman R. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 1990, 41(6): 391-407.

[14] Sahami M, Heilman T D. A web-based kernel function for measuring the similarity of short text snippets. In *Proc. the 15th International Conference on World Wide Web*, May 2006, pp.377-386.

[15] Bu F, Zhu X, Li M. Measuring the non-compositionality of multiword expressions. In *Proc. the 23rd International Conference on Computational Linguistics*, Aug. 2010, pp.116-124.

[16] Pantel P, Crestan E, Borkovsky A, Popescu A M, Vyas V. Web-scale distributional similarity and entity set expansion. In *Proc. the 2009 Conference on Empirical Methods in Natural Language Processing*, Aug. 2009, pp.938-947.

[17] Andrzejewski D, Zhu X, Craven M. Incorporating domain knowledge into topic modeling via Dirichlet Forest priors. In *Proc. the 26th Annual International Conference on Machine Learning*, Jun. 2009, pp.25-32.

[18] Zhao S, Liu T, Li S. A topical document clustering method. *Journal of Chinese Information Processing*, 2007, 21(2): 58-62. (in Chinese)

[19] Elsner M, Charniak E, Johnson M. Structured generative models for unsupervised named-entity clustering. In *Proc. the 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, May 31-Jun. 5, 2009, pp.164-172.

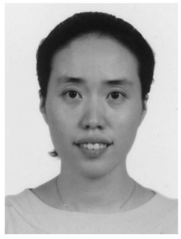
[20] Andrews N, Eisner J, Dredze M. Robust entity clustering via phylogenetic inference. In *Proc. the 52nd Annual Meeting of the Association for Computational Linguistics, Vol. 1: Long Papers*, Jun. 2014, pp.775-785.

[21] Green S, Andrewst N, Gormleyt M R, Dredzet M, Manning C D. Entity clustering across languages. In *Proc. the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Jun. 2012, pp.60-69.

[22] Chen J, Zhao Z, Ye J, Liu H. Nonlinear adaptive distance metric learning for clustering. In *Proc. the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Aug. 2007, pp.123-132.

[23] Li F, Han C, Huang M, Zhu X, Xia Y, Zhang S, Yu H. Structure-aware review mining and summarization. In *Proc. the 23rd International Conference on Computational Linguistics*, Aug. 2010, pp.653-661.

[24] Zhang Y, Zhu W. Extracting implicit features in online customer reviews. In *Proc. the 22nd International Conference on World Wide Web Companion*, May 2013, pp.103-104.



Yu Zhang now is an associate professor of Zhejiang Sci-Tech University, Hangzhou. She received her Ph.D. degree in computer science and technology from Zhejiang University, Hangzhou, in 2009. She is a member of CCF. Her current research interests include data mining, recommender system, and sentiment analysis.

Miao Liu now is a master candidate of Zhejiang Sci-Tech University, Hangzhou. She received her Bachelor's degree in computer science and technology from Hebei University of Economics and Business, Shijiazhuang, in 2013. Her research interests include data mining and recommender system.

Hai-Xia Xia now is a lecturer of Zhejiang Sci-Tech University, Hangzhou. She received her Ph.D. degree in electrical engineering from Zhejiang University, Hangzhou, in 2007. Her current research interests include pervasive computing, motor systems, finite element method and its application.