Computers and Mathematics with Applications 62 (2011) 2770–2778

Contents lists available at SciVerse ScienceDirect

Computers and Mathematics with Applications

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

A novel approach for clustering sentiments in Chinese blogs based on graph similarity

Shi Feng [a,*], Jun Pang [b], Daling Wang [a], Ge Yu [a], Feng Yang [a], Dongping Xu [b]

[a] Northeastern University, Shenyang 110819, China

[b] Wuhan University of Technology, Wuhan, China

ARTICLE INFO

Keywords:
Blog mining
Sentiment analysis
Blog clustering
Graph-based representation

ABSTRACT

Blog clustering is an important approach for online public opinion analysis. The traditional clustering methods, usually group blogs by keywords, stories and timeline, which usually ignore opinions and emotions expressed in the blog articles. In this paper, an integrated graph-based model for clustering Chinese blogs by embedded sentiments is proposed. A novel graph-based representation and the corresponding clustering algorithm are applied on the Chinese blog search results. The proposed model SoB-graph considers not only sentiment words but also structural information in blogs. Experimental results show that comparing with the traditional graph-based document representation model and vector space document representation model, the proposed SoB-graph model has achieved better performance in clustering sentiments in Chinese blog documents.

1. Introduction

Nowadays, with the fast development of Web 2.0 technology, more and more people are willing to write blog articles to express their personal feelings, emotions and attitudes on their daily activities and hot topics on the Internet. According to the reports from CNNIC (China Internet Network Information Center), by June 2009, the number of Chinese blogs has researched 300 million with 181 million blog writers, and it is reaching one fourth of the total Internet users in China [1].

Since there are a huge amount of useful blog data on the Web, how to collect, monitor and analyze the sentiment information contained in these blogs have attracted a lot of attention from both computer researchers and sociologists. Clustering blog articles by embedded sentiments means partitioning the blogs into groups according to bloggers' emotions and attitudes such as "*like*", "*agree*", "*hate*", and "*disagree*" on a certain topic. The blogs in the same group should have similar emotion and attitude, and the ones in different groups should have different emotions and attitudes. At present, most existing blog clustering methods focus on the topics, keywords, stories and timeline in blog entries, and there is limited literature on clustering the author's sentiments in blogs, especially in Chinese blog articles. Moreover, the majority of previous studies on sentiment classification usually partition the sentiments of documents into two (positive and negative) or three (positive, neutral and negative) classes, which are not enough to describe bloggers' rich emotions and attitudes contained in the blog articles. For example, there is a burst of blogs writing about the famous Chinese 110 m hurdler Liu Xiang's withdrawing from Beijing Olympics Games in late August, 2008. Traditional topic-oriented clustering techniques can generate story coherent clusters for Liu's events. However, these tools could not provide users with a summarization or guideline about people's sentiments and opinions for Liu's behavior in blogs.

In this paper, we focus on Chinese blog search results, and an integrated graph-based approach is introduced to represent and cluster Chinese blog search results by embedded sentiments. Graph-based representation [2–4] (GBR for

* Corresponding author.
E-mail address: fengshi@ise.neu.edu.cn (S. Feng).

0898-1221/$ – see front matter
doi:10.1016/j.camwa.2011.07.043

short) considered both words and their structural information such as the locations of words in a document and the relation between words. To reflect the embedded sentiments in blog articles, in this paper, we proposed an integrated GBR model for blog representation, and a K-Medoids clustering algorithm based on the new representation [5] is applied to partition blog search results into sentiment clusters. Based on the proposed approach, the clustering algorithm can divide sentiments of the blogs into more than three classes, which reflects the people's abundant emotions and attitudes contained in the blog search results on the given topic. The proposed integrated GBR document representation model can describe more useful sentiment information embedded in blogs, and thus generates more accurate clustering results.

The rest of the paper is organized as follows. Section 2 provides a brief review of related work. In Section 3, we describe the problem about clustering Chinese blogs according to bloggers' sentiments. In Section 4, we propose integrated graph representation and the K-Medoids clustering algorithm based on graph similarity. Section 5 demonstrates the experimental results. Finally, we present concluding remarks and future work in Section 6.

2. Related work

2.1. Blog mining

Blogs have recently attracted a lot of interest from both computer researchers and sociologists. One direction of blog mining focuses on analyzing the contents of blogs [6]. Glance et al. [7] gave a temporal analysis on blog contents and proposed a method to discover trends across blogs. In [8], the similarity between two blogs was calculated at the topic level, and Shen et al. presented the approach to find the latent friends who shared the similar topic distribution in their blogs.

Some papers have been published on blog clustering. Qamra et al. [9] proposed a Content–Community–Time model that can leverage the content of entries, their timestamps, and the community structure of the blogs, to automatically discover story clusters. Bansal et al. [10] observed that given a specific topic or event, a set of keywords will be correlated. They presented efficient algorithms to identify keyword clusters in large collections of blog posts for specific temporal intervals. Agarwal et al. [11] proposed a collective wisdom method to cluster blogs by label information.

Our work is quite different from the previous studies on blogs. Most of the existing work focuses on developing topic-based clustering methods or conducting content analysis for blogs. We propose a novel method to group blogs into sentiment clusters, which could facilitate public opinion monitoring for governments and business organizations.

The unsupervised clustering techniques have many potential applications for Web data. However, there are limited papers reporting on how to cluster the sentiment embedded in blogs [12–14]. Feng et al. [15] proposed an emotion-oriented clustering approach according to the sentiment similarities between blog search result titles and snippets. But only English blogs are studied by this paper. In [16], the authors used PLSA model to perform Chinese blog clustering by sentiment. However, no structural information was considered during the clustering approach. And our work in this paper considers not only sentiment words but also structural information.

2.2. Sentiment analysis

Sentiment analysis is the main task of opinion mining, and most of the existing work focused on determining the sentiment orientations of documents, sentences and words [17,18]. In document level sentiment analysis, documents were classified into positive and negative according to the overall sentiment expressed in them [19,20]. However, the sentiments that bloggers want to express are usually much more complex. Therefore, it would be too simplistic to classify the document into just positive or negative categories. In [21], a PLSA based method was used to model the hidden sentiment in the blogs, and an autoregressive sentiment-aware model was presented to predict movie box office. Lu et al. [22] used semi-supervised PLSA to solve the problem of opinion integration. In [23,24], the authors learned hidden topics from large external resources to enrich the representation of short text, which shared the similar idea of our paper. The classification and clustering approaches with the new representation have generated promising results. However, the authors focused on the topics of each text, and they did not consider the emotion information in the texts that we concern.

Different from the traditional classification approaches for sentiment analysis, in this paper, we propose a PLSA-based sentiment clustering method for blogs. An interactive sentiment clustering method on movie reviews has been proposed in [25]. Users needed to participate in the clustering approach and the results were highly relevant to the users' experiences. Besides that, the emotions expressed in blogs are more complex than in movie reviews. Our method can model the multifaceted nature of sentiments and group the blogs according to sentiments they contain.

2.3. Graph-based representation for documents

Traditional vector space model treats a document as the bag of words and employs the term vector to represent each document in the dataset. Generally a term is a word appearing in the document, and the value of the term is the absolute frequency or relative frequency of the word in the document. However, in the vector representation, a document D_1 consisting of terms a, b, c, d and e is the same as another document D_2 consisting of terms a, d, e, c and b, but D_1 maybe different from D_2 in expression because of the different order and relation between the terms. The graph-based

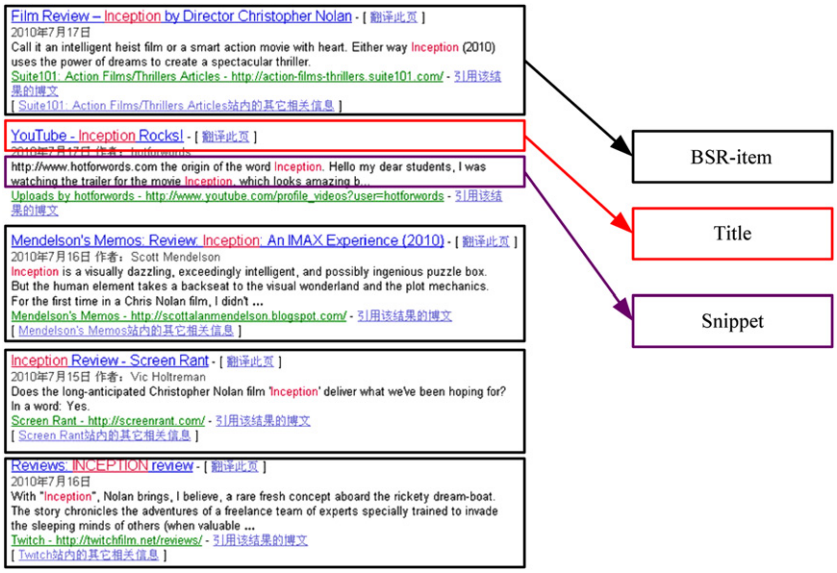


Fig. 1. An example of titles and snippets of BSR items.

representation overcomes the above shortcomings and considers the relations between the adjacent words [2–4]. The graph-based representation treated each document as a digraph and the node of the digraph denotes the word occurring in the document. The directed edge of the digraph represents the adjacent two words in the document. Therefore, the similarity between two documents can be calculated by using the similarity between two digraphs representing the documents.

Schenker et al. proposed the graph-based representation for the Web documents and performed experiments using different graph distance measures as well as various document representations that utilize graphs. The experiment results show that the graph-based approach can outperform traditional vector-based methods [4]. Hossain et al. proposed the GDClust system, which presented text documents as hierarchical document-graphs and utilized an Apriori paradigm to find the frequent subgraphs [26]. The discovered frequent sub-graphs are then utilized to generate sense-based document clusters.

Different from the graphs in [2–4], our graph-based representation for Chinese blogs can include more structural information for the embedded sentiments in blog articles.

3. Preliminaries

Given query words, the blog search engine (e.g. Google Blog Search [27]) can return a set of items. Each blog search result item includes title, snippet, URL, and other information. According to the URL, a corresponding full blog article can be crawled. Blog articles contain rich emotions of the bloggers, but with more scattered topics. For example, in a blog entry, a blogger writes down his/her review about a movie, and moreover he/she may also talk about other relevant and/or irrelevant topics in the same article, such as that day's bad weather, a delicious dinner, the sudden appearance of an old friend and so on.

For the blog search result item, its title and snippet are usually highly relevant to the query words and contain the bloggers' sentiments and opinions about the query topic. Therefore, in this paper we utilize blog search result items (BSR-item for short) instead of full blog articles to analyze bloggers' sentiments and opinions. BSR-items facilitate us to collect and analyze the public opinion about the given topic in the blogosphere. Here we give the formal definition of BSR-item.

Definition 1 (*BSR-Item*)**.** Let $R = \{r_1, r_2, \ldots, r_n\}$ represent the blog search results. Each r_i $(i = 1, 2, \ldots, n)$ is a BSR-item and can be represented as $r_i = \langle ti_i, sn_i \rangle$, where t_i and s_n mean title and snippet, respectively.

Fig. 1 demonstrates the BSR-items of the query words "*Inception*" that is a popular movie in 2010. We can see from Fig. 1 that the BSR-items have the following characteristics.

(1) The title and snippet of each item are very short, may be just one or two sentences, and sometimes may be just several words.
(2) The titles and snippets are highly relevant to the given query word. This is because the blog search engine employs sophisticated and mature Web search techniques to get the most topic relevant articles and generated snippets.
(3) The titles and snippets contain the bloggers' sentiments and opinions. As the search results are highly topic-coherent, the sentiment words in titles and snippets mainly reflect the bloggers' own opinions about the given query key word.

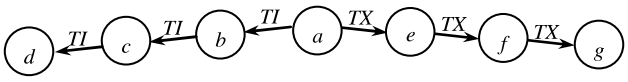


Fig. 2. An example of graph representation for document.

Fig. 3. An example of SoB-graph representation for the blog search result.

In fact, a BSR-item can contain more information besides title and snippet, but in this paper we only use contents of these two parts for extracting bloggers' sentiments.

Our purpose is to partition R into m groups according to the sentiment contained in every BSR-item (i.e., the sentiment of the corresponding blogger) and make the sentiment similarity intra group more than the sentiment similarity inter groups. Moreover, we extract the representative sentiment terms in every group. Several challenges need to be tackled during the clustering progress:

(1) How to represent every BSR-item $r \in R$?
(2) How to partition R into sentiment groups?
(3) How to show the key sentiment for each group?

To tackle these challenges, first we utilize a digraph model to represent the sentiment words and their relation contained in each BSR item (title and snippet). Then, a clustering technique based on graph similarity is applied on the proposed new representation. Finally, for the third challenge, we extract the key sentiment words in the centroid of every cluster to represent the bloggers' major emotions in each group.

4. Clustering BSR-items based on graph representation

4.1. The graph-based representation for blog search results

Schenker et al. [2–4] treated a document as a digraph for representing both words, and the corresponding relations between them (i.e., the structural information of the document). In detail, a word occurring in the document corresponds to a node in the digraph, which is marked with the word. Note that the word happening more than once only corresponds to one node of the digraph. The adjacent two words in the document correspond to a directed edge of the digraph. Meantime, the words occurring in the title part and the text part of the document are marked with *TI* and *TX*, respectively. For example, there is a document D which is made up of title "*abcd*" and text "*aefg*". Here a, b, c, d, e, f, and g are seven different words in D. So there are seven nodes separately marked with a, b, c, d, e, f, g, and six directed edges in the corresponding digraph, which is shown in Fig. 2.

Because the sentiment words play an important role in the sentiment analysis of the document, in this paper we propose an integrated graph representation with sentiment words as nodes to model the embedded emotions in BSR-items. We first conduct segmentation on every sentence in BSR-items using Chinese natural language processing tool, and then extract sentiment words by means of sentiment lexicons (e.g. NTUSD and HowNet). For providing more structural information, the proposed graph representation integrates multi graphs in [2–4] and is different from Fig. 1. The integrated graph includes every sentiment word of BSR-item and their relations for representing the sentiment of the BSR item, so we call the graph as SoB-graph.

Definition 2 (*SoB-graph*)**.** A BSR-item is modeled as a SoB-graph and each sentence in the BSR-item is modeled as a sub-graph of the SoB-graph. Each node of the SoB-graph denotes a sentiment word in the BSR-item and the nodes are marked with the sentiment word and an identifier such as "*TI*", "*SN*" or "*BOTH*" when the sentiment word occurs in the title part, the snippet part or both of them.

In SoB-graph, the sentiment word happening more than once corresponds to one node of the SoB-graph. When two sentiment words occur in one sentence unit separated by punctuation, there exists an edge between the two corresponding nodes in SoB-graph and the direction of edge is from the former word to the latter one.

In the Sob-graph, each edge is marked with three tags: (1) Position Tag "*TI*", "*SN*" or "*BOTH*" which reflect the two words occurring in the title part, the snippet part or both of them of the BSR-item; (2) Interval Number Tag which reflects the number of words between the two sentiment words; (3) Frequency Tag which denotes the occurring number of the edges in the graph.

For example, there is a BSR-item r_i which has the same words as the example shown in Fig. 1. Supposing that the word a, c, d, f and g in r_i are sentiment-bearing words and other words are not sentiment words. The non-sentiment words are eliminated, so there are five nodes and four directed edges in the corresponding SoB-graph, which is shown in Fig. 3.

By this way, a BSR-item can be represented by some sub-graphs of the SoB-graph according to the number of sentences in the BSR-item and the sentiment word distribution in the title and snippet. In this paper, the proposed BSR sentiment clustering method is based on the new SoB-graph representation, and the sentiment similarity between BSR-items is measured by the similarity between SoB-graphs.

4.2. Clustering blog search results based on graph similarity

The distance between the data is the critical part for the clustering algorithm. In the graph-based representation, the distance between two graphs is defined as

$$GDistance(G_1, G_2) = 1 - \frac{|mcs(G_1, G_2)|}{\max(|G_1|, |G_2|)} \tag{1}$$

where G_1, G_2 are two graphs, $GDistance(G_1, G_2)$ is the distance between G_1 and G_2. $|G_i|$ is the size of graph G_i, namely the sum of the number of all the nodes and edges in the graph G_i. $\max(\cdot)$ denotes the maximum of the values in the parenthesis, and $mcs(G_1, G_2)$ represents "the max common sub-graph" of G_1 and G_2. $|mcs(G_1, G_2)|$ denotes the size of "the max common sub-graph" of G_1 and G_2. The max common sub-graph of G_1 and G_2 is defined as [4] follows.

Definition 3 (*Max Common Sub-Graph*)**.** A graph G is a maximum common sub-graph (mcs) of graphs G_1 and G_2, denoted $mcs(G_1, G_2)$, if: (1) $G \in G_1$ (2) $G \in G_2$ and (3) there is no other sub-graph G' ($G' \in G_1$, $G' \in G_2$) such that $|G'| > |G|$.

Since people can use different words to express the same sentiments in blogs, different from traditional definition of the max common sub-graph, the nodes in the max common graph of SoB-graph SG_1 and SG_2 are based on the semantic similarity of nodes between G_1 and G_2. An external knowledge base HowNet is utilized to calculate the semantic similarity between the nodes. Here we give the definition of identical nodes and edges in SoB-graphs.

Definition 4 (*Identical Nodes in SoB-graphs*)**.** The two nodes g_1, g_2 are identical to each other in SoB-graphs in the following situations: (1) If g_1, g_2 have the same sentiment word sw and have the same identifier, g_1, g_2 are identical; (2) If g_1, g_2 have the same sentiment word, and one of their identifiers is "*BOTH*", g_1, g_2 are identical; (3) Supposing g_1, g_2 have the same identifier (also including one of their identifiers is "*BOTH*"), their similarity s will be calculated by the semantic similarity measurement proposed by Liu [28]; (4) If $s > \lambda$, g_1, g_2 are identical.

Definition 5 (*Identical Edges in SoB-graphs*)**.** The two directed edges are identical to each other when they have the same direction and the three tags (Position, Interval, and Frequency).

Based on the above definition, the max common sub-graph in SoB-graphs is represented by $s\text{-}mcs(SG_1, SG_2)$. The distance in SoB-graphs SG_1, SG_2 is defined as

$$SGDistance(SG_1, SG_2) = 1 - \frac{|s - mcs(SG_1, SG_2)|}{\max(|SG_1|, |SG_2|)}. \tag{2}$$

In this paper, we apply K-Medoids clustering algorithm based on SoB-graph representation. The detail of the sentiment clustering algorithm for BSR-items is described as follows.

Algorithm 1. Sentiment Algorithm for BSRs based on SoB-graph representation

Inputs: n BSR-items, the number of clusters k and similarity threshold λ;
Outputs: k clusters of BSR-items, the key sentiment words in centroids of each cluster;
Process:
(1) Conduct segmentation on the BSR-items, and eliminate the non sentiment words according to sentiment lexicons;
(2) Utilize the SoB-graphs to represent the BSR-items;
(3) Choose k SoB-graphs of the n SoB-graphs as the initial centroids;
(4) **Repeat**
(5) Compute the distances between the initial centroids and the other SoB-graphs and assign each SoB-graph into the group whose centroids has the closest *SGDistance* to the SoB-graph;
(6) Re-compute and choose the new centroids;
(7) **Until** the centroids do not change;
(8) Extract the key sentiment words in the centroids as the description of each clustering.

In the clustering process, the centroid of the cluster is a SoB-graph which has the smallest average distances to all the other SoB-graphs in the same cluster. The centroid of a set of SoB-graphs is defined as

$$SG_c = \arg\min_{\forall s \in S} \left(\frac{1}{m} \sum_{i=1}^{m} SGDistance(s, SGi) \right) \tag{3}$$

Table 1
The statistics of the labeled blog search result datasets.

Dataset	Query words	Total	Relevant	Positive	Negative	Neutral
FAR	The founding of a republic	1000	841	510	125	206
Liu	Liu Xiang	1000	977	723	54	200

where S is a set of SoB-graphs, $S = \{SG_1, SG_2, \ldots, SG_m\}$ and $|S| = m$, and SG_c is the centroid of the cluster. The distance $SGDistance(s, G_i)$ is computed using Formula (2). In the case where the centroid is not unique, i.e., there is more than one SoB-graph that has the same smallest average distance, we select one of those SoB-graphs at random as the centroid for the K-Medoids algorithm.

4.3. Extracting description words for each sentiment cluster

If the BSR-items are demonstrated directly as the results of the clustering, the sentiments and attitudes contained in the blogs in different clusters are not clear. Therefore, k centroids are exacted to describe these sentiments and attitudes contained in the blog search results of each cluster. The extracted key sentiment words will provide convenience for users to explore blogs in the blogosphere.

5. Experiments

The experiment is conducted using a commodity PC with Windows XP, Intel Core 2 Duo CPU 2.33 GHz and 2 GB memory.

5.1. Data collecting

Since there is no standard dataset to do this work, we collect real world dataset using Google Blog Search. Additionally, determining the orientation of movie reviews is very common and fundamental work for sentiment analysis. 1000 blog search results about the movie of "The Founding of a Republic" were crawled to be analyzed, with publishing date ranging from September 16th to September 30th, 2009. Another 1000 search results about Liu Xiang were crawled to analyze public opinion on Liu's comeback to the track, with publishing date ranging from September 20th to September 30th, 2009. Then the two thousand results are labeled into two groups, i.e., relevant or irrelevant to the topic of the query words. And the relevant results are divided into three categories, positive, negative and neutral. First, two volunteers separately label the results. Then the third one modifies and confirms the tags labeled by the first two volunteers. The statistic labeling results are shown in Table 1.

For evaluating the proposed method, we randomly select 100 positive, 100 negative and 100 neutral data from FAR dataset, and randomly select 100 positive, 100 neutral and all the 54 negative data from Liu dataset. Unlike English and Spanish, there is no delimiter to mark word boundaries and no explicit definition of words in Chinese languages. So the preprocessing steps need to segment Chinese text into unique word tokens. ICTCLAS [29] is a Chinese lexical analysis system, which is able to make Chinese word segmentation and part-of-speech tagging with about 98% precision.

After segmentation, the emotion words are extracted using sentiment lexicons. There are some previous papers on building sentiment lexicons [30,31]. We obtained the Chinese sentiment lexicon NTUSD used by Ku et al. [31], which contains 2812 positive words and 8276 negative words in Chinese. We also collect the data from HowNet Sentiment Lexicon (HowNet for short) [30], which contains 4566 positive words and 4370 negative Chinese words.

5.2. Evaluation metrics

The semantic similarity between sentiment words is computed with the tool, called the software package of HowNet-based vocabulary semantic similarity calculation and based on [30]. The vocabulary semantic similarity is defined as a real number with the interval between 0 and 1. The bigger semantic similarity value indicates that the two words are more similar to each other at the semantic level. Clustering performance is measured using three performance metrics.

The first performance metric is the precision, which is defined as follows. There is a dataset of n data points. The data points are labeled into k categories, $c_1, c_2, \ldots, c_k$. The dataset is partitioned into k clusters, $C_1, C_2, \ldots, C_k$ after clustering. The number of data points in C_i is n_i $(i = 1, 2, \ldots, k)$, containing n_{i1} data points with the category c_1, n_{i2} data points with the category c_2 and so forth. So the precision of the C_i is defined as

$$\Pr(C_i) = \frac{\max(n_{i1}, n_{i2}, \ldots, n_{ik})}{n_i} \quad (i = 1, 2, \ldots, k). \tag{4}$$

And the precision of the clustering is defined as

$$\Pr = \sum_{i=1}^{k} \left(\frac{n_i}{n} \Pr(C_i) \right). \tag{5}$$

Table 2
The performance comparison of the three models about the FAR dataset.

Method	Dictionary	Precision	Entropy	Rank index	Threshold
VSM	HowNet	0.366	1.572	0.479	None
	NTUSD	0.365	1.572	0.456	None
GBR	HowNet	**0.398**	1.568	0.333	None
	NTUSD	0.396	**1.557**	0.338	None
SoB-graph	HowNet	0.377	1.569	**0.526**	0.85–1.0
	NTUSD	0.375	1.570	0.494	0.90–1.0

Table 3
The performance comparison of the three models about the Liu dataset.

Method	Dictionary	Precision	Entropy	Rank index	Threshold
VSM	HowNet	0.422	1.510	0.481	None
	NTUSD	0.437	1.508	0.466	None
GBR	HowNet	0.442	1.507	0.352	None
	NTUSD	0.435	1.505	0.352	None
SoB-graph	HowNet	0.447	1.506	**0.533**	0.95–1.0
	NTUSD	**0.473**	**1.489**	0.504	0.85–1.0

The second performance metric is the entropy. Similarly, the entropy of the C_i is defined as

$$entropy(C_i) = -\sum_{j=1}^{k}(\Pr_i(c_j) * \log_2 \Pr_i(c_j)) \tag{6}$$

where $\Pr_i(c_j)$ is the proportion of class c_j data points in c_i. The total entropy of the whole clustering (which considers all clusters) is:

$$entropy_{total}(C) = \sum_{i=1}^{k}\left(\frac{n_i}{n} * entropy(C_i)\right). \tag{7}$$

The last performance metric is the Rand metric which is defined as

$$R = \frac{a+b}{a+b+c+d}. \tag{8}$$

For a given dataset of n elements $D = \{d_1, d_2, \ldots, d_n\}$, the manual annotation set $M = \{c_1, c_2, \ldots, c_k\}$ and a clustering set $C = \{C_1, C_2, \ldots, C_k\}$. a is the number of pair of elements in D that are in the same set in M and in the same set in C; b is the number of pairs of elements in D that are in different sets in M and in different sets in C; c is the number of pairs of elements in D that are in the same set in M and in different sets in C; d is the number of pairs of elements in D that are in different sets in M and in the same set in C. The Rand metric has a value between 0 and 1. The bigger the Rand metric is, the more similar the sets M and C are.

5.3. Evaluation results

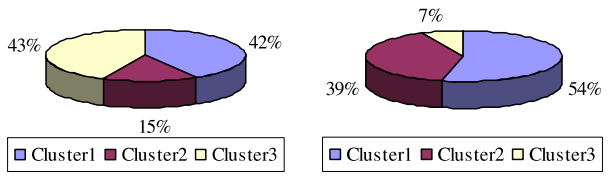
At first, 300 relevant data (100 positive data, 100 negative data and 100 neutral data) about the movie "The Founding of a Republic" are separately represented using VSM, GBR [2], and SoB-graph model and then clustered using K-Medoids algorithm. Here k equals 3 because there are only three classifications' manual annotation data to estimate the results of clustering since documents are hard to be labeled into more than three categories. We employ VSM, GBR [2], and SoB-graph model for representing BSR-items, respectively and apply K-Medoids algorithm for clustering the BSR-items. Now we show the comparison result of the three models for FAR dataset in Table 2.

From Table 2, we can see that the rank index obtained with the SoB-graph is higher than that obtained with GBR and VSM regardless of whether HowNet or NTUSD is used. At the same time, the precision obtained with the SoB-graph is less than that obtained with GBR but higher than that obtained with VSM, and the entropy is bigger than that obtained with GBR and VSM, regardless of whether HowNet or NTUSD is used.

Second, 254 relevant data (100 positive data, 54 negative data and 100 neutral data) from Liu dataset are separately represented using VSM, GBR, and SoB-graph model and then clustered using K-Medoids algorithm. Similarly, in the clustering algorithm, k equals 3. The comparison result of the three models for FAR dataset is shown in Table 3.

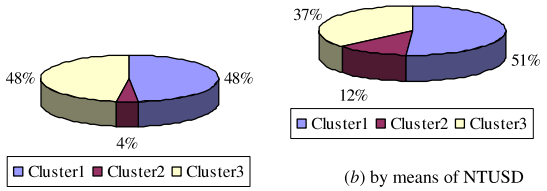
From Table 3, it is clear that the precision and rank index obtained with the SoB-graph is higher than that with GBR and VSM, and entropy is less than them.

Comparing with Table 2, the data in Table 3 show that representation with SoB-graph model is better than VSM and GBR, but the data in Table 2 has not shown such a result. This phenomenon tells us that it may have different performance values



(a) By means of HowNet. (b) By means of NTUSD.

Fig. 4. The clustering results for FAR dataset.



(a) By means of HowNet. (*b*) by means of NTUSD

(b) By means of NTUSD.

Fig. 5. The clustering results for Liu dataset.

Table 4
The extracted key sentiment words of each centroid about FAR dataset.

Dictionary	Clusters	Sentiment words
HowNet	Cluster1	Long-term, leading, determined, right
	Cluster2	Peaceful, democratic, central
	Cluster3	Freedom, white, report, across
NTUSD	Cluster1	Constantly, hold, beneficial
	Cluster2	Many, entertain
	Cluster3	Reach, reality

Table 5
The extracted key sentiment words of each centroid about Liu dataset.

Dictionary	Clusters	Sentiment words
HowNet	Cluster1	Peaceful, right, orderly
	Cluster2	Be good at
	Cluster3	Bored, to say, desire, available
NTUSD	Cluster1	A little bit, great
	Cluster2	Becoming, expectation, fantasy
	Cluster3	Unwanted, hero, golden, implementation

with the same method when the topic of the blog is different. Through analyzing the two datasets, we think that for the Liu dataset, the corresponding sentiment can discover trends across three classes. However, for the FAR dataset, corresponding sentiment should present more classes. Here we only cluster the sentiment into three clusters, so we could not achieve better performance. In fact, we can subdivide more than three sentiment classes, although the experiment only generates three sentiment classifications because it is hard to label data into more than three sentiment categories for evaluating the clustering performance.

5.4. Case study

The distribution of clustering results based on SoB-graph representation for FAR dataset is shown in Fig. 4 and the extracted key sentiment words of each centroid are shown in Table 4.

The distribution of clustering results based on SoB-graph for Liu dataset is shown in Fig. 5, and the sentiment words of each centroid are shown in Table 5.

6. Conclusion and future work

This paper proposes a new method to cluster Chinese blog search results by bloggers' sentiments. Graph-based representation and K-Medoids algorithm are applied. This document representation considers not only sentiment words but also the structural information in the blog search results, namely the word occurring in the title or snippet, the order of words and the distance of the adjacent two words here. This method subdivides the sentiment polarities, may subdivide

sentiments of document into more than three categories and could describe public rich emotions and attitudes contained in the blogs on a certain topic.

In this study, the sentiment words and their position in title and snippets are considered for the graph-based representation. In the future work, more linguistic structure information could be integrated for the representation. Moreover, we will apply the proposed sentiment clustering method to public opinion extraction studies, which will help the government to guide their propaganda programs and help individual users to make decisions.

Acknowledgments

This work is supported by National Natural Science Foundation of China (No. 60973019, 60973021) and National 863 Project (No: 2009AA01Z150). We also wish to thank the reviewers for their very useful comments and suggestions.

References

[1] CNNIC, http://www.cnnic.net.cn.
[2] A. Schenker, M. Last, H. Bunke, A. Kandel, Comparison of distance measures for graph-based clustering of documents, in: Proc. of GbRPR, New York, UK, 2003, pp. 202–213.
[3] A. Schenker, H. Bunke, M. Last, A. Kandel, A graph-based framework for web document mining, in: Proc. DAS, Florence, Italy, 2004, pp. 401–412.
[4] A. Schenker, H. Bunke, M. Last, A. Kandel, Clustering of web documents using graph representations, Applied Graph Theory in Computer Vision and Pattern Recognition (2007) 247–265.
[5] L. Kaufman, P. Rousseeuw, Finding Group in Data: An Introduction to Cluster Analysis, John Wiley & Sons, New York, 1990.
[6] J. Bar-Ilan, An outsider's view on "topic-oriented blogging", in: Proc. of WWW Alternate Papers Track, New York, NY, USA, 2004.
[7] N. Glance, M. Hurst, T. Tornkiyo, Blogpulse: automated trend discovery for weblogs, in: Proc. of WWW Workshop on the Weblogging Ecosystem, New York, NY, USA, 2004.
[8] D. Shen, J. Sun, Q. Yang, Z. Chen, Latent friend mining from blog data, in: Proc. of ICDM, Hong Kong, China, 2006, pp. 552–561.
[9] A. Qamra, B. Tseng, E. Chang, Mining blog stories using community based and temporal clustering, in: Proc. of CIKM, Washington, DC, USA, 2004, pp. 58–67.
[10] N. Bansal, F. Chiang, N. Koudas, F. Tompa, Seeking stable clusters in the blogosphere, in: Proc. of VLDB, University of Vienna, Austria, 2007, pp. 806–817.
[11] N. Agarwal, M. Oliveras, H. Liu, S. Subramanya, Clustering blogs with collective wisdom, in: Proc. of ICWE, Yorktown Heights, New York, USA, 2008, pp. 336–339.
[12] R. Kuo, H. Wang, T. Hu, S. Chou, Application of ant K-means on clustering analysis, Computers & Mathematics with Applications, 50 (10–12) 1709–1724.
[13] P. Karamolegkos, C. Patrikakis, N. Doulamis, P. Vlacheas, I. Nikolakopoulos, An evaluation study of clustering algorithms in the scope of user communities assessment, Computers & Mathematics with Applications 58 (8) (2009) 1498–1519.
[14] C. Pluempitiwiriyawej, N. Cercone, X. An, Lexical acquisition and clustering of word senses to conceptual lexicon construction, Computers & Mathematics with Applications 57 (9) (2009) 1537–1546.
[15] S. Feng, D. Wang, G. Yu, C. Yang, N. Yang, Sentiment clustering: a novel method to explore in the blogosphere, in: Proc. of APWeb/WAIM, Suzhou, China, 2009, pp. 332–344.
[16] S. Feng, D. Wang, G. Yu, C. Yang, N. Yang, Chinese blog clustering by hidden sentiment factors, in: Proc. of ADMA, Beijing, China, August 17–19, 2009, pp. 140–151.
[17] M. Efron, Using cocitation information to estimate political orientation in web documents, Knowledge Information System 9 (4) (2006) 492–511.
[18] T. Fan, C. Chang, Sentiment-oriented contextual advertising, Knowledge Information System (2009).
[19] B. Pang, L. Lee, S. Vaithyanathan, Thumbs up? sentiment classification using machine learning techniques, in: Proc. of EMNLP, Philadelphia, PA, USA, 2002, pp. 79–86.
[20] P. Turney, Thumbs up or thumbs down? semantic orientation applied to unsupervised classification of reviews, in: Proceedings of ACL, Philadelphia, PA, USA, 2002, pp. 417–424.
[21] R. Kumar, J. Novak, P. Raghavan, A. Tomkins, Structure and evolution of blogspace, Communications of the ACM 47 (12) (2004) 35–39.
[22] Y. Lu, C. Zhai, Opinion integration through semi-supervised topic modeling, in: Proc. of WWW, Beijing, China, 2008, pp. 121–130.
[23] C. Nguyen, X. Phan, S. Horiguchi, T. Nguyen, Q. Ha, Web search clustering and labeling with hidden topics, ACM Transactions on Asian Language Information Processing 8 (3) (2009) 1–40.
[24] X. Phan, M. Nguyen, S. Horiguchi, Learning to classify short and sparse text & web with hidden topics from large-scale data collections, in: Proc. of WWW 2008, Beijing, China, 2008, pp. 91–100.
[25] R. Bekkerman, H. Raghavan, J. Allan, K. Eguchi, Interactive clustering of text collections according to a user-specified criterion, in: Proc. of IJCAI, Pune, India, 2007, pp. 684–689.
[26] M. Hossain, R. Angryk, GDClust: a graph-based document clustering technique, in: Proc. of ICDM Workshops, 2007, pp. 417–422.
[27] Google Blog Search, http://blogsearch.google.com/.
[28] Q. Liu, S. Li, Word similarity computing based on Hownet, Computational Linguistics and Chinese Language Processing (2002) 59–76.
[29] ICTCLAS, http://www.ictclas.org.
[30] HowNet, http://www.keenage.com/.
[31] L. Ku, H. Chen, Mining opinions from the web: beyond relevance retrieval, Journal of the American Society for Information Science and Technology (JASIST) 58 (12) (2007) 1838–1850.