7-2010

Extracting common emotions from blogs based on fine-grained sentiment clustering

Shi FENG
Northeastern University

Daling WANG
Northeastern University

Ge YU
Northeastern University

Wei GAO
Singapore Management University, weigao@smu.edu.sg

Kam-Fai WONG
Chinese University of Hong Kong

Knowl Inf Syst (2011) 27:281–302
DOI 10.1007/s10115-010-0325-9

REGULAR PAPER

Extracting common emotions from blogs based on fine-grained sentiment clustering

Shi Feng · Daling Wang · Ge Yu · Wei Gao · Kam-Fai Wong

Received: 31 October 2009 / Revised: 22 April 2010 / Accepted: 1 July 2010 /
Published online: 21 July 2010

Abstract Recently, blogs have emerged as the major platform for people to express their feelings and sentiments in the age of Web 2.0. The common emotions, which reflect people's collective and overall sentiments, are becoming the major concern for governments, business companies and individual users. Different from previous literatures on sentiment classification and summarization, the major issue of common emotion extraction is to find out people's collective sentiments and their corresponding distributions on the Web. Most existing blog clustering methods take into account keywords, stories or timelines but neglect the embedded sentiments, which are considered very important features of blogs. In this paper, a novel method based on Probabilistic Latent Semantic Analysis (PLSA) is presented to model the hidden sentiment factors and an emotion-oriented clustering approach is proposed to find common emotions according to the fine-grained sentiment similarity between blogs. Extensive experiments are conducted on real-world datasets consisting of different topics. The results show that our approach can partition blogs into sentiment coherent clusters and the extracted common emotion words afford good navigation guidelines for embedded sentiments in each cluster.

Keywords Opinion mining · Sentiment analysis · PLSA

S. Feng (✉) · D. Wang · G. Yu
Institute of Computer Software and Theory,
Northeastern University, No.3-11 Wenhua Road,
Heping District, Shenyang, China
e-mail: fengshi@ise.neu.edu.cn

D. Wang
e-mail: wangdaling@ise.neu.edu.cn

G. Yu
e-mail: yuge@ise.neu.edu.cn

W. Gao · K.-F. Wong
Department of Systems Engineering and Engineering Management,
The Chinese University of Hong Kong, Shatin, NT, Hong Kong

1 Introduction

As the rise of Web 2.0 technologies, more and more people can freely express their comments, views, and emotions on the Web rather than just passively browse and accept information. Blogs are often online diaries published and maintained by bloggers, reporting on their daily activities and feelings. The content of blogs includes commentaries or discussions on a particular subject, ranging from mainstream topics (e.g., food, music, products, politics, etc.), to highly personal interests [19]. Blogs have provided a rich source of information and platform where Web mining practitioners could extract and summarize people's emotions and attitudes for better serving the decision-making process in various sectors of our society.

Currently, there are a great number of blogs on the Internet. Take Chinese blogs as an example, according to the newest reports from CNNIC (China Internet Network Information Center), the number of Chinese blog writers has reached 329 million by June 2009, and it has exceeded half of the total Internet users in China. Among them, about 62.7% of the bloggers regularly update their blogs [7]. With more and more valuable blogs out there, government agencies, business firms and individual users are inclined to expect a better understanding on these common sentiments or opinions from general public and study how to make the best values out of the trend of public thinking for decision-making.

Most recently, various sentiment mining techniques have been proposed to analyze people's emotions in blogosphere. Previous studies on sentiment analysis in blogs commonly focus on identifying the orientation (or polarity) of attitude of blog entries [6,23], but not exploiting the underlying emotional relatedness among different blogs. The existing studies on sentiment summarization can generate a short abstract or a list of words of the major sentiments in a close text dataset on a given topic [35]. However, for common emotion extraction task, which aims to find people's collective and overall sentiments about a given topic or a group of people, the extracted results should not only contain the summary of emotions but also include the distribution of each common emotion, i.e. the proportions of the number of people associated with each different emotion. On the other hand, previous literatures about blog clustering concentrated on exploring the keywords, stories, and timelines in blog entries, ignoring the analysis of sentiments which are important features for blogs.

Although previous studies showed promising results on classifying the polarity of blogs into positive, negative and neutral categories, a finer-grained level of emotions should be extracted considering different type of information need of users and the target of analysis. Figure 1. demonstrates an example of common emotions extraction results in blogosphere toward the accidental withdrawal of the former world champion of 110-meter hurdler Liu Xiang from Beijing Olympics [26]. Different from traditional sentiment analysis task, the

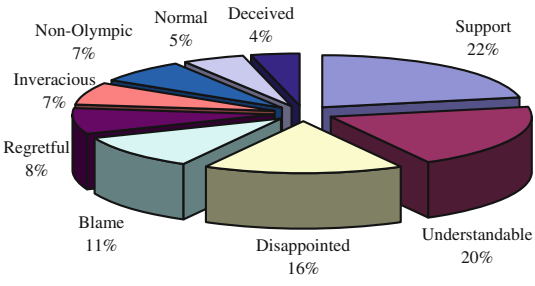


Fig. 1 Common emotions toward Liu Xiang's withdrawal from Beijing Olympic Games

major issue of this special kind of sentiment extraction is to find out the people's common attitudes toward the event and their corresponding proportional distribution on the Web. In Fig. 1, there come nine different categories of common emotions and each one reflects a typical sentiment about Liu's withdrawal, where we can see about 22% bloggers support his decision, and about 16% bloggers feel disappointed, and about 62% bloggers are among others. Clearly, the extracted common emotions and the distribution are potentially useful for many applications, such as public opinion monitoring, marketing intelligence and policy making. However, it's tough for analysts to get this report because most of the summarizing work is labor costly and needs to be done by hand without the aid of automatic tools.

From the discussion earlier, we know that common emotion extraction is not just a sentiment classification task and not the same as sentiment summarization task. There are obvious defects in existing methods, which could not totally meet the need for extracting common emotions in blogosphere. For this reason, we propose a novel finer-grained sentiment clustering approach which provides the detail insights on embedded sentiments in blogs. As in Fig. 1, the fine-grained sentiment clustering should find people's nine different common emotions and their corresponding proportions of people associated with each common emotion. The two major challenges include:

Fine-grained emotion: Human's emotions are very subjective and complex. Setting just positive, negative and neutral categories is too coarse to get the detailed common emotions toward specific information in blogs. For example, in Fig. 1, there are nine different common emotions dataset about the topic.

Synonymous emotion: Different sentiment words can express the same state of emotion. For example, the word "*bewildered*", "*doubtful*" and "*hesitant*" can express a state of confused ; the word "*energetic*", "*engrossed*" and "*interested*" can express a state of eager. More and more new words are emerging on the Web, bringing about the difficulty to completely define all this emotion relationship between words in a systematic manner.

In order to tackle these challenges, in this paper, a Probabilistic Latent Semantic Analysis (PLSA)-based method is introduced to model the hidden sentiment factors in blogs. The relatedness between blogs is calculated at the emotion state level, and a fine-grained sentiment clustering method is proposed, which goes beyond the traditional trichotomy on the categories. Common emotions are extracted by their sentiment contribution to the cluster. Evaluation results show that our proposed method can not only extract common emotions effectively but also shed light on their corresponding distribution. By this fine-grained sentiment clustering method, we can categorize blogs into groups to allow for better organization and easy navigation. For business companies and governments, they can quickly collect people's attitudes on their products and services. For individual users, our methods can provide them an exploration guide when they surf the blogosphere.

The rest of the paper is organized as follows. Section 2 introduces the related work. Section 3 analyzes the characteristics of blogs and gives the problem definition of this paper. Section 4 models blogs using hidden sentiment factors. Section 5 describes the blog sentiment clustering algorithm based on fine-grained emotion state similarity. Section 6 provides experimental results on real world datasets and finally we present concluding remarks and future work in Sect. 7.

2 Related work

There are two types of previous literatures relevant to our work. One is about blog mining and the other is about sentiment analysis.

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 [4]. Glance et al. [12] gave a temporal analysis on blog contents and proposed a method to discover trends across blogs. In [33], 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 literatures have been published on blog clustering. Qamra et al. [31] 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. [3] 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. [1] proposed a collective wisdom method to cluster blogs by label information.

Our work is quite different from the previous studies on blogs. Most of 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.

2.2 Sentiment analysis

Sentiment analysis is the main task of opinion mining, and many existing work focused on determining the sentiment orientations of documents, sentences and words [9,10]. In document level sentiment analysis, documents were classified into positive and negative according to the overall sentiment expressed in them [27,36]. However, the emotions that bloggers want to express are much more complex. Therefore, it would be too simplistic to classify the document into just positive or negative categories. In [20], 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. [21] used semi-supervised PLSA to solve the problem of opinion integration. In [25] and [30], the authors learned hidden topics from large external resources to enrich the representation of short text, which shared 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 [5]. 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.

3 Blog characteristics and problem

In this section, first we discuss the characteristics of blogs. Then we give our formal definition of common emotion extraction task.

Table 1 The blog search result items from the query word "*Liu Xiang*" (translated from Chinese)

Title	Snippet
A tragic hero–Liu Xiang	Liu Xiang is a real hero, and he perfectly embodies the Olympic Spirit. No pain, no gain. I hope he will recover soon, and return to the arena.
Continue to support Liu Xiang	Ask your elementary school P.E. teacher: why Liu cannot run because of Achilles tendon injury. I will always support you, Liu Xiang!
I feel disappointed about Liu	It is said that Liu has advanced training methods which keep him healthy. So why he got injured? I feel so disappointed about him!

3.1 Blog entries and blog search results

A blog entry is a full article that records blogger's feelings, emotions and opinions [24], which contains rich sentiments of the author. However, due to the informal and erratic style of blogs, there may be various topics in a single blog entry. For example, in a blog entry, 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.

From the discussion earlier, we know that a blog entry contains rich knowledge of emotions of bloggers but may be associated with more scattered topics. In this paper, we utilize the titles and snippets of **b**log **s**earch **r**esults, BSRs for short, to extract people's common emotions about a given topic. To exhibit the characteristics of BSRs, we issue the sample query word *Liu Xiang* to Google Blog Search [13] and restrict the date to the duration of Beijing Olympic Games. Table 1 shows three examples of search result in terms of titles and snippets for this query.

The examples in Table 1 contain keywords showing some bloggers' sentiments toward Liu Xiang's injury, such as "understandable", "supporting" and "disappointed". Following extensive analysis, we observe that the BSR items have the following characteristics:

(1) The titles and snippets are highly relevant to the given query word. This is because blog search engine employs sophisticated and mature Web search techniques to retrieve the most topic-relevant articles and snippets;
(2) The titles or 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 emotions toward the given query word;
(3) Google patents indicate that a blog will be ranked higher if it has many subscribers, many click hits and a higher PageRank value. This implies that the top ranked BSRs are from popular blog sites or may be from underlying opinion leaders [34]. Therefore, we can get a good coverage of common emotion using a relatively small dataset.

Based on this observation, we can search topic words through a blog search engine, collect the BSR items and cluster BSRs according to their embedded sentiments, in order to summarize and extract the common emotion of the bloggers on the given topic.

3.2 Problem description

Given a topic, our goal is to cluster blogs by sentiments and extract common emotions from the clusters, i.e. find people's overall attitude or opinion about the given topic. In Sect. 3.1, we show that BSRs are good source of information for extracting common emotions. Therefore, topic words are used as the query q submitted to a blog search engine and the top m

BSR items $R = \{r_1, \ldots, r_m\}$ are collected for further processing. Here, we give the formal definitions of this problem.

Definition 1 (*Sentiment Cluster (SC)*) A sentiment cluster SC_i is a group of blogs that share similar emotion meanings in the given dataset. According to the discussion in Sect. 3.1, we utilize BSRs instead of blogs to get more topic-coherent sentiment clusters.

Definition 2 (*Common Emotion Word (CEW)*) For a given sentiment cluster SC_i, the words which can reflect its major emotion meaning are the common emotion words of SC_i.

Definition 3 (*Common Emotion Distribution (CED)*) The distribution of a common emotion is the quantitative measurement of BSRs which have similar sentiments with the common emotion. In this paper, CED is calculated by $Num(SC_i)/m$, where $Num(SC_i)$ is the number of items in cluster SC_i.

The common emotion extraction task of a given topic includes finding the common emotions of the topic and estimating their corresponding distributions. To extract common emotions, we take the following 4 steps: (1) submit query q to the blog search engine; (2) cluster BSRs into SC; (3) extract CEW to label each cluster; (4) finally calculate the CED of each cluster. Note that we give common emotion a broad sense, which can reflect people's personal feelings or their sentiments on specific targets.

The existing blog clustering methods commonly partition blog entries based on keywords, stories or timelines (see Sect. 2). The granularity of existing sentiment classification methods is too coarse to reflect the complex sentiments of bloggers. To address this problem, we introduce a fine-grained sentiment clustering algorithm which provides insights on the embedded sentiments in BSR items. The similarity is calculated at emotion state level, which goes beyond just positive, negative and neutral categories. In the next section, we describe a PLSA-based method to model the hidden sentiment factors in BSRs, which can help us to discover the emotion relatedness between BSR items.

4 Modeling the hidden sentiment factors in BSRs

In this section, we propose a PLSA-based approach to model the hidden sentiment factors in BSRs. Note that although our methods and experiments are implemented on Chinese BSRs, there is nothing specific to our technique that is language dependent. When employing the sentiment lexicon of a different language, our algorithm can be directly applied to the blogs written in that language.

4.1 Sentiment lexicon acquisition and preparation

Vector Space Model (VSM) is used to represent the blog entries and each element in the vector is the weight of term frequency in blog entries. Since our goal is to cluster BSRs by their embedded sentiments, we employ Chinese sentiment lexicon to give the BSRs a new sentiment-bearing representation.

There are some previous literatures on building sentiment lexicons [16,18]. We obtained the Chinese sentiment lexicon NTUSD used by Ku et al. [18], which contains 2,812 positive words and 8,276 negative words in Chinese. We also collect the data from HowNet Sentiment Lexicon (HowNet for short) [16], which contains 4,566 positive words and 4,370 negative Chinese words.

4.2 BSRs preprocessing

With the sentiment lexicon available, the preprocessing method in this study can be described in the following steps.

(1) Given a Chinese BSR item r, we segment the sentences into words using Chinese text processing tools. Sentiment lexicon is added into the user defined vocabulary of the tools, so the precision of segment can be improved.
(2) The words, which are not in the sentiment lexicon, are filter out, so it is straightforward that we can get the set of sentiment words $W = \{w_1, \ldots, w_n\}$ in r.
(3) The frequencies of the words in W are counted, and the BSR r is represented as feature vector. Thus given a BSR set $R = \{r_1, \ldots, r_m\}$, we can use a $m \times n$ matrix $A = (f(r_i, w_j))_{m \times n}$ to describe the blogs, and $f(r_i, w_j)$ is the occurring frequency of sentiment word w_j in BSR r_i.

After preprocessing, the Chinese BSRs can be represented as sentiment matrix A and each element denotes a sentiment word and its frequency in one BSR item.

4.3 A PLSA-based approach for modeling hidden sentiment factors

The traditional methods of sentiment analysis usually focus on classifying the documents and sentences by their emotion orientation (polarity). Different from the previous work, we attempt to find the fine-grained sentiment similarities between BSRs and cluster BSRs by hidden sentiment factors.

4.3.1 Hidden sentiment factors (HSF)

The sentiments that people expressed in blogs are complex, and multi-emotions can coexist in just one blog. For example, in one BSR item, the blogger can feel sad and regretful for Liu Xiang's injury, and he or she may also express hopeful attitudes and encourage Liu to survive this injury and recover as soon as possible. Various words can be used for bloggers to express one state of emotion. The words *glad, joyful, pleased* and *delightful* can express a state of happiness; the words *sore, sorrowful, woeful* can express a state of sadness. Here, we consider the BSRs as being generated under the influence of a number of hidden sentiment factors and each hidden factor may reflect one fine-grained state of emotion embedded in the BSRs. Based on HSF, we can build an emotion state layer in BSRs, which could facilitate us calculating similarity from a new point of view.

4.3.2 Probabilistic latent semantic analysis (PLSA)

Probabilistic latent semantic analysis (PLSA) [15] and its extensions [22] have recently been applied to many text mining problems with promising results. The PLSA model can be utilized to identify the hidden semantic relationships among general co-occurrence activities. Given a BSR set R, R is generated from a number of hidden sentiment factors $Z = \{z_1, \ldots, z_k\}$. There are two basic assumptions for PLSA model. (a) The distribution of sentiment words given a hidden sentiment factor is conditionally independent of the blog; (b) The observation data pair is generated independently. Suppose $P(r)$ denotes the probability of picking a BSR r from R and w denotes the sentiment word in r. According to probability theory, we have the following formula:

$$P(w, r) = P(r)P(w|r) \tag{1}$$

We model the embedded emotions as hidden sentiment factor $Z = \{z_1, \ldots, z_k\}$ and the probability $P(w|r)$ can be rewritten by latent variables z as:

$$P(w, r) = P(r) \sum_z P(w|z) P(z|r) \tag{2}$$

where $P(w|z)$ represents the probability of choosing a word w from the sentiment word set W; $P(z|r)$ denotes the probability of picking a hidden sentiment factor z from Z. Applying Bayes rule, we can transform Formula (2) as follows:

$$P(w, r) = \sum_z P(w|z) P(r) P(z|r) = \sum_z P(w|z) P(z) P(r|z) \tag{3}$$

The EM algorithm [38] is used to estimate the parameters in PLSA model. The probability that a word w in a BSR item r is explained by the latent sentiment corresponding to z is estimated during the E-step as:

$$P(z|r, w) = \frac{P(z) P(r|z) P(w|z)}{\sum_{z'} P(z') P(r|z') P(w|z')} \tag{4}$$

And the M-step consists of:

$$P(w|z) = \frac{\sum_r f(r, w) P(z|r, w)}{\sum_{r,w'} f(r, w') P(z|r, w')} \tag{5}$$

$$P(r|z) = \frac{\sum_w f(r, w) P(z|r, w)}{\sum_{r',w} f(r', w) P(z|r', w)} \tag{6}$$

$$P(z) = \frac{\sum_{r,w} f(r, w) P(z|r, w)}{\sum_{r,w} f(r, w)} \tag{7}$$

After several iteration steps, the algorithm converges when a local optimal solution is achieved. Using the Bayes rule, we can compute the posterior probability $P(z|r)$ as follows:

$$P(z|r) = \frac{P(r|z) P(z)}{\sum_z P(r|z) P(z)} \tag{8}$$

The result $P(z|r)$ in Formula (8) represents how much a hidden sentiment factor z "contributes" to the BSR item r and the hidden factor probability set $\{P(z|r)|z \in Z\}$ can reflect the embedded sentiments in r. In the next section, the sentiment similarity between BSRs can be computed based on the hidden factor probabilities, and a fine-grained BSR sentiment clustering method is proposed.

5 Extracting common emotions from BSRs

5.1 BSR sentiment similarity measurement

When computing similarity, the traditional term vector model will be faced with sparseness problem. Figure 2 shows term vector representation of two BSR items. $r_1 = \{w_1, w_2, w_5, w_6, w_9, w_{12}, w_{16}, w_{17}, w_{20}, w_{22}, w_{23}\}$ and $r_2 = \{w_3, w_7, w_8, w_{11}, w_{15}, w_{18}, w_{19}, w_{21}\}$. There is no overlap between two vectors. However, the two BSRs may be similar from the emotion state point of view, that is to say, the words in r_1 and r_2 can reflect the same sentiment

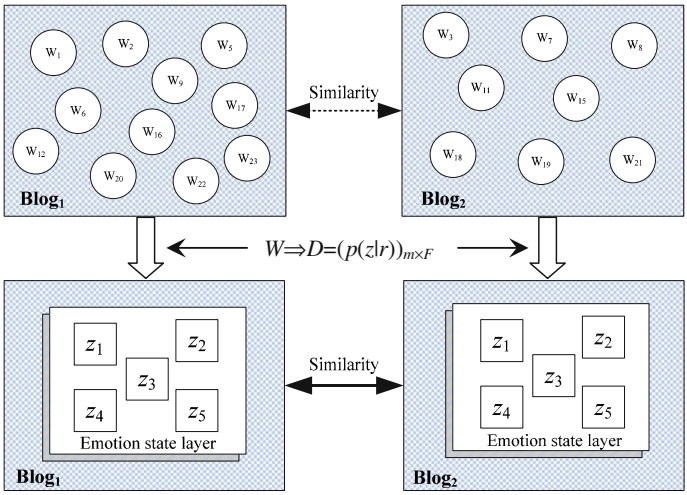


Fig. 2 The similarity measurement by hidden sentiment factors

meanings. Using Formulas (1–8), we can model bloggers' emotions by hidden sentiment factors. As can been seen from Fig. 2, the HSF has built a emotion state layer for the blog datasets, and the sentiment similarity can be calculated at fine-grained level even if there are no overlap or few overlap sentiment words between two BSR items.

From the discussion earlier, we know that the set $\{P(z|r)|z \in Z\}$ can conceptually represent the probability distribution over the hidden sentiment factor space for BSR set R. Suppose m represents the number of BSRs, F represents the number of hidden sentiment factors, then we can build a $m \times F$ matrix $D = (p(z_j|r_i))_{m\times F}$ to reflect the relationship between BSRs and latent sentiment factors, and $p(z_j|r_i)$ denotes the probability of the z_j hidden sentiment factor in BSR r_i. We let $\vec{e}$ denote the emotion state vector of BSR r_i. The distance between two BSR emotion state vectors can represent the sentiment similarity between them. Therefore, we define emotion state similarity between BSRs by applying classic cosine metric as:

$$eSim(\vec{e}_i, \vec{e}_j) = \frac{\vec{e}_i \cdot \vec{e}_j}{||\vec{e}_i|| \times ||\vec{e}_j||} \tag{9}$$

where $\vec{e}_i \cdot \vec{e}_j = \sum_{m=1}^{F} p(z_m|r_i)p(z_m|r_j)$, $||\vec{e}_i|| = \sqrt{\sum_{m=1}^{F} p(z_m|r_i)^2}$

The similarity can also be calculated directly by the sentiment word vectors of BSR items. Similarly, we have:

$$wSim(\vec{r}_i, \vec{r}_j) = \frac{\vec{r}_i \cdot \vec{r}_j}{||\vec{r}_i|| \times ||\vec{r}_j||} \tag{10}$$

where

$$\vec{r}_i \cdot \vec{r}_j = \sum_{l=1}^{n} w_{i,l} w_{j,l}, \; ||\vec{r}_i|| = \sqrt{\sum_{l=1}^{n} w_{i,l}^2}$$

Finally, the combined version of the similarity measure is defined as:

$$SentiSim(BSR_i, BSR_j) = \lambda \times eSim + (1 - \lambda) wSim \tag{11}$$

5.2 BSRs sentiment clustering and CEW extraction

The *SentiSim* function in Formula (11) provides us a new way to measure the similarity between Chinese BSRs considering both emotion state information and word matching information. Based on *SentiSim*, we can employ the existing clustering methods to do fine-grained sentiment clustering tasks. In this paper, we cluster Chinese BSRs based on K-Medoids algorithm, which has already been proven to be an effective text clustering method. Two concrete problems during the clustering are discussed as follows.

The number of clusters: Basic emotions can be defined in many ways. For most existing work on opinion mining and sentiment classification, there is a more generalized definition of people's sentiment, such as *Positive, Negative* and *Neutral*. However, for a finer-grained level, there's no authoritative definition of emotion categories in the related work of emotion detection and analysis. For example, in Quan's paper, the authors selected eight emotion classes including *Expect, Joy, Love, Surprise, Anxiety, Sorrow, Angry* and *Hate* [32]. In another paper, Yang et al. defined emotion categories as *Awesome, Heartwarming, Surprising, Sad, Useful, Happy, Boring* and *Angry* [39].

Besides the differences in the controversial definition of emotion categories, the emotions in sentences or snippets are always mixed up and most theorists appear to take a combinatorial view of emotion states in text unit of articles. For example, Plutchik talked about "mixed states", and of "dyads" and "triads" of primary emotions [29]. Averill argued for compound emotions based on more elementary ones [2].

In this paper, we intend to cluster blogs by people's common emotions of a given topic. From the aforementioned discussion, we know that each cluster could not contain only just one kind of basic emotion. Using PLSA, we have modeled the BSRs as generated from a number of hidden sentiment factors, i.e. emotion states, which is consistent with the characteristics of the mixture states of emotions in blogs. Therefore, for each cluster the extracted common emotion is mixed up by basic emotions and is widely contained in the members of the cluster.

For finer-grained sentiment clustering, we empirically set the number of cluster k to be eight. Too big k value may bring in confusion when people browsing the clustering results and too small k could not reflect the fine-grained feature of people's emotions. Note that there is no one-on-one mapping relationship between eight clusters and the eight basic emotion classed defined above, because there may be multiple basic emotions in just one cluster.

The common emotion words: Our intention is to facilitate user to get the collective emotions that bloggers express in the BSR dataset. Since we have employed a combined similarity measurement considering both emotion state vectors and term vectors, it is not very comprehensive to extract common emotion words only by the learned hidden sentiment factors from the blogs. We assume that if a word appears frequently in one cluster and seldom appears in other clusters, it will be a good discriminative and representative label for the cluster. Given a sentiment cluster SC_i, we have a document frequency and inverse cluster frequency measurement to calculate the weight of each sentiment word:

$$DFICF(w) = \frac{DF(w)}{|SC_i|} \times \log_2 \frac{k+1}{n_w + 1} \tag{12}$$

where $DF(w)$ denotes the document frequency of word w in cluster SC_i, $|SC_i|$ denotes the number of BSRs in cluster SC_i, and n_w represents the cluster frequency of the sentiment word w.

Table 2 The query words and corresponding intention for constructing the dataset

ID	Query words	Date range	Intention
Liu08	"刘翔" (Liu Xiang)	2008.8.18–2008.8.20	Extract common emotions about Liu's withdrawing from the Olympic Games
Liu09	"刘翔" (Liu Xiang)	2009.9.16–2009.9.25	Extract common emotions about Liu's back on track
FAR	"建国大业" (The Founding of A Republic)	2009.9.16–2009.9.30	Extract common emotions about the movie
Obama	"奥巴马 诺贝尔" (Obama Nobel)	2009.10.9–2009.10.12	Extract common emotions about President Obama winning Nobel prize

Algorithm 1. Fine-grained Sentiment Clustering for BSRs based on HSF (**FSC**)
Input: a set of BSR R; $P(w|z)$, $P(r|z)$, $p(z)$; the number of hidden sentiment factors F;
Output: a set of k sentiment clusters, CEWs for each cluster,
Process:
(1) for each BSR r_i, compute probability distribution $p(z|r_i)$:
$P(z|r_i) = \frac{P(r_i|z)P(z)}{\sum_z P(r_i|z)P(z)}$, where $z \in (z_1, \ldots, z_k)$
(2) construct the BSR-hidden sentiment factor matrix D;
(3) arbitrarily choose k items in D as the initial cluster centers;
(4) **repeat**
(5) compute the centroid of each cluster and assign each BSR to the cluster according to $SentiSim(r, \text{centroid})$,;
(6) recalculate centroid of each cluster;
(7) **until** no change;
(8) extract top three words of each cluster as CEWs by their corresponding *DFICF*

Here we simply use the *DFICF* in the cluster to rank the candidate sentiment words and the top three words are extracted as CEWs of the cluster. The sentiment clustering algorithm based on hidden sentiment factors is described as in Algorithm 1, which we call it as FSC algorithm.

Using Fine-grained Sentiment Clustering method (FSC), we can group the BSRs into k sentiment clusters and the BSR items in each cluster reflect bloggers' similar sentiments. Therefore, each cluster contains bloggers' coherent sentiments about a certain topic and CEWs can represent the common emotions of each cluster.

6 Experiment evaluations

6.1 Data collections

In paper [11], we pay more attention to the blog entries which contain bloggers' rich emotions. In this paper, we focus on BSRs which include more topic-relevant sentiments of the bloggers. Since there are no standard benchmarks for common emotion extraction task, in this work we collected Chinese BSRs using Google Blog Search. The four query words used in the experiments are shown in Table 2.

According to the properties of blogosphere, there are usually temporal bursts of blogs about hot topics on the Web [37]. Therefore, we restrict the published date of the blog entries within a certain range, so that we can find more topic-relevant BSRs in blogosphere. Furthermore, different date range can be used to track the evolution of bloggers' sentiments about public figures or events. For example, we issue the same query words "*Liu Xiang*" to the blog search engine with different date ranges in order to extract common emotions about Liu's withdrawing from the Games (in 2008) and his back on the track (in 2009).

For each query, the top 1,000 BSR items are extracted and processed for further analysis. The characteristics of Google Blog Search can guarantee that the top ranked BSR items are high topic relevant and from popular blog sites, which provide good sources for common emotion extracting task.

Three graduate students major in opinion mining are asked to annotate the datasets by the following rules:

(1) The first two students tag the BSR item as "Relevant" if it contains the contents that are relevant to the query words;
(2) The first two students tag the BSRs as "Positive", "Negative" and "Neutral";
(3) The first two students extract zero to three words of each BSR item that could reflect the blogger's main emotions and opinions in the BSR item;
(4) If there's a disagreement between the first two students, the third student will determine the final category and key sentiment words.

By this annotating method, the BSR datasets are classified into three categories (Positive, Negative, and Neutral). The BSRs containing the emotions such as compliment, blessing, and encouragement are classified into "Positive" category; the BSRs expressing a state of emotion such as condemnation, disappointment and sadness are classified into "Negative" category. The BSRs expressing authors' opinions, which are hard to determine orientation, are classified into "Neutral" category. Finally, the tagged BSR items are used to validate the proposed sentiment clustering algorithm when there are coarse-level predefined categories.

Like many other annotation tasks in opinion mining and sentiment analysis, the annotation of emotions in blogs is also very subjective. For orientation annotation, the two students have 54, 73, 58, 61% agreement for Liu08, Liu09, FAR and Obama datasets, respectively. For key sentiment words annotation, the two students have 58, 65, 34, 51% BSR items that has at least one key sentiment word in common. Note that there will be a discussion among the three annotators if the third annotator feels confused to decide the final annotation results. Still further clustering method will be conducted on the datasets considering more fine-grained sentiments. We employ the labeled key sentiment words to reflect the major emotion meanings contained in each BSR item and the extracted common emotion words will be compared to the labeled key words. The detail of the evaluation methods are discussed in Sect. 6.2.

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 [17] is a Chinese lexical analysis system, which is able to make Chinese word segmentation and part-of-speech tagging with about 98% precision. The result words with part-of-speech adjective, verb, proper noun, adverb and conjunction are selected for further processing and finally a sentiment word matrix $A = (f(r_i, w_j))_{m \times n}$ to represent the BSRs of each dataset.

6.2 Evaluation methods

For coarse level sentiment clustering task, i.e. when cluster number equals three, we have annotated the BSR datasets with three orientation labels. Supposing the orientation label of a BSR item is l, the result of FSC algorithm is evaluated by external clustering evaluation metrics, such as cluster entropy. For each emotion cluster SC_i, the cluster entropy E_{SC_i} is computed by:

$$E_{SC_i} = -\sum_j \frac{n(l_j, SC_i)}{n(SC_i)} \log \frac{n(l_j, SC_i)}{n(SC_i)} \tag{13}$$

where $n(l_j, SC_i)$ is the number of the paragraphs in cluster SC_i with a predefined label l_j and $n(SC_i)$ is the total number of paragraphs in clusterSC_i. The overall cluster entropy E_{SC} is then given by a weighted sum of individual cluster entropies by:

$$E_{SC} = \frac{1}{\sum_i n(SC_i)} \sum_i n(SC_i) E_{SC_i} \tag{14}$$

Other external evaluation metrics such as class entropy, normalized mutual information and cluster purity are also used to evaluate the performance of the clustering results.

For fine-grained sentiment clustering task, it is difficult to directly evaluate the clustering results, because there is no fine-grained emotion annotation for the items in our four BSR datasets. We intend to employ a relative evaluation metric to measure the compactness and separation of the clustering results. Also estimating the sentiment relatedness between BSRs and extracted common emotion words is very subjective for human beings. In this section, we introduce a quantitative evaluation method based on the semantic distance between CEWs and BSR items to measure the overall sentiment cohesiveness and consistency of the clustering results and extracted CEWs.

Since there are no external fine-grained emotion class labels for the four BSR datasets, we utilize the cluster compactness and separation to measure the emotion clustering results [14]. Given a clustering results C and a sentiment cluster SC_i, we have:

$$v(C) = \sqrt{\frac{1}{m} \sum_{i=1}^{m} d^2(r_i, \overline{r})} \tag{15}$$

where m is the number of BSRs in the dataset, d is the distance function of the two vectors, and $\overline{r}$is the centroid BSR of the dataset. The definition of cluster compactness is given as:

$$Cmp = \frac{1}{k} \sum_{i=1}^{k} \frac{v(SC_i)}{v(C)} \tag{16}$$

The definition of cluster separation is given as:

$$Sep = \frac{1}{k(k-1)} \sum_{i=1}^{k} \sum_{j=1, j \neq i}^{k} \exp\left(-\frac{d(\overline{r}_{SC_i}, \overline{r}_{SC_j})}{2\sigma^2}\right) \tag{17}$$

where $\overline{r}_{SC_i}, \overline{r}_{SC_j}$ is the centroid BSR of cluster SC_i and SC_j, σ is a Gaussian constant and in this paper, we set $2\sigma^2 = 0.25$. Finally, the overall clustering quality is measured by:

$$Ocq = \beta Cmp + (1 - \beta) Sep \tag{18}$$

The cluster compactness score evaluates the homogeneity of the clusters, and cluster separation score measures the dissimilarity among the output clusters. The lower score of *Ocq* indicates a better clustering result.

Second, we measure the sentiment distance between a CEW cw_i and a human tagged sentiment words sw_j. Directly applying word matching method cannot discover semantic relationship between words. For example, the word "like" and "love" do not match, but they share similar emotional meanings. Furthermore, the words used in blogs are usually not very standardized, so the lexicon-based measurement may suffer from the coverage problem. Normalized Google Similarity Distance (NGD) is a new method for measuring similarity between terms based on information distance and Kolmogorov complexity [8]. NGD used in this paper is based on search hits in blogosphere, defined as:

$$NGD(kw_i, sw_j) = \frac{\max\{\log h(cw_i), \log h(sw_j)\} - \log h(cw_i, sw_j)}{\log CB - \min\{\log h(cw_i), \log h(sw_j)\}} \tag{19}$$

where $h(w_a)$ denotes the Google Blog Search hits using word w_a; $h(w_a, w_b)$ denotes the hits using w_a and w_b together; CB means the total number of Chinese blog entries indexed by Google. According to the statistics, there are more than 30 billion Chinese blog articles on the Web [7]. Finally, we set $CB = 20$ billion to estimate the number of indexed Chinese blog entries.

Given a CEWs set CW of cluster C_i, let SW denotes the set of manually tagged sentiment words in BSR item r. The distance between CW and r is defined as:

$$D_{CW-r}(CW, r) = \underset{cw_i \in CW, sw_j \in SW}{avg} \{NGD(cw_i, sw_j)\} \tag{20}$$

Therefore, the cohesiveness of a cluster C_i can be defined as:

$$Cohen(C_i) = \frac{1}{|C_i|} \sum_{j=1}^{|C_i|} D_{CW-r}(CW_i, r_j) \tag{21}$$

The total quality of the clustering results is defined as:

$$Q(C) = \frac{1}{k} \sum_{k} Cohen(C_i) \tag{22}$$

The $Q(C)$ function evaluates the overall cohesiveness in the clustering results. The lower $Q(C)$ value means that the BSR items in each cluster are more coherent and the CEWs are better emotion summarization for the given cluster.

6.3 BSRs vs. Blog entries

The blog articles are written in an informal and erratic style and bloggers usually do not confine themselves to just one topic [28]. Even in one blog entry may contain many topics and corresponding opinions of the blogger. On the other hand, BSRs contain titles and topic-coherent sentences of a blog post, which provide a better platform for common emotion extraction of a given topic. Here, we analyze the percentage of the human tagged "Relevant" BSRs in the four datasets. We also ask annotators to randomly select 10% items of each BSRs dataset and read their corresponding full blog articles. The average number of topics in these full blog articles is shown in Table 3.

The "Relevant" BSRs are the blog search results that contain the content for our common emotion extraction task. For example, in Liu08 dataset, if the BSR item contains people's

Table 3 The percentage of "Relevant" BSRs and average number of topics in full blog articles

Dataset ID	"Relevant" BSRs (%)	Average # of topic in full blog articles
Liu08	97.8	1.53
Liu09	97.7	1.73
FAR	84.1	2.21
Obama	90.5	1.95

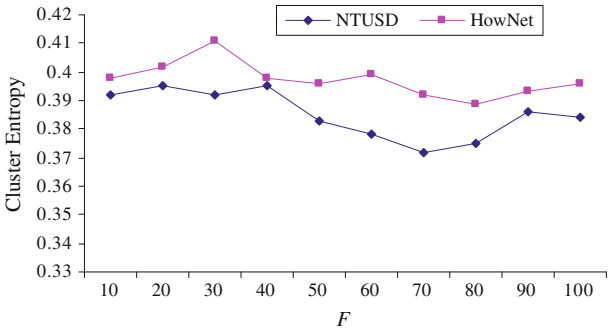


Fig. 3 Cluster entropy for dataset "Liu08" with different number of hidden sentiment factors

sentiment about Liu's withdrawing from the Games, we regard it as "Relevant". The average number of topics estimates how many topics are there in each full blog articles. If a blog entry contains more topics, it may contain more irrelevant emotions to query words. Table 3 validate our assumption that BSRs are more topic coherent than full blog entries. And also processing full blog pages are more time consuming, which could bring in more noisy emotions that are not on the topic of the query words.

6.4 Evaluation results on BSR datasets

6.4.1 Coarse level sentiment clustering

First of all, we employ the proposed FSC algorithm on four BSR datasets for a coarse level sentiment clustering task. The objective number of clusters is set to be three, which is corresponding to Positive, Negative and Neutral labels of the four BSR datasets. For this evaluation, we intend to find out if the hidden sentiment factors could catch the emotional feelings for bloggers in the give four BSR datasets. We set the number of hidden sentiment factors F ranging from 10 to 100 to check the impact of F. Since the performance of EM algorithm is relevant to the initial parameter settings, for each F, we run the sentiment clustering algorithm twenty times to get the average value. Here, we empirically set the parameter $\lambda = 0.5$ of Formula (11). Two different Chinese sentiment lexicons are employed and the final clustering performances of cluster entropy are shown in Figs. 3, 4, 5 and 6.

Above figures depict that the clustering performance is relevant to the number of hidden sentiment factors. The following tables summarize our observations for the proposed sentiment clustering algorithm compared to K-Means clustering method with cosine function between term vectors as similarity measurement.

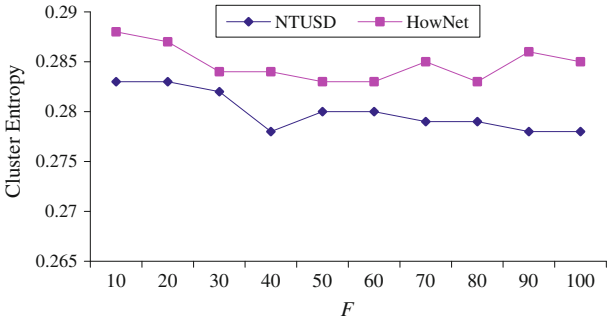


Fig. 4 Cluster entropy for dataset "Liu09" with different number of hidden sentiment factors

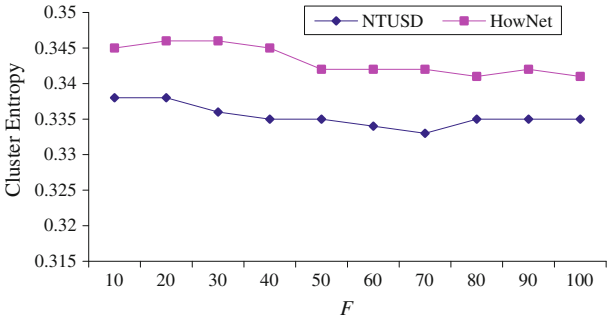


Fig. 5 Cluster entropy for dataset "FAR" with different number of hidden sentiment factors

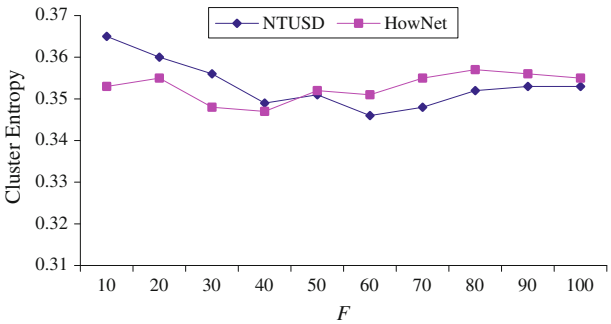


Fig. 6 Cluster entropy for dataset "Obama" with different number of hidden sentiment factors

Table 4 summarizes the best performance that we achieve for the coarse level sentiment clustering task. Generally we can get better clustering performance when using FSC algorithm. A better clustering performance is achieved on "Liu09" and "FAR" datasets than on the other two datasets. However, our FSC method achieves less improvement on Liu09 dataset. After in-depth analysis of the four datasets, taking Liu09 for example, we find that the sentiment words used in Liu09 dataset are more convergent, which reflects people's more coherent emotions about Liu's back on the track. It can be seen from above tables that blog sentiment clustering is really a hard problem (we do not get high performance results comparing to topic

Table 4 The clustering performance of different methods in four BSR datasets

DataSet	Method	Cluster entropy	Class entropy	NMI	Purity
Liu08	FSC	0.372	0.348	0.159	0.496
	K-Means	0.41	0.36	0.148	0.452
Liu09	FSC	0.278	0.283	0.202	0.64
	K-Means	0.281	0.272	0.213	0.595
FAR	FSC	0.333	0.231	0.126	0.671
	K-Means	0.368	0.297	0.122	0.531
Obama	FSC	0.346	0.26	0.285	0.489
	K-Means	0.372	0.295	0.293	0.456

Table 5 The *Qcq* score of the four datasets using FSC methods

	Liu08	Liu09	FAR	Obama
Ocq score	0.362	0.259	0.342	0.286

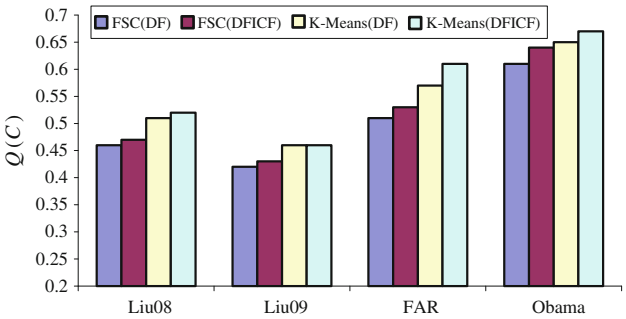


Fig. 7 The $Q(C)$ function for each dataset

clustering problems), and this may be because people's emotions expressed on controversial events are much more complex than topic words that used in the blogs.

6.4.2 Extracting common emotions using fine-grained sentiment clustering

Since there are no fine-grained emotion labels for BSR datasets, for finer-grained sentiment clustering task, we intend to measure the clustering performance using Formula (18). Here we set the parameter $\lambda = 0.5$ of Formula (11) and the number of clusters $k = 8$. We set the distance function $d = 1 - SentiSim$ and the parameter β of Formula (18) is set to be 0.5 which equally treats the weight of cluster result's compactness and separation. NTUSD is employed as sentiment lexicon and the final results are shown in Table 5.

In above table, the *Ocq* score is calculated by the average value of runs with different number of hidden sentiment factors. In Table 5, the small *Ocq* value indicates the partition result has clusters with acceptable variance and the clusters in the result are distinct. However, we cannot directly use Formula (18) to compare the proposed FSC method with term vector K-Means algorithm, because the similarity metrics of these two methods are different. Therefore, the Formula (22) is employed to measure the semantic similarity extracted CEWs and annotated key sentiment words and the result is shown in Fig. 7.

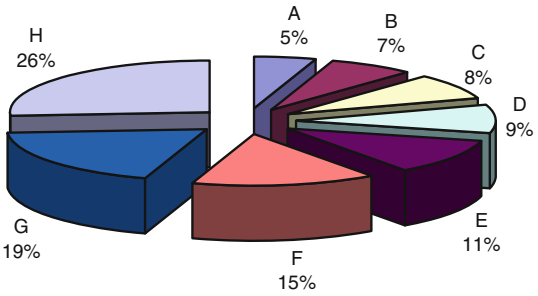


Fig. 8 The FSC clustering results for Liu08 dataset

Table 6 CEWs extracted from each cluster of Liu08 dataset

Clusters	Common emotion words
A	“期望” (expectation) ,“希望”(hope), “接受”(accept)
B	“痛苦的”(painful), “心痛”(broken-hearted), “责怪”(blame)
C	“赞同”(approve), “谢谢”(thank), “陶醉”(intoxicated)
D	“失误”(mistake), “可怜”(pathetic), “伤害”(hurt)
E	“眼泪”(tear), “没想到”(unexpected), “危机”(crisis)
F	“相信”(believe), “了解”(understandable), “勇敢的”(brave)
G	“失望”(disappointed), “受伤的”(injured), “无奈”(helpless)
H	“遗憾”(regretful), “慰问”(console), “郁闷”(depressed)

Figure 7 illustrates the $Q(C)$ function of each dataset. When using term vector-based K-Means method, the feature with a higher DF or $DFICF$ value is extracted as the CEWs of each cluster. The figures shows that the proposed FSC method can generally achieve a lower $Q(C)$ value compared to term vector-based K-Means algorithm, which means that sentiment clustering results in each cluster are more semantic related using FSC algorithm. *DF* in Fig. 7 means directly using document frequency to extract CEWs of each cluster. Using *DF* we can get a better $Q(C)$ value. However, when analyzing the extracted words, we find that *DF* sometimes generates duplicate CEWs. Therefore, we regard *DFICF* as a better indicator of cluster’s descriptive and discriminative emotion words, because *DFICF* give penalty to the words that appears frequently in different clusters.

A case study: We conduct FSC algorithm on Liu08 dataset as a case study to demonstrate the extracted common emotion words and their corresponding distribution in the BSR datasets, as shown in Fig. 8 and Table 6.

Figure 8 shows that the Liu08 dataset has been partitioned into eight clusters. We use the Formula (12) to extract CEWs in each sentiment cluster and the corresponding CEWs are shown in Table 6. In Table 6, cluster H contains an emotion state of “*regret*”; cluster A expresses a state of “*expectation*” and cluster G shows a state of “*disappointment*”. We can see from above table that CEWs provide a very brief summarization of the sentiments in each cluster. However, there are still some confusing clusters with CEWs that are hard to tell the

cohesive sentiment meanings. For example, the specific emotion meanings in cluster C are not very comprehensive.

From the aforementioned experiments, we can see that sentiment clustering is a hard problem for blog data. This is because that people usually express multi-facet sentiments in blogs on controversial topics. Grouping the blogs into just three categories (Positive, Negative, Neutral) may lose many detailed emotions embedded in blogs. The proposed clustering algorithm based on hidden sentiment factors can reflect the complexity of emotions and extracted common emotions are effective navigation guidelines for users to quickly get the bloggers' sentiments on given topics.

7 Conclusion and future work

The common emotions, which reflect people's common and overall sentiments, are important factors for governments, enterprises and individuals to judge situation and make proper decisions. This paper studied how to extract common emotions from Web blogs by clustering methods. Traditional blog clustering methods usually partition blogs by topics or keywords. In this paper, we propose fine-grained sentiment clustering algorithm to group blog search results according to hidden sentiment factors, which is modeled by Probabilistic Latent Semantic Analysis (PLSA). Experimental results demonstrate that we can generate correct clusters by their embedded emotions and extracted common emotion words are good summarization of the collective sentiments for each cluster.

In future work, more linguistic information may be considered in the new representation of blog search results. Since the performance of clustering results is relevant to the sentiment lexicons, we intend to build an appropriate Chinese sentiment lexicon for emotion analysis task. And also, the product reviews are useful data source for both individual customers and business companies. We will use fine-grained sentiment clustering method to analyze people's opinions about certain product and help latent customers and company leaders to make decisions.

Acknowledgments this work is partially supported by National Natural Science Foundation of China (No.60973019, 60973021), MOST's National 863 Project (No: 2009AA01Z150) and HKSAR ITF (No. GHP/036/09SZ).

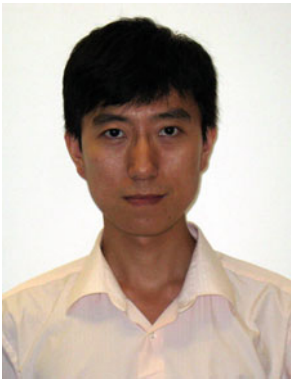
References

1. Agarwal N, Oliveras M, Liu H, Subramanya S (2008) Clustering blogs with collective wisdom. In: Proceedings of the eighth international conference on web engineering (ICWE 2008). Yorktown Heights, New York, USA
2. Averill J (1975) A semantic atlas of emotional concepts. JSAS Catalog of Selected Documents in Psychology, 5530 Ms. No. 421
3. Bansal N, Chiang F, Koudas N, Tompa F (2007) Seeking stable clusters in the blogosphere. In: Proceedings of 33rd international conference on very large data bases. University of Vienna, Austria
4. Bar-Ilan J (2004) An outsider's view on "Topic-oriented" Blogging. In: Proceedings of 13th international conference on world wide web alternate papers track. New York, NY, USA
5. Bekkerman R, Raghavan H, Allan J, Eguchi K (2007) Interactive clustering of text collections according to a user-specified criterion. In: Proceedings of 20th international joint conference on artificial intelligence. Hyderabad, India
6. Chesley P, Bruce V, Li X, Rohini S (2006) Using verbs and adjectives to automatically classify blog sentiment. In: AAAI spring symposium technical report SS-06-03
7. China Internet Network Information Center (CNNIC), http://www.cnnic.cn/en/index

8. Cilibrasi R, Vitányi P (2007) The google similarity distance. IEEE Trans Knowl Data Eng 19(3): 370–383
9. Efron M (2006) Using cocitation information to estimate political orientation in web documents. Knowl Inf Syst (KAIS) 9(4):492–511
10. Fan T, Chang C (2009) Sentiment-oriented contextual advertising. Knowl Inf Syst (KAIS) 23(3):321–344
11. Feng S, Wang D, Yu G, Yang C, Yang N (2009) Chinese blog clustering by hidden sentiment factors. In: Proceedings of 5th international conference on advanced data mining and applications (ADMA 2009). Beijing, China
12. Glance N, Hurst M, Tornkiyo T (2004) Blogpulse: automated trend discovery for weblogs. In: Proceedings of WWW 2004 workshop on the weblogging ecosystem. New York, NY, USA
13. Google Blog Search, http://blogsearch.google.com
14. He J, Tan A, Tan C, Sung S (2002) On quantitative evaluation of clustering systems. Information Retrieval and Clustering. Kluwer Academic Publishers, Dordrecht
15. Hofmann T (1999) Probabilistic latent semantic indexing. In: Proceedings of 22nd international ACM SIGIR conference on research and development in information retrieval (SIGIR 1999). Berkeley, CA, USA
16. HowNet, http://www.keenage.com/html/e_index.html
17. ICTCLAS, http://www.ictclas.org
18. Ku L, Chen H (2007) Mining opinions from the web: beyond relevance retrieval. J Am Soc Inf Sci Technol 58(12):1838–1850
19. Kumar R, Novak J, Raghavan P, Tomkins A (2004) Structure and Evolution of Blogspace. In: Commun. ACM, 47(12): 35–39
20. Liu Y, Huang X, An A, Yu X (2007) ARSA: a sentiment-aware model for predicting sales performance using blogs. In: Proceedings of 30th international ACM SIGIR conference on research and development in information retrieval (SIGIR 2007). Amsterdam, The Netherlands
21. Lu Y, Zhai C (2008) Opinion integration through semi-supervised topic modeling. In: Proceedings of 17th international conference on world wide web (WWW 2008). Beijing, China
22. Mei Q, Zhai C (2006) A mixture model for contextual text mining. In: Proceedings of twelfth ACM SIGKDD international conference on knowledge discovery and data mining (KDD 2006). Philadelphia, PA, USA
23. Melville P, Gryc W, Lawrence R (2009) Sentiment analysis of blogs by combining lexical knowledge with text classification. In: Proceedings of the 15th ACM SIGKDD international conference on knowledge discovery and data mining (KDD 2009). Paris, France
24. Nardi B, Schiano D, Gumbrecht M, Swartz L (2004) Why we blog. Commun ACM 47(12):41–46
25. Nguyen C, Phan X, Horiguchi S, Nguyen T, Ha Q (2009) Web search clustering and labeling with hidden topics. ACM Trans Asian Lang Inf Process 8(3):1–40
26. Online Opinion Channel. http://yq.people.com.cn
27. Pang B, Lee L, Vaithyanathan S (2002) Thumbs up? sentiment classification using machine learning techniques. In: Proceedings of 2002 conference on empirical methods in natural language processing (EMNLP 2002). Philadelphia, PA, USA
28. Pew (2006) Internet and the American Life Project. http://www.pewinternet.org/PPF/r/186/report_display.asp
29. Plutchik R (1962) The emotions: facts, theories and a new model. Random House, New York
30. Phan X, Nguyen M, Horiguchi S (2008) Learning to classify short and sparse text & web with hidden topics from large-scale data collections.In: Proceedings of the 17th international conference on world wide web (WWW 2008). Beijing, China
31. Qamra A, Tseng B, Chang E (2004) Mining blog stories using community based and temporal clustering. In: Proceedings of thirteen ACM conference on information and knowledge management (CIKM 2004). Washington, DC, USA
32. Quan C, Ren F (2009) Construction of a blog emotion corpus for Chinese emotional expression analysis. In: Proceedings of the 2009 conference on empirical methods in natural language processing (EMNLP 2009). Singapore
33. Shen D, Sun J, Yang Q, Chen Z (2006) Latent friend mining from blog data. In: Proceedings of 6th IEEE international conference on data mining (ICDM 2006). Hong Kong, China
34. Song X, Chi Y, Hino K, Tseng B (2007) Identifying opinion leaders in the blogosphere. In: Proceedings of the sixteenth ACM conference on information and knowledge management (CIKM 2007). Lisbon, Portugal
35. Titov I, McDonald R (2008) A joint model of text and aspect ratings for sentiment summarization. In: Proceedings of 46th meeting of association for computational linguistics (ACL08). Columbus, OH, USA

36. Turney P (2002) Thumbs up or thumbs down? semantic orientation applied to unsupervised classification of reviews. In: Proceedings of 40th annual meeting of the association for computational linguistics (ACL 2002). Philadelphia, PA, USA
37. Wong K, Xia Y, Li W, Yuan C (2006) An overview of temporal information processing. J Comput process orient lang 18((2):137–152
38. Wu X, Kumar V, Quinlan J, Ghosh J, Yang Q, Motoda H, Mclachlan G, Ng A, Liu B, Yu P, Zhou Z, Steinbach M, Hand D, Steinberg D (2008) Top 10 algorithms in data mining. Knowl Inf Syst (KAIS) 14(1):1–37
39. Yang C, Lin K, Chen H (2007) Building emotion lexicon from weblog corpora. In: Proceedings of 45th annual meeting of the association for computational linguistics (ACL 2005). Prague, Czech Republic

Author Biographies



Shi Feng is a Ph.D. student in the department of computer science at Northeastern University, China. He received his B.S. and M.E. degree from the Northeastern University, China. His research interests include opinion mining, sentiment analysis and emotion detection. He is currently a research assistant at department of systems engineering and engineering management at Chinese University of Hong Kong.

Daling Wang is a professor at School of Information Science and Engineering, Northeastern University. She received her Ph.D. degree in computer software and theory from Northeastern University of China in 2003. Her main research interests include data mining, machine learning and information retrieval.

Ge Yu is a professor at School of Information Science and Engineering, Northeastern University. He received his Ph.D. degree in computer science from Kyushu University of Japan in 1996. He is a member of IEEE, ACM, and a senior member of China Computer Federation. His research interests include database theory and technology, distributed and parallel systems, embedded software, network information security.

Wei Gao received his Ph.D. in Information Systems from the Chinese University of Hong Kong in 2010. His research interests include Information Retrieval, Natural Language Processing and Data Mining. His work is focused on effective techniques in cross-lingual query processing, and cross-lingual and cross-domain adaptive ranking for Web search. He also works on opinion retrieval and mining, and is obsessed in a long run by machine translation and automatic transliteration of named entities. His publications appear in ACM TOIS, SIGIR, ECIR and ACL.

Kam-Fai Wong obtained his Ph.D. from Edinburgh University, Scotland, in 1987. He was a post doctoral research scientist at Heriot-Watt University (Scotland), UniSys (Scotland) and ECRC (Germany). At present he is a professor in the department of systems engineering and engineering management, the Chinese University of Hong Kong (CUHK). His research interest focuses on Chinese computing, database and information retrieval. He is a member of the ACM, and IEE (UK).