Appl Intell (2014) 40:441–452
DOI 10.1007/s10489-013-0463-3

Sentiment analysis based on clustering: a framework in improving accuracy and recognizing neutral opinions

Gang Li · Fei Liu

Published online: 15 September 2013

Abstract Clustering-based sentiment analysis is a novel approach for analyzing opinions expressed in reviews, comments or blogs. In contrast to the two traditional mainstream approaches (supervised learning and symbolic techniques), the clustering-based approach is able to produce basically accurate analysis results without any human participation, linguist knowledge or training time.

This paper introduces new techniques designed to extend the capability of the clustering-based sentiment analysis approach in two aspects: firstly by applying opposite opinion contents processing and non-opinion contents processing techniques to further enhance accuracy; and secondly by using a modified voting mechanism and distance measurement method to conduct fine-grained (three classes) sentiment analysis. According to the experiment results, the clustering-based approach is proven to be useful in performing high quality sentiment analysis result, and suitable for recognizing neutral opinions.

Keywords Sentiment analysis · Opinion mining · Clustering · Semantic web

1 Introduction

The World Wide Web has become the most popular way to obtain information. If someone wants to purchase goods, find a travel destination or select a service, they generally seek an online review to assist their decision making. According to two surveys [1, 2], 81 % of Internet users have conducted online research on a product at least once; and, of the readers of online reviews of restaurants, hotels, and various services (e.g., travel agencies or doctors), between 73 % and 87 % reported that online reviews had a significant influence on their purchase [3].

Sentiment orientations of online opinion expressing documents, which including comments, feedback, critiques, reviews and blogs, are believed to be important sources of information for effective decision making [4]. Currently sentiment detection is a discipline at the crossroads of Natural Language Processing and Information Retrieval, and as such, it shares a number of characteristics with other tasks such as information extraction and text mining [5]. A large body of research has attempted to find a method to automatically analyse the sentiment orientation of these types of documents. This research can be roughly classified into two classical academic streams: supervised machine learning (classification) approaches and symbolic techniques [6]. Although supervised machine learning approaches enjoy a relatively high accuracy, its process requires human participation. Symbolic techniques, on the other hand, do not demand human involvement; its accuracy is limited. In order to address the drawbacks of these approaches, a novel clustering-based sentiment analysis approach was proposed by Li & Liu [7]. The framework was established, and it was proven to be more practical for real applications on the aspect of accuracy and efficiency.

While the clustering based sentiment analysis approach performs well, there is still room for its improvement. For example, we could further explore the potential of clustering based sentiment analysis framework to improve the accuracy level of binary analysis. Additionally, the basic clustering based sentiment analysis framework [7] has not demonstrated its ability of handling non-binary sentiment analysis which has been discussed in supervised learning

G. Li (✉) · F. Liu
Department of Computer Science and Computer Engineering, La Trobe University, Melbourne, Australia
e-mail: g9li@students.latrobe.edu.au

approaches [10]. The contribution of this paper is to introduce new techniques to extend the current clustering based framework with the aim to further enhance the accuracy of the framework and to address the issue of non-binary sentiment analysis.

The rest of the paper is organized as follows. In Sect. 2, we briefly review the two mainstream approaches, with a discussion of the existing achievement of the clustering-based approach. The extending directions and methods of the clustering-based approach are established in Sect. 3. In Sects. 4, 5 and 6, the experiment results and analysis are presented in order to evaluate the performance of the new techniques. Section 7 concludes the paper and suggests future research directions.

2 Related work

As mentioned in Sect. 1, there were two mainstream approaches to conducting sentiment analysis [5] before the clustering-based approach was proposed. All these approaches are started with addressing the binary sentiment analysis issue. The first one is the supervised machine learning (classification) approach. Pang and Lee (2002) conducted the most representative research on supervised learning [8]. They further extended their supervised learning framework in 2004 [9] and 2005 [10] by applying subjective sentence extraction and considering a 3-point or 4-point scale, respectively. Another mainstream approach is the symbolic technique or semantic analysis, with notable research in this area conducted by Cesarano [11], Kamps and Marx [12] and Turney et al. [13]. Both of these approaches are processed under the classic vector space model [14] by converting the collection of object opinion expressing documents into an $m * n$ matrix, where m is the number of documents, and n is the total number of terms in these documents [15].

Cluster analysis (clustering) is a technique which divides data into groups (clusters) that are meaningful, useful, or both. The process of clustering aims to discover natural groupings, and thus presents an overview of the classes in a collection of documents [16]. Since clustering analysis grouping objects based merely on the internal properties, which means no additional or supervised information are required, it satisfies our requirement for a prospect sentiment analysis method. Cluster analysis techniques can be categorized into hierarchical and partitional clustering. Hierarchical algorithms create a hierarchical decomposition of data set [17]. If we permit clusters to have subclusters, then we obtain a hierarchical clustering, which is a set of nested clusters that are organized as a tree. In contrast, a partitional clustering simply divides the set of data objects into non-overlapping subsets such that each data object is in exactly one subset. In the application of binary sentiment analysis, there is no nested subcluster in both positive and negative document groups. Therefore, hierarchical clustering technique is not suitable for sentiment analysis. On the other hand, partitional clustering is more appropriate for binary sentiment analysis because it only looks to divide the review corpus into two sentiment groups. K-means is a typical partitional clustering algorithm, which attempts to find a user-specified number of clusters (K).

2.1 Supervised machine learning approach

Supervised machine learning is also called classification when the class feature is discrete. There are many popular classification algorithms, for instance SVM, decision tree and Naive Bayes. All of these algorithms require pre-tagged training data to identify a model that best fits the relationship between the attribute set and class label of the input data [18].

The most widely known research to apply the classification approach on document-level sentiment analysis was conducted by Pang and Lee [8] in 2002. They used 700 positive and 700 negative pre-tagged documents as training data to build a model with Naive Bayes, maximum entropy and SVM, respectively. They also discussed the performances under different feature types (unigrams, bigrams, etc.) and data forms (frequency or presence), obtaining accuracy rates between 72.8 % and 82.9 %. The best analysis result is obtained by using SVM with the feature of unigrams and represented in the form of presence of data.

Pang and Lee's classification framework was extended by many researchers, including themselves. In 2004, Pang and Lee introduced a subjective sentence extraction method to reduce the dimension of the data set and improve the accuracy rate [9] to about 85 %. Recently, some newly emerged research based on supervised learning approaches, such as Xia et al. [19], Tan [20] and Bai [21]. Xia et al. added ensemble techniques to the data set. This improved the accuracy of all Naive Bayes, Max Entropy and Support Vector Machine over 85 %. Bai also applied Naive Bayes classifiers into this area, but he paid more attention on finding a new feature selection method. Two new features: information gain and two stage Markov blanket classifier enhance the accuracy rate up to 92 %. Tan applied another classifier, K-nearest neighbor (KNN), to achieve accuracy over 90 %.

On the other hand, in 2005, Pang and Lee extended their research from binary classification (positive and negative only) to a 3-point or 4-point scale [10].

2.2 Symbolic techniques

In contrast to machine learning approaches, symbolic techniques consider the sentiment analysis problem from a linguistic perspective, assigning each term a sentiment score.

The score is a measurement of the direction and intensity of the term on a scale of positive or negative. Once the scores for each term are obtained, the score for the whole document can be calculated by applying aggregation functions. Usually, the function is an average or sum.

The performance of symbolic techniques relies heavily on scoring methods. There are many approaches to measure the term score. For example, Cesarano et al. introduced a scoring method in [11], where people were asked to give a score; Kamps and Mar [12] scored terms by applied WordNet synsets; and Truney's PMI-IR approach relied on Web search [13].

There has not yet been found any research which applied symbolic techniques to deal with three or more classes sentiment analysis problem.

2.3 Clustering-based approach

After reviewing and evaluating these two mainstream approaches, Li and Liu [7] proposed a new binary sentiment analysis approach based on clustering to overcome several of the drawbacks of the previous methods. Although the accuracy rate of the supervised learning approach is good (around 80 %), the costs are high for several reasons. Firstly, supervised learning requires a large volume of tagged training data. Signing labels can be a slow and expensive process because manual inspection and domain expertise are needed [22]. Additionally, the supervised learning approach is highly dependent on the domain of training data. In other words, it lacks generality [23]. Symbolic techniques sometimes also require human participation and rely on human summarized superficial linguistic knowledge, instead of mining the information from document objects. Consequently and more importantly, the accuracy rates of symbolic-based approaches are usually lower than 70 %.

The clustering-based sentiment analysis approach applies an unsupervised learning algorithm: k-means. Initially, it neither requires any human tagged training data, nor time for training. Therefore, compared with the supervised learning approach, it has a competitive advantage from the outset. From another point of view, the clustering-based approach does not rely on but also does not reject linguistic knowledge. Nonetheless, since the research adopts k-means as the specific clustering algorithm, a new challenge must be faced, that is, the instability issue of the clustering results caused by the random selection of initial centroids. K-means is extremely sensitive to the choice of initial centroids [24]. Therefore, some researchers are focused on this issue in specialty, which include Laszlo and Mukherjee [25] who applied a genetic algorithm to detect global optimum. Moreover, Poomagal and Hamsapriya [26], Menendez and Camacho [27] also contributed to this research.

A series of experiments had been conducted by Li and Liu [7] based on a set of movie review data which is also used in Pang and Lee's supervised machine learning research [8]. This document set consists of 1000 positive and 1000 negative movie reviews. These review documents are pre-tagged based on their sentiment classes. To speed up the experiment, we randomly chose 300 positive and 300 negative documents to construct a balanced distributed experiment data set.

Adjectives and adverbs are selected as the terms to construct the document matrix in the form of frequency and presence. Preliminarily, by directly running the k-means algorithm with these two matrixes, the average accuracy rate for both frequency and presence of data are around 55 %. This accuracy level is far lower than the supervised learning or even symbolic techniques, which is unacceptable in real applications. Also, as anticipated, the results are unstable. In 20 times running of clustering processes, the accuracy fluctuates, with a standard division of 2.6 % for frequency and 4.9 % for presence.

To achieve a more acceptable accuracy level, the TF-IDF (Term Frequency-Inverse document Frequency) weighting method was applied on the documents matrixes. This is a weighting method to evaluate how relevant a word in a corpus is to a document [28]. Mathematically, the TF-IDF weight W_i of a term i can be expressed as:

$$W_i = tf_i * \log(D/df_i)$$

In this expression, tf_i is the term frequency of term i in a document, D is the number of documents in the corpus, and df_i is the document frequency or number of documents containing term i. By applying this weighting method, the accuracy rates for both frequency and presence data are dramatically increased to 72.2 % and 73.1 %, respectively.[1]

However, the issue of instability becomes more serious after this. The standard division values increased to 4.02 % and 6.7 %. To solve this problem, Li and Liu set a voting mechanism. Under this mechanism, the final group $f(d)$ of a document d is not determined by any individual clustering process, but was voted by running results of clustering multiple times (see the following equation, where n(Vj=positive) is the times of document d voted as positive, vice versa).

$$\text{f(d)} = \begin{cases} \textit{positive} \text{ if } n(\text{Vj} = \textit{positive}) \geq n(\text{Vj} = \textit{negative}) \\ \textit{negative} \text{ if } n(\text{Vj} = \textit{positive}) < n(\text{Vj} = \textit{negative}) \end{cases}$$

This mechanism greatly improved the stability of the clustering result. The standard division value reduced to 1.55 % and 0.8 %, meanwhile, the accuracy rates also increased by about 3 % to around 76 %.

[1] Once the TF-IDF weights are calculated by using frequency of data, the weigh values are also able to be applied on presence of data.

Table 1 Performance analysis for three streams of approaches

	Accuracy (Binary)	Efficiency	Human participation	Multiple Class Analysis Ability
Symbolic Techniques	65.83 %	Very fast	Mostly No	Not proved yet
Supervised Learning	77 %–82 %	Slow on training & fast on test	Yes	Yes
Clustering-based Approach	77.17 %–78.33 %	Fast	No	Not proved yet

As mentioned, the clustering-based approach does not reject linguistic knowledge. The clustering outcome obtained further improvement by importing term scores. Specifically, all terms measure the distance (WordNet synonym distance) to reference words 'good' and 'bad', the distance denoted by X. Consequently, a weight W can be calculated by an experimentally derived (thresholds values are verified by experiments) function:

$$W = \begin{cases} 1.2 - (X-1) * 0.02 & \text{if } X \leq 8 \\ 1 - (X-1) * 0.1 & \text{if } 8 < X \leq 11 \end{cases}$$

Therefore the terms which are closer to any reference word obtain a larger weight value than those far away to reference words. After applying the weight to document matrixes, accuracy rates were finally achieved to the level of 77.88 % and 77.25 % for two forms of data, with low standard division (0.4 % and 0.9 %), which indicates that they are very stable. Additionally, more than half the terms which are not able to link to the reference words (non-sentiment expressing terms) are detected and eliminated. In other words, the dimension of the matrix was greatly reduced. As described in [29], the complexity of k-means is $(n * K * I * d)$, where n is the number of vectors, K is the number of clusters, I is the number of iterations and di is the number of dimensions. Thus, complexity dramatically decreased and the clustering process became more effective.

In summary, up to now, clustering based analysis approach is a well performing framework for addressing the binary sentiment analysis issue.

2.4 Performance analysis of current sentiment analysis approaches

Hereby, we summarize and compare the performances of the current existing three streams of sentiment analysis approaches to analyze theirs advantages and disadvantages respectively. The results of the analysis are listed in Table 1.

All of these approaches are initially designed for binary sentiment analysis. Among these approaches, the supervised learning approaches generate the highest accuracy, but the cost is also high in terms of time and human participation. On the other hand, symbolic approaches run very fast, but the accuracy rates are usually poor. Our clustering based sentiment analysis approach performed balance on the aspects of both accuracy and cost, and its performance is advanced compared to the average performance of supervised learning and symbolic techniques.

For the issue of multiple classes sentiment analysis, up to this stage, only supervised machine learning approaches demonstrated competency [10].

Therefore, for our object of research, clustering based sentiment analysis approach, there are two aspects that may need further improvement. The first one is to achieve the same level of accuracy as those of supervised learning approaches. The second one is to establish an approach for multiple class sentiment analysis, though it may be difficult to improve the efficiency to compete with symbolic techniques. This is because the difference of complexity levels is substantial.

3 Extensions of clustering-based sentiment analysis

As discussed in the previous section, the clustering-based sentiment analysis approach has many advantages compared to traditional approaches. However, our clustering-based sentiment analysis approach is not fully developed, and there are many as yet unexplored opportunities to extend this research in order to enhance accuracy and perform finer grained analysis.

3.1 Extension direction

Firstly, the accuracy rate of the clustering-based approach is about 77 %, and taking into account its advantage in terms of low cost and effectiveness, this accuracy rate is acceptable. Nevertheless, with the development of advanced computing techniques, especially cloud computing techniques, the issue of computing cost is not as serious as it was before. Users care more about analysis accuracy than effectiveness. Therefore, it is important to find a method to further enhance accuracy.

Secondly, similar to most of the existing research, clustering-based sentiment analysis currently only focuses on the binary distinction of positive vs. negative. But it is believed that having more information than this binary sentiment analysis [10] will be helpful. Some opinion expressing documents are neutral, since the ratio of positive and negative content is almost balanced.[2] The clustering-based sentiment analysis approach has not yet proved that it is competent to address this issue. Finding the appropriate technique to solve fine-grained (three classes) sentiment analysis is an important extension, since this could provide more precise sentiment analysis results to support decision making.

3.2 Overview of the solutions of accuracy enhancement

In an ideal scenario, a document expresses either a positive or negative opinion exactly using a bag of positive or negative words without any other content, allowing the document vectors to perfectly cluster into two non-intersecting groups. However, this generally does not happen in the real world. Usually, there are many opposite opinion (positive word in negative document or reverse) or non-opinion (objective sentences) contents in a real opinion expressing document.

Case 1 "*Director Martin Scorsese has crafted a sumptuous and dazzling visual feast from Edith Wharton's novel of social propriety and repressed desire*, *but one that suffers from an overlong second act and a lack of character depth*."

Case 2 "*On the day of their engagement announcement*, *Archer is re-introduced to May's cousin Ellen Olenska* (*Michelle Pfeiffer*), *an unhappily married countess who has left her European husband for the support of her family in New York*."

The sentence in case 1 is an excerpt from a negative review of a movie. Although this sentence expresses a negative opinion of the movie, the words '*sumptuous*', '*dazzling*', '*feast*' and '*propriety*' express a strong positive sense.

Case 2 is also a description of a movie plot. It does not present the writer's opinion of the movie, however, it contains the word 'unhappily' which expresses a negative sentiment.

It is believed that this sort of content obstructs the judgment of our clustering-based sentiment analysis system. In other words, the accuracy rate is expected to increase if we can properly process and filter this content. Therefore, for non-opinion expressing content, the strategy is simply to eliminate them to make the opinion contents more notable. For opposite opinion content, we can convert them to their opposite side to enforce the real opinion direction.

3.2.1 Opposite sentiment contents processing

Usually, the opposite sentiment contents can be identified by key words. After inspecting ten movie reviews, we found that more than 90 % opposite sentiment words appear in transitional sentences or are followed by negation words. For example, the sentence in case 1 is a typical transitional sentence, as the transition conjunction 'but' splits the sentence into two sections. The second section expressed the real sentiment (negative) of the writer; meanwhile, the first section expressed the opposite sentiment (positive). Transitional sentences are usually in such a two-section structure and they usually can be identified by transitional conjunction words such as 'but', 'though', 'despite' etc.

On the other hand, negation words also cause the appearance of opposite sentiment contents. Das and Chen [20] claimed that whenever a negation word appears in a sentence, it usually causes the meaning of the sentence to be the opposite of that without the negation word. They handle negation by detecting some pre-defined key words, such as 'not' 'never', etc. And then they tagged the contents after a detected negation key word in a sentence as negated contents which represent opposite sentiment. Therefore, by appropriately identifying numbers of key words and adopting Das and Chen's negation definition approach, we can capture the majority of opposite sentiment contents.

3.2.2 Non-opinion contents processing

It is not possible to distinguish non-opinion content from opinion content simply by linguistic characters (key words or sentence structure) as there is no significant grammatical difference between them. This is quite a complicated issue, involving a specific field of research known as *subjectivity detection*. Interestingly, similar to sentiment analysis, since it is also a binary distinction analysis problem, the solutions of subjectivity detection also have two streams: symbolic techniques and the machine learning approach. It seems that the clustering approach is not appropriate for this task, since there is no significant feature that could make sentences gather into clusters by their objective or subjective shape.

Yu and Hatzivassiloglou proposed several subjectivity detection approaches in [30] at both the document and sentence level. At the sentence level, their first approach is a symbolic-based method which is under the assumption of "Opinion sentence will be more similar to their sentence than to factual sentence". Furthermore, they use SIMFINDER [31] which is a system for measuring sentence similarity based on shared words, phrases and WordNet synsets. Their second approach is based on a Naive Bayes classifier with some pre-tagged opinions or face document as training data. Finally, they apply multiple Naive Bayes classifiers to boost classification accuracy.

[2] In this paper, documents with large proportions of objective content are not regarded as neutral documents. The object of the study is opinion expressing documents, though they usually involved small proportions of objective content.

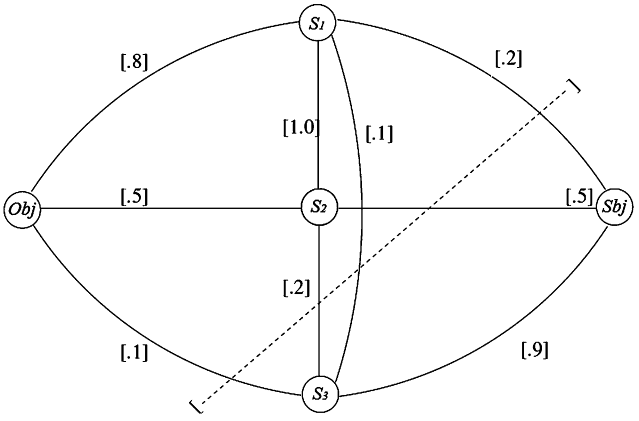


Fig. 1 An example of graph G, where the numbers in square brackets are the individual scores and association scores. The *dotted line* expresses the best cut, the cost of this cut is $0.2 + 0.5 + 0.1 + 0.2 + 0.1 = 1.1$, which is the minimal. In this case, $S1$ and $S2$ are Objective and $S3$ is Subjective

Pang and Lee introduced the cut-based classification approach to improve their supervised learning approach in [9]. The cut-based classifier can be regarded as a combination of Yu and Hatzivassiloglou's two methods. For a set of sentences $(S1, \dots, Sn)$, it not only measured the probability of an individual sentence belonging to a certain class (individual score, denoted as $ind1(Si)$ and $ind2(Si)$), it also took into account the importance of two sentences in the same class (association score, denoted as $assoc(Si, Sk)$). They use 5000 subjective sentences and 5000 objective sentences to train a Naive Bayes model. The individual score can be the Naive Bayes' (or other supervised classifiers') estimation of probability that sentence Si is a subjectivity or objectivity sentence, and the association score can be the distance between two sentences. Once the individual scores and association scores of each sentence are obtained, a graph G (see Fig. 1) can be constructed to find the best cut. A cut of G is a partition of sentence nodes into two groups, in which some nodes are on the objective side and the remainders are on the subjective side. To find the best cut, we need to ensure the sum weight of cut edges is minimal. Algorithm 1 explains this technique in detail.

This cut-based classification technique could provide higher than 90 % accuracy for subjective detection. When applying this technique to Pang and Lee's sentiment classification system, accuracy was improved to about 86 %. This technique resulted in a 4 % improvement over the supervised learning system, but it has never been used in clustering based sentiment analysis before. Therefore, we expect that this technique could increase the accuracy of our clustering-based sentiment analysis approach.

Algorithm 1 Cut-based Subjectivity Detection

```
Input: A corpus $C$ includes $n$ opinion express documents $\{di, \dots, dn\}$.
Output: A new corpus $C'$ includes $n$ documents $\{d1', d2', \dots, di', \dots dn'\}$ only contains subjective sentences.
Methods:
Begin
foreach document $di$
    split $di$ into $Mi$ sentences $\{si1, si2, \dots sij \dots siMi\}$ by detecting punctuations ('.', '?', or '!') to construct a sentence corpus with size of total $\sum_{i=1}^{n} M_i$
endfor
use external dataset (e.g. 5000 subjectivity sentences and 5000 objectivity sentences) to train a Naive Bayes model
apply this model to predict the subjectivity/objectivity probability $PSij$ and $POij$ of sentence $sij$ ($PSij + POij = 1$) ($PSij$ and $POij$ can be regard as the individual score of sentence $sij$, $ind1(Sij)$ and $ind2(Sij)$)
foreach pair of sentences $sa$ and $sb$
    calculate the distance (e.g. Cosine distance) $d(sa, sb)$ between them, this distance can be regarded as the association secore $assoc(sa, sb)$
endfor
construct a graph G, with the vertices $\{s1, s2, \dots, stotal, Obj, Sbj\}$; link $Obj$ with $\{s1, s2, \dots, stotal\}$; link $Sbj$ with $\{s1, s2, \dots, stotal\}$; link each pair of two sentence vertices $\{si, sj\}$.
Assign individual scores and association scores to edges as the cost.
Find the cut with minimum cost, assign each sentence a $Obj$ or $Sbj$ label.
Foreach sentence $sij$
If $sij \in Obj$
Delete $sij$ from document $di$
End
```

3.3 Overview of the solution of three *classes* analysis

Due to the fact that a large amount of expressing documents do not express a very clear tendentiousness, positive and negative contents ratio are relatively balanced. Therefore, we need to find an approach to classify a review corpus into three classes: positive, negative and neutral. Intuitively, this is simply a superficial change of the target class from two to three. For our cluster-based approach, we only need to add another randomly selected centroid. However, this is different to topic-based text clustering [32], in which the clusters' meaning can be regarded as irrelevant. There is an ordinal relation between the positive, neutral and negative groups.

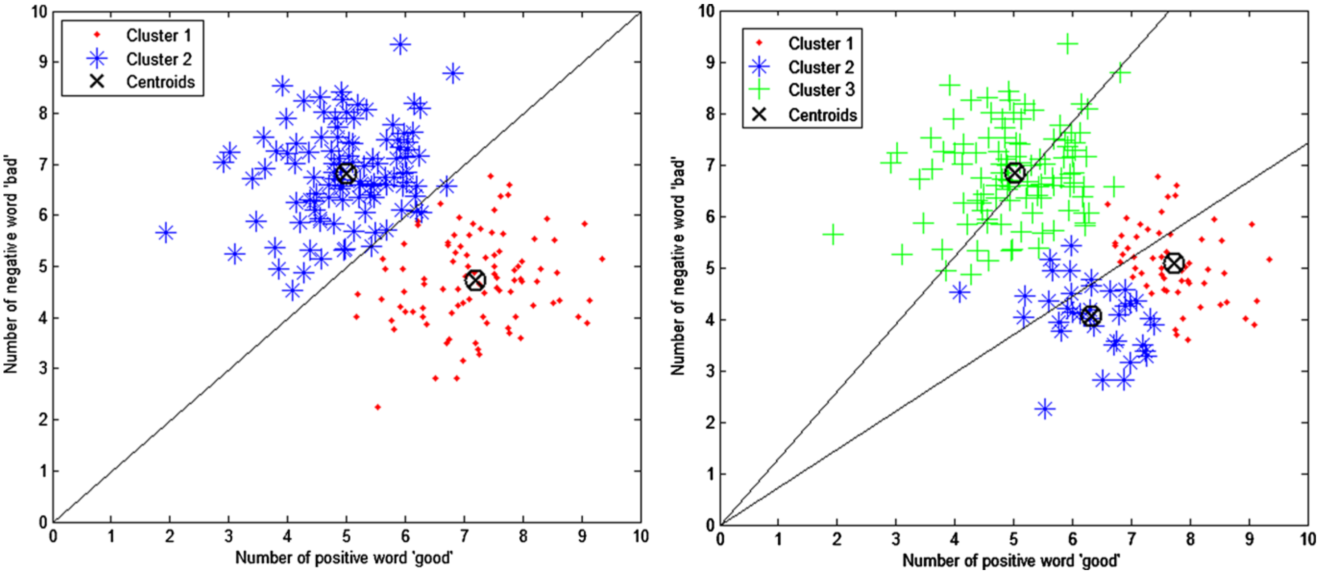


Fig. 2 (**a**) Binary clustering in 2-dimensional space (*Left*). (**b**) Three classes clustering in 2-dimensional space (*Right*)

Table 2 Performance of Pang and Lee's three-classes classification

	Author a	Author b	Author c	Author d
Best accuracy	65 %	60 %	75 %	62 %
Baseline	40 %	38 %	48 %	42 %
Size	1770	902	1307	1027

For example, a document with a positive sentiment is theoretically closer to neutral documents than negative ones.

Different to Koppel and Schler [33], who applied a standard n-ary classifier or regression to this problem, Pang and Lee [10] consider this problem as a metric labeling [34] task. They applied their One-vs-All (ova) and Regression with a novel similarity measurement method positive-sentence percentage (PSP) on four authors' movie review data sets. Their best accuracy rates (see Table 2) for each authors' reviews were always yielded by ova + PSP. It should be noted that best accuracy fluctuates with the baseline. Consequently, we can assume there is a positive correlation between the accuracy rate and base line, and that the accuracy could lower if the data set is completely balance distributed (33.3 % for each class). Although it is hard to find a completely balanced distributed corpus in practice, we can assert that the accuracy tends to decrease when the data distribution tends to balance. This is a disadvantage of the current existing approach.

As mentioned, for our clustering based approach, it is not easy to obtain an acceptable result by directly conducting three centroids clustering. Hence, we assume a simple scenario where all documents only consist of two words: 'good' and 'bad', and that the sentiment of any document is determined by the ratio of 'good' and 'bad'. Thus, all documents can be expressed as a point in a two-dimensional space, as shown in Fig. 2(a). When the binary class clustering was conducted, most points were assigned to the correct centroid, and clearly distributed to the two sides of a theoretical bound. However, in most conditions, it is difficult to make sure the centroids move to the appropriate position in the three-group clustering (Fig. 2(b)). This is because the neutral group is usually in a non-globular shape which k-means algorithm is essentially not sensitive to [18].

Fortunately, the two centroids of binary clustering can be stably located at the end of a clustering process. We can draw support from the two stably-located centroids to indirectly identify the neutral group, because theoretically, the ratio of positive and negative content in a neutral document tends to balance. Therefore, ideally, they should be distributed between the two centroids. For example, in Fig. 3, the points between two slant border lines can almost be regarded as neutral. Based on this understanding, we designed two strategies to address this issue.

3.3.1 Modified voting mechanism

A voting mechanism was introduced in [7], which indicates the final group of a document is determined by the major group of multiple clustering processes running. We can modify the voting mechanism to fit the requirements of three classes sentiment analysis. It is well known that the positions of binary clustering centroids are only basically stable, but not absolutely. This can be verified by the fact that the accuracy rates of multiple running processes only basically remain level, but still float up and down. Therefore, we can easily imagine that some points (see Fig. 3, especially points

Algorithm 2 Voting Mechanism (for three classes)

```
Input: A corpus $C$ includes $n$ opinion express documents $\{d1, d2, \ldots, di, \ldots, dn\}$.
Output: For each document $di (1 \leq i \leq n)$ in $C$, assign a label mark $pi, pi \in \{positive, neutral, negative\}$
Begin
Run binary clustering process $N$ times; assign each document $di$ $N$ temporary labels ($POSi$ times of positive labels and $NEGi$ times of negative labels, $POS_i + NEG_i = N$).
Select a threshold $T$ $((N/2) < T < N)$
foreach document $di$
    if $POS_i > T$
        $pi = positive$
    else if $NEG_i > T$
        $pi = negative$
    else
        $pi = neutral$
endfor
end
```

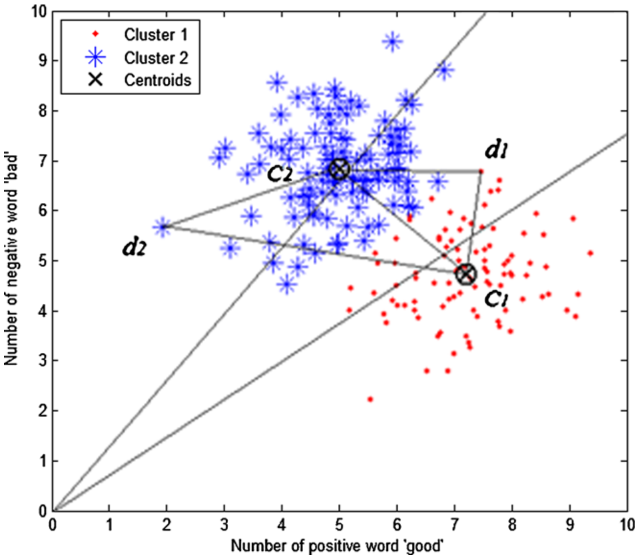


Fig. 3 Using centriods of binary clustering to support three classes sentiment analysis

between bound lines) fluctuate between two clusters in different clustering process running. Inversely, we can assume that the points which are frequently changing their group are neutral documents. Specifically, for each document di in a corpus, we conduct N $(N > 0)$ time clustering processes, and the number of times di was assigned to a positive or negative document was counted and denoted as $POSi$ or $NEGi$. A threshold T $((N/2) < T < N)$ was selected to determine the final group of di, which is denoted as pi.

Table 3 Best threshold values for five sets of test

N	5	10	15	20	25
T	4	8	13	18	22

A series of experiments were conducted to determine the relationship between the experiment number N and the threshold T. We first tested the modified voting mechanism for each author's corpus on both balanced and original data sets. To find a proper threshold value, we ran 5 sets of binary clustering processes with $N = \{5, 10, 15, 20, 25\}$. Table 3 records the best value of T for a different number of experiments N.

By applying regression, we established the linear relationship between N and T as

$$T = N - 2$$

This means that a document needs to assign to a positive or negative group more than $N - 2$ times to be grouped as non-neutral. The reason for the threshold to be high is because the system is currently very stable (standard division is usually lower than 1 %), which makes it unlikely for documents to shift between clusters. Setting a high threshold makes it more difficult to identify a non-neutral document, and easier to capture neutral documents.

3.3.2 *Distance measurement approach*

It is possible to identify neutral documents by measuring the distance between a document point and the centroids. For each point di, the distance to the two centroids can be measured and denoted as $dis(di, C1)$ and $dis(di, C2)$. Usually, the points between these two bound lines (e.g. $d1$) have mostly equal distance to the two centroids, meanwhile the points which are outside the bound lines (e.g. $d2$) have significantly different distances to the two centroids. For example, in Fig. 3, $C1$ and $C2$ are the two centroids after running a clustering process; $d1$ has more equal distances to these two centroids than $d2$. Hence, we can select an appropriate threshold T' to distinguish the final group pi of these points use the following Algorithm 3. If a document is non-neutral, we simply retain the origin group of the document.

4 Experiment of accuracy enhancement methods

4.1 Results of opposite contents processing

We firstly verify the performance of the opposite sentiment contents processing techniques by using the same data used in [7], which consists of 600 corpuses of 300 positive and 300 negative movie reviews, respectively. Opposite sentiment contents were captured by searching key words (6 transitional conjunctions and 8 negation words). As discussed

Algorithm 3 Algorithm of distance measurement approach (For three classes)

```
Input: A corpus $C$ includes $n$ opinion express documents $\{d1, d2, \ldots, di, \ldots, dn\}$.
Output: For each document $di(1 \le i \le n)$ in $C$, assign a label mark $pi, pi \in \{positive, neutral, negative\}$
Begin
    Run binary clustering process; assign each document di a temporary label $p'i \in (positive, negative)$. And obtained two centroids C1 and C2.
    Select a threshold $T'$ $(0 < T < 1)$
    foreach document $di$
    measure the distances between two centroids $dis(di, C1)$ and $dis(di, C2)$
    if $|dis(d_i, C_1) - dis(d_i, C_2)| > T'$
    $pi = p'i$
    else if $|dis(d_i, C_1) - dis(d_i, C_2)| > T'$
    $pi = neutral$
    endfor
end
```

Table 4 Results before and after applying opposite sentiment contents processing

	Highest	Lowest	Average	Standard deviation
Af	78.33 %	77.17 %	77.88 %	0.4 %
Ap	78.33 %	76.17 %	77.25 %	0.9 %
Af'	79.16 %	77.33	78.39 %	0.6 %
Ap'	79.33	77.83 %	78.58	0.5 %

in Sect. 3.2, this content is located after negation words, or in the negation section of a transitional sentence. Similar to previous work, adjectives and adverbs are extracted to construct the matrix by counting the frequency. The special step is for terms in opposite sentiment contents sections, and we count them as −1 instead of +1 to convert them to their opposite sentiment side. We conducted the exact same process for this matrix, which yielded the following results in Fig. 4, where Af and Ap are the ten times accuracies applied the technique before, and Af' and Ap' are the new ones.

Generally, the new lines in both Figures are slightly above the old ones, which shows that the processing of opposite contents contributed to the accuracy. Table 4 shows that the accuracy for highest, lowest and average all increased, though only by about 1 %. Meanwhile, the standard division values remain at a very low level.

4.2 Result of non-opinion contents processing

Next, we applied the subjectivity detection technique to our clustering system. The experiment was conducted on a new movie review data set (scale data set v1.0) which was provided by Pang and Lee. This data set contains two sub-data sets: a set of raw movie reviews and the subjective-extracted data for correspondent reviews. We intended to contrast the difference between the accuracies of these two sub-data sets. If the subjective-extracted data performs better than the raw data, the contribution of non-opinion contents processing (subjectivity extraction) is proven.

Table 5 Results before and after applying non-opinion contents processing

	Highest	Lowest	Average	Standard deviation
NAf	81.33 %	80.5 %	80.98 %	0.3 %
NAp	81.33 %	80.5 %	80.91 %	0.3 %
NAf'	89.17 %	87.17 %	88.7 %	0.5 %
NAp'	89.67 %	88 %	88.9 %	0.8 %

To avoid the effect of data set changing, we firstly verified the performance of the raw movie review data of scale data set v1.0. We randomly selected 600 reviews without considering the author, of which 300 review scores were greater than 0.7 (positive) and the other 300 review scores were lower than 0.4 (negative). We applied the same preprocessing steps on this data set. Then we obtained the final accuracies of the (denotes as NA_f and NA_p) around 81 % (see Fig. 5 and Table 5). This indicates the quality of this data set is slightly better than the previous one.

Next, we selected the same 600 reviews from the scale data set v1.0, which are subjectivity extracted, and applied our clustering process on them. The results (in Fig. 5 and Table 5) show a significant increase to above 88 %. Although this accuracy rate was obtained on a better quality data set, a 7 % improvement brought about by the subjective extraction technique is remarkable.

5 Experiment on three classes clustering

We use the same data as that used by Pang and Lee [10], which are 4 corpuses containing 4 authors' movie reviews. The data set has been preprocessed of subjective extraction and ranked by scores. Documents lower than 0.4 are regarded as negative, those greater than 0.7 are regarded as positive and the rest are regarded as neutral. The distribution of all 4 corpuses is unbalanced. The k-means algorithm is more appropriate to balance distributed data, since the centroid is easy to shift if the cluster sizes are different. To compare the performance between balanced and unbalanced data, we made another balance distributed data set by extracting equal numbers of positive, neutral and negative documents from each author's corpus.

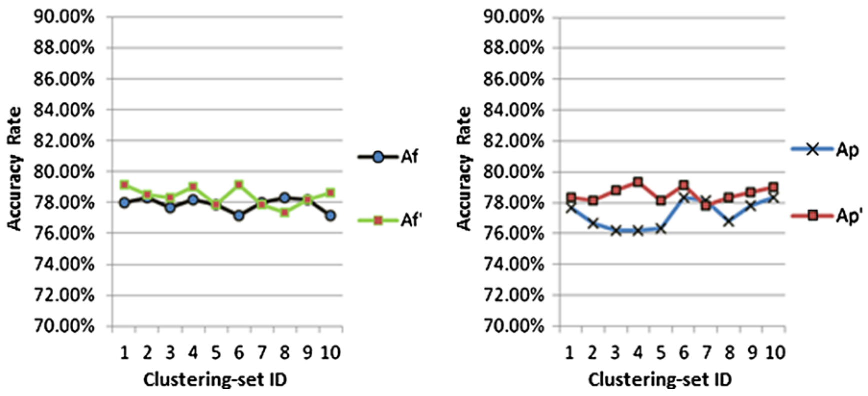


Fig. 4 (**a**) Comparison of accuracy of frequency of data before and after applying opposite contents processing (*Left*). (**b**) Comparison of accuracy of presence of data before and after applying opposite contents processing (*Right*)

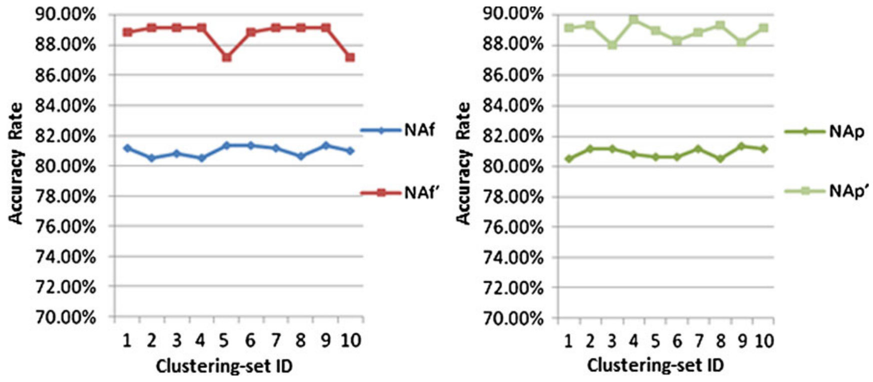


Fig. 5 (**a**) Accuracy comparison of frequency of data before and after applying non-opinion contents processing (*Left*). (**b**) Accuracy comparison of presence of data before and after applying non-opinion contents processing (*Right*)

5.1 Results of modified voting mechanism

We set $N = 20$ to test the accuracy of each corpus. The threshold T can be calculated as $T = N - 2 = 18$. The results (A_{voting}) are listed in Table 6, where B expresses balanced data and O expresses original data. For the purpose of clarity, we only list the average accuracy rates of ten results of frequency data.

Clearly, the clustering-based method is more suitable for a balanced distributed data set. The accuracy rates of balanced data are generally 7 % to 10 % higher than the unbalanced original data. For balanced data, the accuracy rates are basically at the level of 57 % to 60 %. It is believed that such an accuracy level is compatible to Pang and Lee's supervised learning approach [10], if making the base line 33.3 %.

5.2 Results of distance measurement approach

Lastly, we evaluated the performance of the distance measurement approach of the three classes analysis. Our clustering process adopts cosine distance, in which the distances between all documents and centroids are in the interval of [0, 1]. By testing each corpus, the average distance difference values of non-neutral documents (0.055 to 0.065) are always greater than the values of neutral documents (0.041 to 0.046). This outcome verified our previous assumption. Therefore, we selected an approximate intermediate value of $T' = 0.05$, as this value performs best accuracy in our test. This test was only applied on the balanced data set. The accuracy rates of this approach are also listed in Table 5 as $A_{distance}$. Compared with the voting mechanism, the accuracy rates all increased by about 3 % to over 61 %.

6 Result discussion

The experiment gave satisfactory results for both the two extension directions of the clustering-based sentiment analysis system.

Firstly, the accuracy rate of binary clustering reached a very high level. Both the opposite contents and non-sentiment contents process methods are proved to be effective, though the former only brought about 1 % improvement. The outcome derived by subjective contents extraction is more significant. Although 88 % accuracy was obtained on a better quality data set, the 7 % improvement displays the importance of data purity. This accuracy rate is almost at the same level as most of recent researches which applied more costly supervised learning approach, such as Pang and Lee [9] (85 %), Xia et al. [26] (85 %), Tan [27] (90 %) and Bai [28] (92 %). In other words, our high accuracy is obtained in a more efficient way with a lower cost. In addition, all results are very stable, which indicates the instability is-

Table 6 Results of three classes sentiment analysis

	Author a		Author b		Author c		Author d	
	B	O	B	O	B	O	B	O
Baseline	33.3 %	40 %	33.3 %	38 %	33.3 %	48 %	33.3 %	42 %
Size	600	1770	600	902	558	1307	513	1027
A_{voting}	58.83 %	50.71 %	57.5 %	49 %	60.22 %	53.76 %	60.42 %	50.34 %
$A_{distance}$	61.17 %	N/A	60.17 %	N/A	63.44 %	N/A	64.13 %	N/A

sue of k-means will not be re-energized by these new techniques.

Secondly, the results of the three classes sentiment analysis demonstrate that the clustering-based method is more suitable for balanced distributed data processing than unbalanced. For balanced data, most of the results are above 60 %, and are very stable. Additionally, in supervised learning, accuracy rates are reduced and the data distribution tends to be balanced. According to a simple calculation with the results shown in Table 2, it is possible for the accuracy rate for balanced data (baseline 33 %) to be far lower than 60 %. Therefore, we believe the clustering based three class sentiment analysis is more powerful on processing balance-distributed data than supervised learning approaches. Meanwhile, similar to binary sentiment analysis, the cost of our clustering based approach is significantly reduced. Finally, the distance-based method performs better than the modified voting mechanism. This fact seems to indicate that the distance measurement approach is more powerful for detecting the potential relationships among documents and more suitable for three classes sentiment analysis.

7 Conclusions and further research directions

The contribution of this paper is to apply the technique of opposite opinion contents processing and modified voting mechanism into clustering-based sentiment analysis approach to enhance its performance. Currently, it is able to produce higher accuracy in binary sentiment analysis by applying opposite opinion contents and non-opinion contents processing. And it is able to handle three classes sentiment analysis with an acceptable accuracy, especially for balance distributed data. The three classes analysis result is generated based on high accuracy binary analysis result. Two techniques modified voting mechanism and distance measurement approach contributed for this task. And the distance measurement approach usually performs better than the modified voting mechanism.

All these achievements are obtained on the basic clustering based sentiment analysis framework, which is comparatively more efficient than supervised machine learning approaches. Moreover, it does not rely on human participation and linguistic knowledge.

Furthermore, the research described in the paper proved that the clustering-based approach is extendable. There are still opportunities to improve its performance. For example, the k-means algorithm has its own limitation or data size, shape and balance.

Finding a good variety of k-means such as Canopy, Spectral Clustering [35] or other clustering algorithms to substitute basic k-means can be a further research direction. For example, Huang's extension to basic k-means with categorical values [36], an Expectation Maximization clustering algorithm introduced by Yang et al. [37] which believed to be helpful to improve the clustering accuracy.

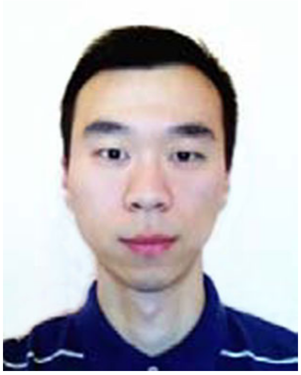
A newly developed clustering category which is different to both partitional and hierarchical clustering, called overlapping clustering [38, 39] is also a potential solution for sentiment analysis.

References

1. Hitlin PLR (2004) The use of online reputation and rating systems. In: Pew Internet & American Life Project Memo. doi:10.1016/j.dss.2005.05.019
2. Group ctK (2007) Online consumer-generated reviews have significant impact on offline purchase behavior. http://www.comscore.com/Press_Events/Press_Releases/2007/11/Online_Consumer_Reviews_Impact_Offline_Purchasing_Behavior
3. Pang B, Lee L (2008) Opinion mining and sentiment analysis. Found Trends Inf Retr 2(1–2):1–135. doi:10.1561/1500000011
4. Chiu C-M (2004) Towards a hypermedia-enabled and web-based data analysis framework. J Inf Sci 30(1):60. doi:10.1177/0165551504041679
5. Tang H, Tan S, Cheng X (2009) A survey on sentiment detection of reviews. Expert Syst Appl 36(7):10760–10773. doi:10.1016/j.eswa.2009.02.063
6. Boiy E, Hens P, Deschacht K, Moens M-F (2007) Automatic sentiment analysis in on-line text. In: International conference on electronic publishing pages, Vienna, Austria, pp 349–360
7. Li G, Liu F (2012) Application of a clustering method on sentiment analysis. J Inf Sci 38(2):127–139. doi:10.1177/0165551511432670
8. Pang B, Lee L, Vaithyanathan S (2002) Thumbs up?: sentiment classification using machine learning techniques. In: Conference on empirical methods in natural language processing (EMNLP), Philadelphia, Pennsylvania, USA, p 79. doi:10.3115/1118693.1118704
9. Pang B, Lee L (2004) A sentimental education: sentiment analysis using subjectivity summarization based on minimum cuts. In: Pro-

ceedings of the 42nd annual meeting on association for Computational Linguistics, Stroudsburg, PA, USA. Association for Computational Linguistics, p 271. doi:10.3115/1218955.1218990

10. Pang B, Lee L (2005) Seeing stars: exploiting class relationships for sentiment categorization with respect to rating scales. In: Proceedings of the 43rd annual meeting on association for computational linguistics. Association for Computational Linguistics, pp 115–124. doi:10.3115/1219840.1219855
11. Cesarano C, Dorr B, Picariello A, Reforgiato D, Sagoff A, Subrahmanian VS (2004) Oasys: an opinion analysis system. In: AAAI spring symposium on computational approaches to analyzing weblogs
12. Kamps J, Marx M, Mokken RJ, De Rijke M (2004) Using wordnet to measure semantic orientations of adjectives. Paper presented at the International conference on language resources and evaluation
13. Turney PD (2002) Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. In: 40th annual meeting of the association for computational linguistics (ACL), Philadelphia, Pennsylvania, USA, p 417. doi:10.3115/1073083.1073153
14. Salton G, Wong A, Yang CS (1975) A vector space model for automatic indexing. Commun ACM 18(11):613–620. doi:10.1145/361219.361220
15. Andrews NO, Fox EA (2007) Recent developments in document clustering. Computer Science, Virginia Tech, Tech Rep
16. Salton G, Buckley C (1988) Term-weighting approaches in automatic text retrieval∗1. Inf Process Manag 24(5):513–523
17. Al-Harbi S, Rayward-Smith V (2006) Adapting k-means for supervised clustering. Appl Intell 24(3):219–226
18. Tan PN, Steinbach M, Kumar V (2006) Introduction to data mining. Pearson Addison Wesley, Boston
19. Xia R, Zong C, Li S (2011) Ensemble of feature sets and classification algorithms for sentiment classification. Inf Sci 181(6):1138–1152
20. Tan S (2008) An improved centroid classifier for text categorization. Expert Syst Appl 35(1):279–285
21. Bai X (2011) Predicting consumer sentiments from online text. Decis Support Syst 50(4):732–742
22. Goldberg AB, Zhu X (2006) Seeing stars when there aren't many stars: graph-based semi-supervised learning for sentiment categorization. Association for Computational Linguistics, pp 45–52
23. Chaovalit P, Zhou L (2005) Movie review mining: a comparison between supervised and unsupervised classification approaches. Paper presented at the Proceedings of the 38th Hawaii international conference on system sciences
24. Shi K, Li L (2012) High performance genetic algorithm based text clustering using parts of speech and outlier elimination. Appl Intell:1–9
25. Laszlo M, Mukherjee S (2007) A genetic algorithm that exchanges neighboring centers for $\langle i\rangle k\langle /i\rangle$-means clustering. Pattern Recognit Lett 28(16):2359–2366
26. Poomagal S, Hamsapriya T (2011) Optimized k-means clustering with intelligent initial centroid selection for web search using URL and tag contents. In: Proceedings of the international conference on web intelligence, mining and semantics. ACM, New York, p 65
27. Menéndez H, Camacho D (2012) A genetic graph-based clustering algorithm. In: Intelligent data engineering and automated learning-IDEAL 2012. Springer, Berlin, pp 216–225
28. Hong T-P, Lin C-W, Yang K-T, Wang S-L (2012) Using TF-IDF to hide sensitive itemsets. Appl Intell:1–9
29. Manthey B, Röglin H (2009) Improved smoothed analysis of the k-means method. In: Society for industrial and applied mathematics, pp 461–470
30. Yu H, Hatzivassiloglou V (2003) Towards answering opinion questions: separating facts from opinions and identifying the polarity of opinion sentences. In: Conference on empirical methods in natural language processing, Stroudsburg, PA, USA. Association for Computational Linguistics, p 129. doi:10.3115/1119355.1119372
31. Hatzivassiloglou V, Klavans JL, Holcombe ML, Barzilay R, Kan MY, McKeown KR (2001) Simfinder: a flexible clustering tool for summarization. In: Citeseer
32. Larsen B, Aone C (1999) Fast and effective text mining using linear-time document clustering. In: KDD-99. ACM, New York, pp 16–22. doi:10.1145/312129.312186
33. Koppel M, Schler J (2006) The importance of neutral examples for learning sentiment. Comput Intell 22(2):100–109. doi:10.1111/j.1467-8640.2006.00276.x
34. Kleinberg J, Tardos E (1999) Approximation algorithms for classification problems with pairwise relationships: metric labeling and Markov random fields. In: IEEE, pp 14–23. doi:10.1109/SFFCS.1999.814572
35. Von Luxburg U (2007) A tutorial on spectral clustering. Stat Comput 17(4):395–416
36. Huang Z (1998) Extensions to the k-means algorithm for clustering large data sets with categorical values. Data Min Knowl Discov 2(3):283–304
37. Yang M-S, Lai C-Y, Lin C-Y (2012) A robust EM clustering algorithm for Gaussian mixture models. Pattern Recognit
38. Yokoyama S, Nakayama A, Okada A (2009) One-mode three-way overlapping cluster analysis. Comput Stat 24(1):165–179
39. Bello-Orgaz G, Menéndez HD, Camacho D (2012) Adaptive k-means algorithm for overlapped graph clustering. Int J Neural Syst 22(05)



Gang Li received the Bachelor of Information System degree (with Honours) from La Trobe University, Australia. He was awarded as the Australia Computer Society best final year student of 2007, and best Honours year student of 2008.

Gang is currently pursuing his Ph.D. degree at La Trobe University. He was awarded La Trobe University Postgraduate Research Scholarship for his Ph.D. study. His research interests include Machine Learning and Natural Language Processing.

Fei Liu received the Bachelor of Science in Mathematics degree from Zhejiang University, China and Masters of Computer Science and Ph.D. degree from La Trobe University, Australia. She was awarded La Trobe University Postgraduate Research Scholarship for her Ph.D. study.

Fei is currently a senior lecturer in the Department of Computer Science & Computer Engineering, La Trobe University. Previously, she worked as a lecturer in the University of South Australia and Royal Melbourne Institute of Technology. She also worked as a software engineer in Ericsson Australia. Her research areas include Automated Reasoning, Semantic Web and Text Mining.

Fei is a member of IEEE (Computer Society). She has authored/co-authored more than 50 conference and journal publications.