Article

Journal of Information Science
1–13
© The Author(s) 2012
Reprints and permission: sagepub.co.uk/journalsPermissions.nav
DOI: 10.1177/0165551511432670
jis.sagepub.com
SAGE

Application of a clustering method on sentiment analysis

Gang Li
La Trobe University, Australia

Fei Liu
La Trobe University, Australia

Abstract

This article introduces a novel approach for sentiment analysis – the clustering-based sentiment analysis approach. By applying a TF-IDF weighting method, a voting mechanism and importing term scores, an acceptable and stable clustering result can be obtained. The methodology has competitive advantages over the two existing types of approaches: symbolic techniques and supervised learning methods. It is a well-performed, efficient and non-human participating approach to solving sentiment analysis problems.

Keywords

sentiment analysis; opinion mining; clustering; semantic web

1. Introduction

The world wide web contains a huge number of documents which express opinions, containing comments, feedback, critiques, reviews and blogs, etc. These documents provide valuable information which can help people with their decision-making. For example, product reviews can help enterprises promote their products; comments on a policy can help politicians clarify their political strategy; event critiques can help the involved parties reflect on their activities, etc. However, the number of these types of documents is huge, so it is impossible for humans to read and analyse all of them. Thus, automatically analysing opinions expressed on various web platforms is increasingly important for effective decision-making [1]. The task of developing such a technique is sentiment analysis or opinion mining. It refers to a broad area of natural language processing, computational linguistics and text mining.

This task is challenging and needs to use symbolic statistical methods, because online reviews are often unstructured, subjective and difficult to digest in a short timeframe [2]. People express their opinions in different ways, and analysing the sentiment in a sentence is difficult using certain statistical approaches. On the other hand, sentiment analysis is very domain-specific when applying supervised learning classification. Many domains lack large amounts of labelled data for fully supervised learning approaches [3]. As a result, when transferred to a new domain without any labelled examples, a sentiment classifier often performs extremely badly [4].

In this article, we propose clustering-based sentiment analysis, which is a novel approach to sentiment analysis. This approach overcomes several drawbacks of currently existing methods (Section 2) relating to aspects of accuracy, effectiveness, human participation and domain transferability. Experiments of the approach are conducted in Section 4. Along with the experiments, numbers of technical challenges are exposed step-by-step. Three techniques for solving these problems – ‘TF-IDF weighting’, ‘voting mechanism’ and ‘enhancement by hybrid with scoring method’ – are, therefore, defined.

Corresponding author:
Fei Liu, Department of Computer Science and Computer Engineering, La Trobe University, Australia.
Email: F.Liu@latrobe.edu.au

2. Related work

2.1. Further definition of sentiment analysis

The task of sentiment analysis aims to predict the sentiment orientation (i.e. positive, negative or neutral) by analysing sentiment or opinion words and expressions in sentences and documents [5]. Consequently, it determines the attitude of a comment writer with respect to a particular topic. The attitude may be their judgement or evaluation, their affective state or the intended emotional communication.

There are two research directions in the area. The first is to classify a large number of opinions into bipolar orientations (positive or negative) [2]. This work was earlier pioneered by Pang et al. [6] and Turney et al. [7]. Another research direction is subjectivity/objectivity identification. This direction is commonly defined as classifying a given text into one of two classes: objective or subjective [8].

These two research directions can mutually influence each other, as in positive/negative polarity classification, filtering out the objective sentence seems be helpful [9]. Inversely, opinion-expressing words or phrases can be an indication of a subjective sentence [10]. The development of these two directions will help people to extract more valuable information from online opinion-expressing texts.

The research which will be introduced in this article commits to the direction of positive/negative polarity classification. The research can be developed at the word, phrase, sentence and document level. Our approach is suitable for document-level analysis.

2.2. Current approaches

As mentioned previously, there are two main academic streams on conducting sentiment analysis: symbolic techniques and supervised machine learning (classification) approaches [11].

Both symbolic techniques and supervised machine learning approaches require the step of pre-converting the raw documents into text vectors. Consequently, a collection of n documents with m unique terms can construct an $m*n$ term-document matrix [12]. This matrix is actually an application of the classic vector space model [13], which was first developed for automatic indexing.

2.2.1. Symbolic techniques. The basic idea of sentiment analysis based on symbolic techniques is to assign each term (feature) 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 score for every term is obtained, the score for the whole document can be calculated by applying aggregation functions. Usually, the function is an average or sum. Clearly, the core step of these kinds of techniques is the way to score terms.

Score by human subjects. The research of Cesarano et al. introduced the simplest scoring method [14], which is to ask a number of human subjects to give scores on opinion-expressing documents. Then, a pseudo-expected-value word scoring strategy, inspired by the concept of expected values in statistics [15], for example, is applied to establish the scored word bank. Another strategy, called pseudo-standard-deviation adjective scoring, was introduced in the same publication. It is believed there are several other strategies to obtain the scored word bank.

However, as all these functions are based on human intuition, they are not sufficiently reliable because of the influence of human subjects' educational and cultural backgrounds. Furthermore, systems that largely rely on human intervention are costly and therefore not suitable for processing a large volume of data.

Score by WordNet. WordNet is a lexical database for the English language [16]. It groups English words into sets of synonyms called synsets; it provides short, general definitions, and records the various semantic relations between the synonym sets. It was created and developed by Princeton University in 1985. It aims to create a combined dictionary and thesaurus that is more intuitively usable and which also supports text analysis and artificial intelligence applications.

The method of scoring words by WordNet was invented by Kamps and Marx [17]. Their work focuses on scoring adjectives only. As WordNet defines the relationship between synonyms, the similarity or distance between two words can be measured. The distance measure was defined using elementary notions from graph theory [18]. All adjectives in the WordNet database were collected, and can be incorporated into a graph. The nodes in the graph are words, and the edges express the synonym relationship. The distance $d\ (w_i, w_j)$ between two words w_i and w_j is the length of the shortest path between w_i and $w_{j.}$ If there is no path between them, the distance is infinite.

For the purpose of scoring adjectives, bipolar words 'good' and 'bad' were selected as the reference words to express positive and negative directions. Therefore, for each adjective w, the distances $d\ (w, good)$ and $d\ (w, bad)$ can be measured. It is believed that adjectives with a shorter distance to *good* are more positive and those which are closer to *bad* are more negative. An expression was given to formally define the score of word *w*:

$$EVA(w) = \frac{d(w, bad) - d(w, good)}{d(good, bad)} \tag{1}$$

Therefore, word w can obtain a score value in the interval $[-1, 1]$, where -1 expresses the 'bad' side and 1 expresses the 'good' side of the lexicon.

Similarly, the method can also be applied to measure adjectives on the dimensions of potency and activity by setting reference words as 'weak', 'strong', 'positive' or 'active'.

The accuracy of the author's experiment is around 70%. This research did not further apply aggregation functions to scoring documents.

Score by web search. The strategy of web search based scoring was introduced by Turney [7]. This strategy also requires distance measurement between words. To define the synonym relationship of words in the web document, Turney discovered and developed an approach to measure the similarity of words [19].

The approach was called Pointwise Mutual Information – Information Retrieval (PMI-IR). It works on the assumption that terms which co-occurred frequently tend to have a similar meaning. The Pointwise Mutual Information score between two words can be expressed as:

$$PMI(w1, w2) = \log_2\left(\frac{p(w1, w2)}{p(w1)p(w2)}\right) \tag{2}$$

where w_1 and w_2 are two words; $\mathrm{p}(w_i)$ is the probability of the occurrence of w_i (i = 1, 2) and $\mathrm{p}(w_1, w_2)$ is the probability that w_1 and w_2 co-occur. In real applications, the probability of terms is calculated by counting the hits (the number of documents retrieved) by the AltaVista Advanced Search Engine.

Similar to the last approach, two reference words 'excellent' and 'poor' were chosen. That is, the score of each term can be calculated by:

$$Score(term) = PMI(term, \text{'}excellent\text{'}) - PMI(term, \text{'}poor\text{'}) \tag{3}$$

Turney conducted the experiments on recognizing document sentiment with the help of term scores. The accuracy rate for four different topics was 84% (automobiles), 80% (banks), 65.83% (movies) and 70.53% (travel destinations). It should be pointed out that the experiment data is unbalanced distributed data. Some 59% of documents are positive, which means that always guessing the major class would yield an accuracy of $\geq$ 59%. It is predictable that the accuracy rate will decrease when adopting this method in a balanced data set.

2.2.2. Supervised machine learning approach. The best-known research using supervised machine learning on sentiment analysis was systematically conducted by Pang et al. In 2002, they introduced basic sentiment classification approaches on movie reviews [7] and presented a method for extracting objective sentences to improve the previous experiments in 2004 [9]. Then, in 2005, they undertook the work of assigning the sentiment value to documents using a three-point or four-point scale [20].

The experimental data consisted of 700 positive-sentiment and 700 negative-sentiment movie reviews. A negation processing technique [21] was firstly adopted on these documents to eliminate the confusions brought by negation of words.

Three classifiers were adopted to conduct the classification work: Naïve Bayes classification (NB), maximum entropy classification (ME) and the support vector machine (SVM). These classifiers have been validated to be effective in other text categorization studies, and all need pre-tagged training data.

For each algorithm, training results were tested by undertaking cross-validation. Different features were selected to process the experiment, including unigrams, bigrams and adjectives, etc. The accuracy level of these was around 80%, with the highest being 82.9% (unigrams' presence by SVM) and the lowest 72.8% (unigrams' frequency by SVM). The adjective data are also worth mentioning, although their accuracy rate is around 77%, which is comparatively low; it could be more efficient with low feature numbers (2633 features).

3. The clustering-based sentiment analysis approach

After examining the two previously mentioned approaches, we believe that there is room for improvement in both. A more effective solution to this issue needs to be created. The clustering-based approach is therefore defined in this section.

3.1. Drawbacks of current techniques

The symbolic techniques completely rely on the score of terms to generate the class of documents. The method of integrating the term's score seems too simple (average or sum), and experimental results show that it obtained relatively low accuracy rates.

On the other hand, although supervised learning approaches obtain good results, they are very costly. The large amount of training data needs to be pre-defined by classes manually. By simulating the training process previously discussed, we find the process is extremely time-consuming. Additionally, the supervised learning approach is highly dependent on the domain of training data. In other words, it lacks generality [2].

This, therefore, motivates us to find a new method which is capable of overcoming the drawbacks of the existing techniques. The new method should satisfy the following requirements:

- It performs well in terms of accuracy and stability.
- It does not require much human participation.
- It needs to be efficient and widely applicable.

3.2. Feasibility analysis of the clustering-based approach

The process of clustering aims to discover natural groupings, and thus presents an overview of the classes in a collection of documents [22]. The most widely used algorithm is basic *k*-means [23]. The algorithm can simply be expressed by the following pseudo code:

```
Select $k$ points as the initial centroids.
repeat
Form $k$ clusters by assigning all points to the closest centroid.
Recompute the centroid of each cluster.
until The centroids don't change
```

The algorithm does not need to pre-know the class of a document and does not need a training process. This means that it is free from human participation.

Based on the requirements mentioned in Section 3.1, clustering-related techniques are a possible solution to the sentiment analysis task. Actually, clustering-related techniques have already been used in similar studies by Hatzivassiloglou and McKeown [24], but they were only able to determine the semantic orientation of adjectives. As far as we know, no research applies clustering-related techniques to solving sentiment analysis at a document level. Possible reasons for this gap can be:

1. Clustering-related techniques cannot generate a model such as one trained in classification, which can be used to predicate the class of any new documents.
2. Clustering groups documents into two parts, but does not indicate which is positive and which is negative.
3. The accuracy can sometimes be relatively low.
4. Clustering results are unstable due to the random selection of centroids in *k*-means.

For the first reason, the concern is unnecessary, because in real applications it is meaningless to predicate a particular document's class, but the proportions of the class distribution are significant. If the document set is updated by adding more new documents, it is not a difficult task to cluster the whole document set again.

Thus, if the last three challenges can be resolved by experiments, we can regard a clustering-based approach as a feasible approach.

3.3. Overview of the clustering-based sentiment analysis approach

This approach is built based on the basic *k*-mean clustering algorithm. Documents are primarily clustered into two clusters which are expected to be a positive group and a negative group. The result is predictably poor in terms of accuracy and stability. In response, we have designed three strategies to solve these problems.

The technique of TF-IDF (term frequency – inverse document frequency) weighting [22] is firstly applied to the raw data to improve accuracy. Then, a voting mechanism will be used to extract a more stable clustering result. The result is obtained based on multiple implementations of the clustering process. Finally, the term score which was mentioned in the introduction to symbolic techniques will be imported to further enhance the clustering result.

The specifications of these strategies and a method of indicating the cluster meaning will be introduced, together with the experiments in the next section.

4. Experiments and analysis

In this section, the experiments to realize the clustering-based sentiment analysis ideas will be described. The final result and the sub-results will also be given to demonstrate the usefulness of each strategy.

4.1. Experimental data

We use the movie review data from Pang et al.'s [6] experiments. It is widely believed that sentiment analysis on movie reviews is more difficult than analysis on other domains [2, 5, 7]. This is because positive movie reviews often mention some unpleasant scenes and negative reviews often mention pleasant scenes. This can be verified by Turney's experimental outcomes (the accuracy rate of movie reviews is the lowest among all domains) which were presented at the end of Section 2.2.1.

This document set consists of 1000 positive and 1000 negative movie reviews. These review documents are pre-tagged based on their sentiment classes. For the same reason mentioned by Pang et al., which is to avoid slowness, we also did not put the whole data set into the experiment as a one-off step. Rather, we randomly chose 300 positive and 300 negative documents to construct a balanced distributed, experimental data set size of 600.

Multiple sets of experiments were conducted with different data sets, each with a size of 600 with a distribution of 300 positive/300 negative. No significant difference was found between these experiments. Therefore, it is believed that the experimental result will not be highly affected by different selections of documents.

4.2. Pre-processing and preliminary test

Directly using the documents set to conduct experiments will obtain a matrix with over 10,000 dimensions. It is inefficient to work in such a high dimensional vector space. Numbers of pre-processes were conducted for the purpose of dimension reduction.

Firstly, the work of stemming was done by applying Porter's algorithm [25]. It contributes to the task of dimension reduction, but only slightly.

As shown by Benamara et al. [26], adjectives and adverbs are good sentiment instructors. Thus, we attempted to extract adjectives and adverbs from documents. A part-of-speech tagger developed by Stanford University was used to tag the documents [27]. Words which were not tagged as being either an adjective or adverb were eliminated. Up to this step, only 6040 adjective and adverb terms remained for all these documents.

Next, we converted them into a vector space with 6040 dimensions in both frequency (counting the number of each feature in each document) and presence (only recording whether a feature is present in a document or not) forms. Before the clustering process, we validate these data by using SVM in the first place, obtaining 75.8% accuracy for frequency of data and 75% for presence of data. This accuracy rate is quite close to Pang's experimental result (75.1% accuracy by adjective in SVM). This indicates that our experiment data mode is essentially similar to Pang's data mode.

The clustering experiments were carried out with the same frequency and presence data. The class tags for each document were moved at first, because they are unnecessary in the *k*-means algorithm. We then applied the *k*-means algorithm with the MatLab ToolBox to cluster the documents into two groups.

In this research, the distance between documents is measured by cosine distance. This is usually used to measure the similarity of the two vectors by measuring the cosine of the angle between them. For our research, this distance measurement method is helpful in eliminating the effect of different document lengths.

Table 1. Confusion matrix

	Group 1	Group 2
Actual positive	*a*	*b*
Actual negative	*c*	*d*

Let **A** = (a_1, a_2, …, a_n) and **B** = (b_1, b_2, …, b_n) be two n-ary vectors and let θ be the angle between **A** and **B**. The cosine distance can easily be derived by using the Euclidean Dot Product formula:

$$A \cdot B = ||A||||B||\cos\theta \tag{4}$$

Then the cosine similarity can be further represented by:

$$\cos(\theta) = \frac{A \cdot B}{||A|| \cdot ||B||} = \frac{\sum_{i=1}^{n} a_i \times b_i}{\sqrt{\sum_{i=1}^{n} (a_i)^2} \times \sqrt{\sum_{i=1}^{n} (b_i)^2}} \tag{5}$$

The 600 documents were divided into two groups: group 1 and group 2. However, we did not know which group was positive and which was negative. Fortunately, we knew the actual class of each document, and therefore a confusion matrix could be constructed as shown in Table 1, where *a, b, c* and *d* are the number of documents in each cell. Thus, if $(a + d) > (b + c)$, group 1 will be regarded as the positive group, otherwise group 2 will be regarded as the positive group. Consequently, the accuracy can be calculated as

$$\text{Accuracy} = \begin{cases} \dfrac{a+d}{a+b+c+d} & \text{if} \quad (a+d) \geq (b+c) \\ \dfrac{b+c}{a+b+c+d} & \text{if} \quad (a+d) < (b+c) \end{cases} \tag{6}$$

Certainly, in the real application, the actual class of document is unknown, and the method of solving this problem will be detailed later. Throughout the remainder of this article, this is the method we use to judge the meaning of the clustered group and calculate the accuracy value.

As mentioned, the clustering results could be unstable. The accuracy rates which are obtained by any two clustering processes could be different. Therefore, it is necessary to set a larger sampling rate to evaluate the performance of the clustering analysis method. In this stage of the research, the clustering process was run 20 times to obtain 20 accuracy rates for both frequency and presence of data, denoted as $Af_n(n \in [1, 20])$ and $Ap_m(m \in [1, 20])$.

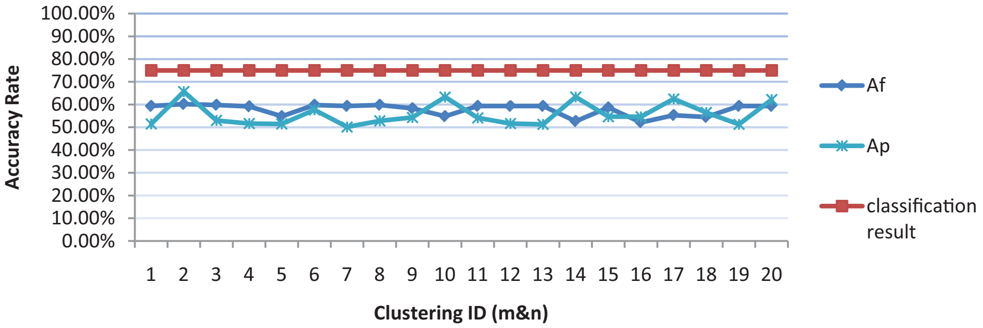

Figure 1. Results of the preliminary test (compared with the classification method).

Table 2. Results of the preliminary test

	Highest	Lowest	Average	Standard deviation
Af_n	60.17% (Af_2)	52.17% (Af_{16})	57.77%	2.6%
Ap_m	65.67% (Ap_2)	50.17% (Ap_7)	55.7%	4.9%

Intuitively, from Figure 1, we can see a very obvious gap between the clustering results (*Af* and *Ap*) and the supervised classification result (75%). The accuracy levels of the clustering results are very low. Moreover, different from the straight line of the classification result, the clustering results are represented by polygonal lines. Lines in such a shape verify the instability which we anticipated really exists.

As shown in Table 2, for both frequency and presence of data, the average accuracy rates are very low (57.77% and 55.7%), which is unacceptable. On the other hand, the standard deviation values of these are 2.6% and 4.9%, which represent the instability of the results. Moreover, the difference between highest and lowest values for both *Af* and *Ap* is greater than 15%. This means that the approach at this stage has the risk of producing very poor results.

In the next two subsections, two strategies will be introduced to solve the problems of low accuracy and the instability of our clustering approach.

4.3. TF-IDF weighting method

TF-IDF is a weighting approach often used in the area of text mining. It is used to evaluate the importance of a term in a document in a corpus. Term frequency in a given document is simply the number of times a given term appears in that document. Inverse document frequency is a measure of the general importance of the term. Mathematically, the TF-IDF weight of a term i can be expressed as:

$$W_i = tfi * \log(D/dfi) \tag{7}$$

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. Thus, $\log(D/df_i)$ is the inverse document frequency.

By applying this weighting method, the importance increases proportionally to the frequency of a term in a document but is offset by the frequency of the term in the corpus. It is believed that the weighting method has contributed to the improvement of clustering accuracy.

Experiments were conducted to validate the effect of the TF-IDF weighting method. Obviously, the weights of terms can only be obtained from the frequency data matrix, but once the weight vector is calculated, it can be applied to both frequency of data and presence of data.

We conducted experiments with TF-IDF weighting 20 times on frequency data and presence data, denoted as $Af'_n(n \in [1,20])$ and $Ap'_m(m \in [1,20])$, respectively. The experimental results up to this stage are illustrated in Figure 2. It is easy to see that the accuracy rate levels for both frequency and presence data are generally increased, but the fluctuations seem more severe.

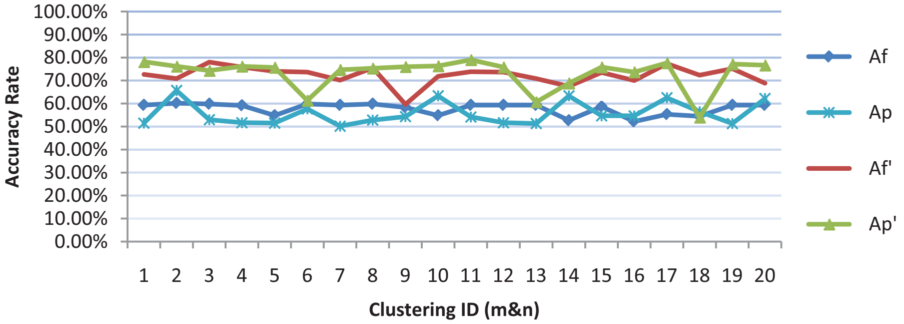

Figure 2. Results after applying TF-IDF weighting.

Table 3. Results after applying TF-IDF weighting

	Highest	Lowest	Average	Standard deviation
Af'_n	78% (Af'_3)	59.67% (Af'_9)	72.2%	4.02%
Ap'_m	79% (Ap'_{11})	54% (Ap'_{18})	73.1%	6.7%

By conducting statistical analysis on the results, it is believed that the TF-IDF weighting method effectively increased the average accuracy rate by more than 15%. Meanwhile, the standard deviations also become greater, which indicates that the results are more unstable (see Table 3).

4.4. Voting mechanism for stabilizing the results

To obtain more stable clustering results, a voting mechanism was designed. Under this mechanism, the final group of a document was not determined by any individual clustering process, but was voted by running results of clustering multiple times. The number of running times needs to be large enough to dilute the effect of outliers and instability. Meanwhile, it cannot be too large to obstruct the efficiency.

Here, we define a clustering set as running clustering 20 times (from tests, 20 times has been found to be large enough to provide stability, but may not necessarily be the smallest number; to find the optimal number, further experiments are required). In each clustering set, a document will obtain 20 votes V_j ($j \in [1, 20]$), and the value could either be positive or negative. Thus, the final voted result of document d could be expressed as:

$$f(d) = \begin{cases} \text{positive if } n(Vj = \text{pos}) \geq n(Vj = \text{neg}) \\ \text{negative if } n(Vj = \text{pos}) < n(Vj = \text{neg}) \end{cases} \tag{8}$$

Under this mechanism, 10 clustering sets were processed on both frequency of data and presence of data. The accuracy rates for the clustering sets were calculated based on the final voted result. From Figures 3 and 4, we can see that the final voted accuracy rates for both frequency of data (Af''_n $(n \in [1,10])$) and presence of data ($Ap''_m(m \in [1,10])$) are generally above 72%, which shows an improvement in accuracy. Meanwhile, in each clustering set, the accuracy rate of the final voted result is always higher than the average of the 20 accuracy rates values in the clustering set.

As Table 4 shows, the average accuracy rates for frequency of data and presence of data increase to 74.33% and 76.85%, respectively. The accuracy rates are comparable to those for the classification method. More gratifyingly, the standard deviation values drastically decrease, which indicates that the stability of the clustering based method is

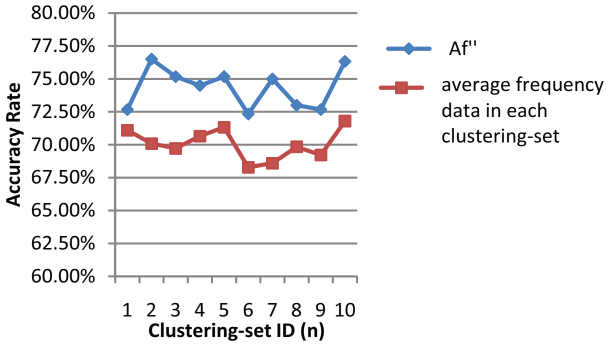

Figure 3. Voted and average results of frequency of data in clustering sets.

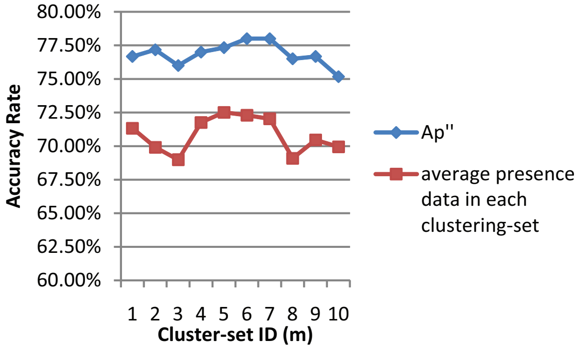

Figure 4. Voted and average results of presence of data in clustering sets.

Table 4. Results after applying the voting mechanism

	Highest	Lowest	Average	Standard deviation
Af''_n	76.5% (Af''_2)	72.33% (Af''_6)	74.33%	1.55%
Ap''_m	78% (Ap''_6)	75.1% (Ap''_{10})	76.85%	0.8%

improved by the voting mechanism. Additionally, up to this stage, the overall performance of the presence of data seems more competent than the frequency of data.

4.5. Hybrid system for sentiment analysis

As discussed, term scores can indicate the positive or negative tendency of documents within a certain range. We attempt to import term scores into our experiments to enhance the performance of our clustering-based method. From another point of view, our method can also be regarded as an improvement of the score-based symbolic techniques.

Firstly, we extract term scores from WordNet by the similar method previously discussed. In our 6040 feature terms, only 2751 terms can be connected to reference words ‘good’ and ‘bad’. The remaining 3289 terms are judged as neutral or non-opinion-expressing words. By directly calculating the average value of the score for each document, and then partitioning by medium value to judge the positive/negative direction, the accuracy is only 66.17%, which is fairly low.

Thus, we attempt to combine the term scores with our clustering-based method to build a hybrid system. Our aim is to create a weighting vector based on the term scores. For a term, the weight value is high if it is close to reference words; the weight value is low if it is far away from reference words. For terms which cannot connect to reference words (there were 3289 such words in this experiment), the weight is zero. Therefore, the dimensions of the data matrix are reduced from 6040 to 2751.

Hereby, regardless of the accuracy of this experiment, the large range of dimension reduction is a positive phenomenon. As described in [28], the complexity of k-means is ($n*k*I*d$), where n is the number of vectors, k is the number of clusters (centroids), I is the number of iterations and d is the number of dimensions (features). Therefore, the complexity is reduced to 2751/6040 of the origin. By testing, we can see that the execution time is reduced, as we expected.

Concretely, for all 2751 adjectives which can connect to the reference words, the maximum distance to the reference words is 11 steps. Let X be the distance of a word; the weight (W) of the word can be calculated as:

$$W = \begin{cases} 1.2 - (X - 1) * 0.02 & \text{if} \quad X \leq 8 \\ 1 - (X - 1) * 0.1 & \text{if} \quad 8 < X \leq 11 \end{cases} \tag{9}$$

This function is derived from experiments. It produces the best result in this research, but we cannot ensure this function is optimal.

As Figures 5 and 6 show, 10 weighted clustering sets were processed for both frequency of data (Af''') and presence of data (Ap'''). Clearly, for frequency of data, the weight is definitely helpful. The polygonal line tends to be straight, which indicates that the accuracy becomes highly stable. However, there seems no clear improvement for the presence of data.

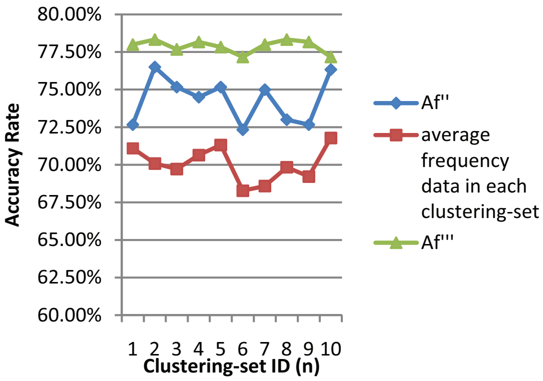

Figure 5. Frequency of data result after importing term score weighting.

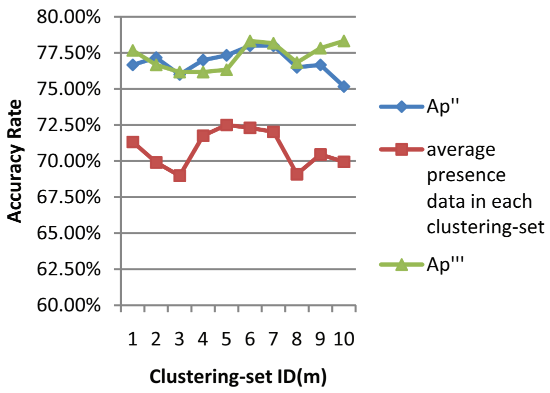

Figure 6. Presence of data result after importing term score weighting.

Table 5. Results after applying the term score weighting

	Highest	Lowest	Average	Standard deviation
Af'''_n	78.33% (Af'''_2)	77.17% (Af'''_6)	77.88%	0.4%
Ap'''_m	78.33% (Ap'''_6)	76.17% (Ap'''_3)	77.25%	0.9%

Statistically, the accuracy rates of both kinds of data exceeded 77% on average, providing evidence that the term scores can enhance the performance of the clustering analysis. Moreover, the standard deviation of the frequency of data is 0.4%, which is an acceptable degree of stability in real applications (see Table 5).

4.6. Judgement of group meaning

As mentioned in subsection B, in real applications it is impossible to utilize documents' actual class to judge the meaning of the clustered group. It is, therefore, necessary to find another strategy to make the judgement.

To achieve this, two seed documents, containing extremely positive or negative content, are added. The seed documents should be ensured to be always correctly clustered by our approach. Consequently, the group which contains the positive seed is the positive group; the group which contains the negative seed is the negative group.

The seed documents are separately built by the 25 collected positive or negative words. After the pre-processing and weighting steps, they are inserted into the original data set. After running a clustering set (20 times), we found these two seed documents were usually in the same group (six times). This indicates that these manually built seed documents are not competent for the task.

Fortunately, we find that over 150 documents in the 600 documents data set never fall into the wrong group in the clustering set. By enlarging the running times of a clustering set to 100, there are always 13 positive and nine negative documents correctly clustered. Although this does not mean that these 22 documents will be correctly grouped forever, they present strong positive and negative polarities which are very unlikely to be incorrectly grouped. Here, we define these documents as solid polarity documents.

Thus, we choose y positive and z negative solid polarity documents to build a positive seed set and a negative seed set. As we verified, the possibility of incorrectly grouping each solid polarity document is lower than 10^{-2}. Thus the meaning of clustered group can be decided by the major direction of the seed sets. This is because the possibility of the incorrect major direction is 10^{-y} and 10^{-z}, which is extremely low when y and z are sufficiently large.

4.7. The algorithm

Up to this stage, almost all the technical challenges we encountered in the experiments have been solved. A complete technical flow has been established for applying the clustering based approach to the sentiment analysis problem. This technical flow or algorithm is now presented as follows.

Complete algorithm of *Clustering-based Sentiment Analysis*

Input: A corpus C includes n opinion express documents $\{d_1, d_2 \ldots d_n\}$.

Output: For each document d_i ($1 \le I \le n$), assign a positive/negative polarity mark $p_{i,}$ p_i $\in${*positive, negative*}.

Method:

```
Begin
//pre-processing
foreach word $w_j$ ($1 \le j \le length(d_i)$) in each document $d_i$
    stemming
    assign part of speech tagging $t_j$
    if $t_j \ne adj$ and $t_j \ne adv$
        delete $w_j$
    endif
endfor

//converting documents to vectors
insert $m$ pre-selected and pre-processed seed documents to $C$, make new corpus $C_1$
    with $m + n$ documents
read and record all words appeared in corpus $C_1$ to a word list $L$ of length $l$
generate an empty matrix M
foreach document $d_i$ ($i \le n+m$) in $C_1$
    create vector $V_i$ ($V_i$ can be frequency vector $Vf_i$ (recommended) or presence vector $Vp_i$)
    insert $V_i$ to the end of matrix $M$
endfor

//weighting vectors
foreach word $w$ in L
    calculate TF-IDF weight of $w$ using equation (7), and store it on the weight list $WL_1$
    calculate the distance between $w$ and reference words using equation (9)
    calculate sentiment express weight, and save it on the weight list $WL_2$
endfor
generate an empty matrix $Mw$
foreach vector $V_i$ in $M$
    weighting vector: $Vw_i=(V_i.*WL_1).*WL_2$
    add $Vw_i$ to the end of matrix $Mw$
endfor

//Clustering processing
select the number of multiple running times $j$.
for $1 \le q \le j$
    run $k$-means clustering with cosine distance, cluster $Mw$ into two groups $G_1$ and $G_2$.
    if $(\sum_{i=1}^{m}(seed_i(positive)) \in G_1 > \sum_{i=1}^{m}(seed_i(positive)) \in G_2)$ and $(\sum_{i=1}^{m}(seed_i(negative)) \in G_1$
        $< \sum_{i=1}^{m}(seed_i(negative)) \in G_2)$
        $G_1 = positive, G_2 = negative$
    else if $(\sum_{i=1}^{m}(seed_i(positive)) \in G_1 < \sum_{i=1}^{m}(seed_i(positive)) \in G_2)$ and
        $(\sum_{i=1}^{m}(seed_i(negative)) \in G_1 > \sum_{i=1}^{m}(seed_i(negative)) \in G_2)$
        $G_1 = negative, G_2 = positive$
    endif
    record the clustering result of this run as $Rq$
endfor

//Voting
foreach document $d_i$ in corpus $C$
    foreach running result $Rq$ ($1 \le q \le j$)
        if $\sum( d_i(Rq) = positive) \ge \sum( d_i(Rq) = negative)$
            $p_i = positive$
        else
            $p_i = negative$
        endif
    endfor
endfor
```

4.8. Results and discussion

Based on the experiment results, we compare the performances of symbolic technique, the supervised machine learning method and our clustering-based approach (see Table 6).

Table 6. Evaluation of the three methods

	Accuracy	Efficiency	Human participation
Symbolic techniques	65.83% [7]	Very fast	Mostly no
Supervised learning	77%–82% [6]	Slow on training & fast on test	Yes
Clustering-based approach	77.17%–78.33%	Fast	No

Regarding accuracy, experiments conducted by Turney and Pang represent symbolic techniques and supervised learning methods, respectively. As mentioned previously, Turney conducted experiments on four different topics, but for our discussion only the movie reviews are applicable, because movie reviews contain a similar language feature style. The accuracy value is relatively low (65.83%), and it is based on imbalanced data. It is believed that accuracy is usually lower if it is applied to balanced data. Combined with our test on directly using the word score (accuracy 66.17%), we believe the accuracy rate of symbolic techniques is low when analysing movie reviews. Pang's experiment results convincingly show that supervised learning is the most powerful method on the aspect of accuracy. Our clustering-based approach still cannot perform as well as supervised learning at this stage, but the accuracy is at an acceptable level.

For efficiency, symbolic techniques work the best. Obtaining term scores is sometimes a one-off task; once the scores of terms are obtained, the score for documents can be calculated very quickly. Supervised learning methods require time to train, and for high dimensional data, the process is highly time-consuming. Our approach requires the same time to obtain the term score, but takes extra seconds to cluster data multiple times. But the time is usually much shorter than training a data set of equal size in supervised learning. After dimension reduction, it becomes even more efficient.

Additionally, both symbolic techniques and our clustering-based approach do not rely on human participation. This is a significant advantage over the supervised learning approach.

5. Conclusions and further research directions

To summarize, our clustering-based sentiment analysis approach overcomes the challenges of low accuracy and instability of results. It can produce sufficiently accurate clustering results in a short time without human participation. Compared with previous techniques, the performance of clustering-based sentiment analysis is the most balanced on the aspects of accuracy, efficiency and human participation. Thus, it is believed that this approach is more practical for real applications than other methods.

This approach is promising, but there are further challenges to face, including the following: the size of the document set may influence the outcome; finding a better way to obtain term scores (WordNet only generates term scores with around 70% accuracy). It is believed that better results can be obtained if these problems can be solved.

References

[1] Chiu C-M. Towards a hypermedia-enabled and web-based data analysis framework. *Journal of Information Science* 2004; 30: 60.

[2] Chaovalit P, Zhou L. Movie review mining: a comparison between supervised and unsupervised classification approaches. In: *Proceedings of the 38th Hawaii international conference on system sciences*, IEEE Computer Society, 2005.

[3] Aue A, Gamon M, Customizing sentiment classifiers to new domains: a case study. *Proceedings of recent advances in natural language processing (RANLP)* 2005, pp. 207–218.

[4] Tan S, Wu G, Tang H, Cheng X. A novel scheme for domain-transfer problem in the context of sentiment analysis. In: *Proceedings of sixteenth ACM conference on information and knowledge management (CIKM)* 2007, pp. 979–982.

[5] Thet TT, Na J-C, Khoo CSG. Aspect-based sentiment analysis of movie reviews on discussion boards. *Journal of Information Science* 2010; 36: 823.

[6] Pang B, Lee L, Vaithyanathan S. Thumbs up? Sentiment classification using machine learning techniques. In: *Conference on empirical methods in natural language processing (EMNLP)*. Philadelphia, Pennsylvania, USA, 2002, p. 79.

[7] Turney PD. 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, 2002, p. 417.

[8] Wiebe JM. Learning subjective adjectives from corpora. In: *Conference on artificial intelligence*, Menlo Park, CA. AAAI Press 2000, pp. 735–741.

[9] Pang B, Lee L. A sentimental education: sentiment analysis using subjectivity summarization based on minimum cuts. In: *Proceedings of the 42nd annual meeting on Association for Computational Linguistics*. Stroudsburg, PA: Association for Computational Linguistics, 2004, p. 271.

[10] Yu H, Hatzivassiloglou V. Towards answering opinion questions: separating facts from opinions and identifying the polarity of opinion sentences. *Conference on empirical methods in natural language processing*. Stroudsburg, PA: Association for Computational Linguistics, 2003, p. 129.

[11] Boiy E, Hens P, Deschacht K, Moens M-F. Automatic sentiment analysis in on-line text. In: *International conference on electronic publishing pages*, Vienna, Austria, 2007, pp. 349–360.

[12] Andrews NO, Fox EA. Recent developments in document clustering. *Computer Science, Virginia Tech, Tech Rep. 2007*.

[13] Salton G, Wong A, Yang CS. A vector space model for automatic indexing. *Communications of the ACM* 1975; 18: 613–620.

[14] Cesarano C, Dorr B, Picariello A, Reforgiato D, Sagoff A, Subrahmanian VS. Oasys: an opinion analysis system. In: *AAAI spring symposium on computational approaches to Analyzing Weblogs*, 2004.

[15] Ross S. *A first course in probability*. Prentice Hall, 1994.

[16] Miller GA, Beckwith R, Fellbaum C, Gross D, Miller KJ. Introduction to wordnet: an on-line lexical database. *International Journal of Lexicography* 1990; 3: 235.

[17] Kamps J, Marx M, Mokken RJ, De Rijke M. Using wordnet to measure semantic orientations of adjectives. In: *International conference on language resources and evaluation* 2004, p. 1115.

[18] Harary F. *Graph theory*. Addison-Wesley, Reading, MA, 1969.

[19] Turney P. Mining the web for synonyms: PMI-IR versus LSA on TOEFL. In: *European conference on machine learning*. Berlin: Springer, 2001, p. 491.

[20] Pang B, Lee L. Seeing stars: exploiting class relationships for sentiment categorization with respect to rating scales. In: *Proceedings of the 43rd annual meeting of the Association for Computational Linguistics*, 2005, pp. 115–124.

[21] Das SR, Chen MY. Yahoo! for Amazon: sentiment extraction from small talk on the web. *Management Science* 2007; 53: 1375–1388.

[22] Salton G, Buckley C. Term-weighting approaches in automatic text retrieval. *Information Processing & Management* 1988; 24: 513–523.

[23] Hartigan JA. *Clustering algorithms*. John Wiley & Sons, Inc., 1975.

[24] Hatzivassiloglou V, McKeown KR. Predicting the semantic orientation of adjectives. In: *35th annual meeting of the Association for Computational Linguistics and eighth conference of the European Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, 1997, pp. 174–181.

[25] Porter MF. An algorithm for suffix stripping. *Program: electronic library and information systems* 1993; 14(3): 130–137.

[26] Benamara F, Cesarano C, Picariello A, Reforgiato D, Subrahmanian V. Sentiment analysis: adjectives and adverbs are better than adjectives alone. *International conference web-logs and social media (ICwsm 07)*, 2007.

[27] Toutanova K, Manning CD. Enriching the knowledge sources used in a maximum entropy part-of-speech tagger. In: *2000 joint SIGDAT conference on empirical methods in natural language processing and very large corpora: held in conjunction with the 38th annual meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 2000, pp. 63–70.

[28] Manthey B, Rglin H. Improved smoothed analysis of the k-means method. In: *Twentieth annual ACM-SIAM symposium on discrete algorithms*. Society for Industrial and Applied Mathematics, 2009, pp. 461–470.