Journal of Intelligent & Fuzzy Systems 27 (2014) 2055–2065
DOI:10.3233/IFS-141171
IOS Press

Investigating association rules for sentiment classification of Web reviews

Yuan Man[a,b,*], Ouyang Yuanxin[c] and Sheng Hao[c]
[a]*China Huarong Asset Management CO. LTD., Beijing, China*
[b]*Research Center on Fictitious Economy & Data Science, Chinese Academy of Sciences, Univiersity of Chinese Academy of Sciences, Beijing, China*
[c]*School of Computer Science and Technology, Beihang University, Beijing, China*

Abstract. Sentiment Classification of web reviews or comments is an important and challenging task in Web Mining and Data Mining due to the increasing social media and e-commerce industry. This paper presents a novel approach using association rules for sentiment classification of web reviews. An optimal classification rule set is generated to abandon the redundant general rule with comparatively lower confidence. In the class label prediction procedure, we proposed a new metric named Maximum Term Weight (MTW) for the evaluation of rules and a multiple metric voting scheme to solve the problem when the covered rules are not adequately confident or not applicable. The final score of a test review depends on the overall contributions of four metrics. Experiments on multiple domain datasets from web site demonstrate that the voting strategy obtains improvements on other rule based algorithms. Another comparison to popular machine learning algorithms also indicates that the proposed method outperforms these strong benchmarks.

Keywords: Association rule, sentiment classification, text categorization

1. Introduction

Sentiment Classification, also referred as Polarity Classification or Binary Classification, aims to determine whether the semantic orientation of the given text is positive or negative. The rise of social media such as blogs and social networks as well as e-commerce suppliers such as Amazon, EBay and Taobao created great demand of sentiment analysis [1]. Automatic detection and analysis of consumer reviews or comments from Web is beneficial to market research, customer management, business intelligence, and recommendation system [2]. For example, when consumers are browsing on the e-commerce website, critical or praiseful reviews or comments from other consumers greatly influence their purchase decisions. Because it is becoming a common practice for a consumer to learn the reputation of a product or how others like or dislike it before buying. The feedback analysis is also important for researchers of recommendation systems to improve the recommending accuracy, and for product manufacturers to keep track of customer opinions on their products to improve user satisfaction [3]. As the number of Web reviews for any product (movies, e-commerce, social network content etc.) grows rapidly, it is hard for a potential consumer to make informed decision when reading hundreds of reviews on a single product. For manufactures or online shopping sites, it also requires more and more effort to manage and keep track of the large scale review dataset [4]. So, sentiment classification is becoming a challenging and interesting topic in text mining area.

Effective sentiment classification relies on multiple disciplines, such as machine learning, natural language processing, linguistic, statistic etc. One of the main

[*]Corresponding author. Yuan Man, China Huarong Asset Management CO. LTD., Beijing, China. Tel./Fax: +86 010 59618543; E-mail: yuanman@chamc.com.cn.

methodologies for sentiment classification is to treat sentiment classification as a special case of Text Categorization [5], which is a well studied field in the last few decades. Text categorization refers to assigning unlabelled natural language documents into a predefined set of semantic categories. Compared with text categorization, the difference in sentiment classification is that the predefined labels only include "positive", "negative" and sometimes "neural", rather than the content topics such as "sports", "entertainment" and "politics". A large number of machine learning and probabilistic based methods have been proposed for text categorization. The most popular methods include Naïve Bayes, decision tree, decision rules, association rules, Rocchi, neural networks, and support vector machines (SVM) [6].

Comparative studies demonstrated that these general techniques provide strong baseline accuracy for sentiment classification and outperforms other methods based on lexicon analysis [7]. However, it is not as good as the performance in general text categorization problems because web review has its own language characteristic. For example, sentiment is often expressed in more subtle and indirect ways: for a given review, sometimes only s small part of the sentences have sentiment orientation, and sometimes both positive and negative words are used in the same sentence, regardless of the rating score. Another reason is in most of the Text Categorization methods, the representation of documents is based on vector space model (VSM) proposed by Salton in 1975 [6]. Since the "bag of words" method of VSM ignores the combinations and orders of terms, one of the drawbacks of this strategy is that it is not sufficient to preserve the semantic and syntactic information.

Association rule based classifiers (associative classifier) originate from association rule mining task in Data Mining. A classification rule is a special association rule which directs to class label. Association rules are derived from frequent items sets. And a frequent item set is a set of items in which the co-occurrence of these items is more than a threshold value called minimum support. Since association rules reflect strong associations between items and includes more underlying semantic and contextual information than individual word, it has been developed within the text mining domain in different aspects [8–10]. Although associative classification has been well studied in text categorization, there is still a lack of applying association rules on sentiment classification problems.

In this paper, we investigate the association rules in sentiment classification problem. The motivation is to convert this general classification approach into a special domain and binary classification problem. The main contributions of this work involve: First, we build an optimal rule set to abandon the redundant rules to construct the associative classifier; Second, we propose a new metric named Maximum Term Weight (MTW) for the evaluation of rules; Third, in rules matching phase, a multiple metric voting scheme is used to solve the problem when matched rules are not applicable or not confident enough.

The rest of this paper is organized as follows. Section 2 gives a brief review of previous works on sentiment classification and associative text categorization. Section 3 describes the proposed method in details. Extensive experimental evaluation on real text data are discussed in Section 4. Section 5 concludes the paper and presents some directions for future work.

2. Related works

2.1. Sentiment classification

Sentiment classification outputs the judgment about whether people like or dislike a product from their reviews or comments recorded in natural language. One of the most popular approaches focused on applying traditional machine learning methods for text categorization on sentiment classification problem. These approaches convert the sentiment analysis problem into a text categorization task in which the predefined labels to assign are "positive" and "negative". Pang [7] firstly tried to classify movie reviews into positive/negative, using several supervised machine learning methods: Naive Bayes, Maximum Entropy and SVM. They tested different feature combinations and the results showed that SVM combined with unigrams outperformed methods based on human-tagged features. In their following work [11], they added in subjectivity detection with minimum cuts algorithm to avoid the sentiment classifier from dealing with irrelevant "objective" sentences. They trimmed out such objective content from movie reviews and used SVM classifier to determine the sentiment polarity of the test reviews. As reported in their work, the classification performance of product reviews is worse than that of normal topical text categorization. One of the main difficulties is that people typically use both positive and negative words in the same review, regardless of the rating score. In analyzing political speeches [12] exploited the argument structure found in speaker reference links to help determine how the members of congress would vote given their congressional floor speeches. The method in [13] also proved that standard machine learning techniques outperform lexicon and rule based approaches. They used bag-of-words (BOW), Part-Of-Speech (POS) information and sentence position as features for analyzing reviews, representing reviews as feature vectors for classifiers such as Naïve Bayes and SVM. But these feature extraction methods also depend on tools like POS Tagger. In [14], the problem of attributing a numerical score (one to five stars) to a review is presented. They use the feature representations of reviews and described it as a multi-label classification problem and present two approaches, using Naïve Bayes and SVM [15]. Compares SVM and ANN (Artificial Neural Networks) for document-level sentiment classification. The experiment results indicated that ANN outperformed SVM on movie review data, but the training time of ANN is too long. Liu et al. [16] proposed an adaptive multi-class SVM model for sentiment classification of tweets. The initial common sentiment classifier is transferred to a topic-adaptive one by optimization, unlabeled data selection and adaptive feature expansion. In [17], a hybrid feature selection method was proposed using RST and Information Gain for sentiment classification problem. SVM and Naïve Bayes was used for classification and the result shows hybrid feature selection method can obtain better results with less number of features.

Other method for sentiment classification mostly use natural language processing tools and linguistic approaches for word/phrase level analysis [18–22] or subjectivity/objectivity detection [23–25], but these methods highly rely on linguistic tools and human knowledge and didn't show obvious advantage.

2.2. *Association rules in text mining and sentiment classification*

Association rule mining is a fundamental task in data mining [9]. The application of association rules in text mining mainly involves two aspects. The first one is using frequent term sets as frequent patterns or features. Because frequent item set reflects strong associations between items, it is naturally expected to contain more underlying semantic and contextual meaning than individual word. Frequent patterns have been explored and proved to be helpful to obtain competitive performance for text categorization and clustering. Frequent patterns in text mining issues [10] can be frequent sequences or frequent itemsets, the difference lies in if the sequential orders of words are considered [26]. Analyzed the frequent patterns for text classification problem and proposed a strategy to set *min_sup* by establishing a connection with feature selection approach. Ahone [27] proposed the first algorithm to find maximal frequent sequence for text document. A maximal frequent sequence (MFS) is a sequence that is not contained or subsequence in other frequent sequence. Consequently, the collection of MFS's can be a compact representation for the original term set. In [28], Edith H. applied MFS in text clustering where each MFS of words correspond to a feature of text document in vector space model (VSM). Then k-means algorithm is employed to group document into clusters. Other text clustering methods based on frequent patterns involve MC [29], CFWS [30] and FTC [31]. Instead of using frequent patterns for text representation, these methods adopted frequent sequence or itemsets in the clustering phase.

Another application is associative classifier which construct classification model with association rules in which the consequent part is the class label [32]. Firstly introduced association rule mining technique into classification problem in which the classifier is built on a subset of association rules called "class association rules" [33]. Then developed this approach in a more specific way and applied it in classifying text documents [8]. Discussed the problem of mining association rules form textual document. More research on association rules in text mining are reported in survey [9]. Based on the existing related works, it can be concluded that although associative classification is a well studied method in text categorization, there is still a lack of investigating association rules on sentiment classification problems.

3. Proposed method

This section describes a new approach for sentiment classification using association rules. The proposed method includes four steps: (1) Data pre-processing and feature selection; (2) Frequent term set extraction and rule mining; (3) Mining optimal classification rules; (4) Predicting test review with multiple metric voting.

3.1. *Data pre-processing and feature selection*

Data pre-processing and feature selection is an essential procedure for most of text processing issues because of the high dimensionality of the natural language text which makes the text data quite noisy and sparse in vector space. Main approaches for text data pre-processing involve stemming which reduce term variations to a single representation and stop-word removal which discard common terms like prepositions and articles. Feature selection is one of the dimension reduction methods that can significantly decrease the computational cost of text categorization and, at the same time, preserve or even increase the classification performance. In this paper, Information Gain (IG) is implemented for feature selection. IG has been proved to be one of the best feature selection methods for text categorization [34]. It measures the entropy decrease of the corpus between the feature is present or absent. Let $\{\mathrm{C_i}\}_{i=1}^{m}$ denote the set of categories, the IG of term t is defined as:

$$IG(t) = -\sum\nolimits_{i=1}^{m} P(c_i)\log P(c_i) + P(t)\sum\nolimits_{i=1}^{m} P(c_i|t)\log P(c_i|t) + P(\bar{t})\sum\nolimits_{i=1}^{m} P(c_i|\bar{t})\log P(c_i|\bar{t}) \quad (1)$$

Before extracting frequent term set, the original terms are selected according to its IG score. In this study, the magnitude of dimensions is reduced from 10^5 to 10^3. Another important reason that makes feature selection essential is that the number of single terms greatly influences the scale of frequent term sets as well as the rule set. When the single terms increases, the number of frequent terms sets grows exponentially and sometimes unreachable. So, the base number of single terms must be restricted to a reasonable scale.

3.2. Frequent term set extraction

Efficient mining of frequent itemsets is a fundamental problem for mining association rules. The original description of association rule mining [35] is as follows: Let $I = \{i_1, i_2, \ldots, i_n\}$ be a set of items and $T = \{t_1, t_2, \ldots, t_m\}$ be a set of transactions in which each transaction contains a subset of the items in I. If itemset $X \subseteq I$, the number of transactions in T that contain X is *Count*(X) and the total number of transactions is n, then the support of X is *Sup*(X) = *Count*(X)/m. An itemset X is called frequent if its support is greater than or equal to a given percentage s which is so called the minimum support (*min-sup*).

In text mining applications, each document d in $D = d_1, d_2, \ldots d_m$ is treated as a transaction and the set of terms $T = t_1, t_2 \ldots t_m$ contained in D corresponds to the items set I. A term set S in T is frequent if $Sup(S) \geq min\text{-}sup$. The *min-sup* constraint of term set is a key measure for frequent sets extraction because it determines the scale and quality of the selected frequent sets. This deserves more consideration for text document because comparing with the classical market basket analysis, the amount of terms in a document is usually much larger than the items in a transaction. The large number of terms will sometimes lead to the exponential growth of frequent term set candidates and make the result unreachable.

When applied to text mining problem, the concept of support count corresponds to Document Frequency (DF). It can be deduced that the support count of a term set is the minimum DF of all the terms in set. However, in classification task, support count cannot be simply substituted by DF because DF only measures the occurrences and this is not sufficient to differentiate the discriminative effect of the frequent term sets. To solve this problem, our previous work [36] proposed a new metric Average Deviation Support (*AD-Sup*), considering the distribution discrepancy of term sets in each document class. Assume the documents set have n classes $\mathrm{class_1}, \ldots\ \mathrm{class_i}, \ldots\ \mathrm{class_n}$ and let *FS* denote the term set and t is the term in *FS*, *AD-Sup* can be formulated as :

$$AD\text{-}Sup(FS) = \frac{\sqrt{\sum\limits_{i=1}^{n} \{Sup(FS)_i - Ave(Sup(FS))\}^2}}{Ave(Sup(FS))} \quad (2)$$

$$Ave(Sup(FS)) = \frac{\sum\limits_{i=1}^{n} Sup(FS)_i}{n} \quad (3)$$

$$Sup(FS)_i = \min\{df(t)_1 \ldots df(t)_m\} \quad (4)$$

The expression of *AD-Sup* in Equation (1) can be deemed as a modified support deviation, where *Sup(FS)*$_i$ means the local support of *FS* in class *i* and *Ave(Sup(FS))* denotes the average value of *Sup(FS)* in all the classes. Additionally, analysis on real data shows that when a term set has a large average support, even if it's distributed quite evenly, sometimes its standard support deviation may still surpass that of the term sets which occur comparatively less but are more distinctive in different classes. However, when a term set has very close support value in different classes, it would not be a valuable feature for classification. Hence in the *AD-*

Sup equation, the standard deviation is divided by the average support to represent the deviation rate instead of the absolute deviation value.

The frequent term extraction procedure is implemented using Apriori strategy [35]. Apriori is one of the best known methods for association rule mining and as a breadth-first-search algorithm, it generates itemsets in a level-wise manner, where each candidate *k*-itemsets in the kth iteration is generated from frequent (*k−1*)-itemsets. In each iteration, the candidates with *support* $\geq min\text{-}sup$ are added into the frequent set until the candidate set is empty. Apriori algorithm is chosen as the extraction method for the characteristic that in Apriori, the algorithm works by scanning the database iteratively and transactions are not stored in memory. This strategy makes Apriori very suitable for the large count of transactions and items in text documents. After obtaining all the frequent term sets (FS), AD-Sup restraint is used to refine the frequent features. The selected FS will involve more term sets that are not only frequent but distributed unevenly in different classes.

3.3. Optimal rule mining

Following the extraction of frequent term sets, an association rule is an implication denoted by $X \Rightarrow Y$, where X and Y are two subsets from a frequent set $Z = X \cup Y$. A rule is considered as "confident" if its confidence restraint $Conf(X \Rightarrow Y) = Count\,(X \cup Y)\,/Count(Y)$ is not less than the threshold value. For classification problem, the consequent of the rules are class labels. The traditional definition of association rule is widely accepted for its simplicity and pruning effectiveness. However, it suffers the following problems: (1) the confidence and support restraint is not always suitable for any mining problem, an adaptation must be taken when association rules are applied in a specific task; (2) the number of association rules are usually too large which makes the pruning quite challenging; (3) the vast amount of association rules involve a lot of redundant information. To overcome these obstacles, many interesting metrics and pruning strategies have been proposed to find "optimal rules". There is no standard definition for "optimal rules". This paper utilizes a similar strategy close to [38]:

Definition (Optimal Association Rule Set): A rule set is optimal with respect to an interestingness metric if it contains all rules except those with no greater interestingness than one of its more general rules. Given two rules $P \Rightarrow C$ and $Q \Rightarrow C$ where $P \subset Q$, the latter is more specific than the former and the former is more general than the latter. Figure 1 is an example of optimal rule set generation where the interestingness measure of rules is confidence. As is shown in Fig. 1, $(A,\ C) \Rightarrow\ Z$, $(B,\ C) \Rightarrow Z$, and $(A,\ B,\ C) \Rightarrow Z$ are pruned because their confidences are smaller than those more general ones.

A theorem that an optimal rule set is a subset of a non-redundant rule set was proved theoretically in [38]. The optimal rules pruning makes use of both support and closure pruning and can greatly eliminate the redundant rules. It can be expected to improve the associative classification on computation complexity and effectiveness which we will discuss in Section 4.

3.4. Predicting sentiment class by association rules

For general association rule based classification, the classifier is a collection of selected rules. The method of predicting class labels by rules can be categorized into two main groups: the first one makes the prediction by a maximum likelihood strategy of single rule; the second one makes use of multiple rules to generate a score by the interestingness metric and the correlations among the rules. However, for sentiment classification, both of the above strategies may fail to predict correctly in some cases. Given covered rule sets with positive and negative classification rules, following examples are hard to predict the sentiment: First, the covered positive rules have the highest confidence but the difference with that of the negative rules is quite small, while the number of negative rules is much more than that of the positive rules. In this case, the test review should be negative rules but it will be predicted to be positive. Second, the number of covered rules and the max-confidence of rules are both equal. Third, the situation can be more complicated, where the negative rules have a higher confidence with a little priority than the positive but besides the highest confidence rules, comparing other covered rules, the positive rules are more persuasive.

To overcome these obstacles, borrowing the idea of democratic regime, we propose a new voting scheme using the following metrics for class prediction. The class label of a test review will be determined by a combination of voting score defined as Equations (5) and (6).

$$Score(test\_review_i) = \sum_{0}^{m} Vote(metric_j) \qquad (5)$$

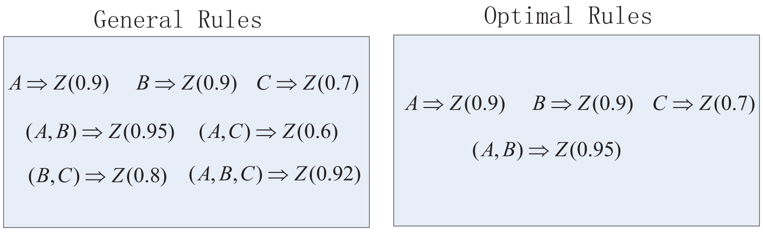
General Rules
$A \Rightarrow Z(0.9)$ $B \Rightarrow Z(0.9)$ $C \Rightarrow Z(0.7)$
$(A,B) \Rightarrow Z(0.95)$ $(A,C) \Rightarrow Z(0.6)$
$(B,C) \Rightarrow Z(0.8)$ $(A,B,C) \Rightarrow Z(0.92)$
Optimal Rules
$A \Rightarrow Z(0.9)$ $B \Rightarrow Z(0.9)$ $C \Rightarrow Z(0.7)$
$(A,B) \Rightarrow Z(0.95)$

Fig. 1. Example of optimal rule set

$$Vote(metric_j) = \{\ 1,\ -1,\ 0\} \qquad (6)$$

The vote on a metric is 1, −1, or 0 depending on the different metric value on covered positive rules (PR) and negative rules (NR). $Vote(metirc_j) = 1$ if *metric* $_j$*(PR)*>metirc$_j$*(NR)*, and respectively, if *metric*$_j$*(PR)* <metirc$_j$(NR), $Vote(metirc_j) = -1$, if *metric*$_j$*(PR)*= *metirc*$_j$*(NR)*, $Vote(metirc_j) = 0$. The assigned class label depends on whether vote score of metrics is a positive number or negative number

The metrics that are evaluated here include:

Definition 1 (Max-conf): the highest confidence value of covered rules.

Definition 2 (Cover-len): the number of rules in the covered rule set.

Definition 3 (Minor-conf): the average confidence of covered rules excluding the highest one.

All the above 3 metrics can be obtained by the recorded value generated in the rule mining procedure. However, sometimes there is no promising rule with a high confidence in the covered rules. The threshold of confidence cannot be too high because that will make the rule set too small and highly increase the unpredictable cases which are not covered by any rules. Consequently, the threshold is usually around 50%. For example, when the Max-conf is 50% or the Max-conf in PR and NR are very close, it is hard to predict if a test review belongs to either category. To solve this problem, we propose a new metric named MTW (Max Term Weight):

Definition 4 (MTW): Maximum Term Weight, the average term weight of each rule clusters in covered rules. In this paper, we use the information gain (IG) of each term for its weight measure. A rule cluster is a collection of rules which contains the same term. The motivation of MTW is to make use of discriminative measurement of single terms contained in the covered rules. Since MTW is extracted from rules clusters, both the maximum and average term weight are considered without redundancy. The measurement of single terms is basis for feature selection which reflects the importance of terms for representing different classes. Besides IG, similar measurement like TF*IDF and χ^2 are also applicable here.

Table 1
Examples of sentiment prediction by metric votes

Test review	Max-conf	Cover-len	Minor-conf	MTW	Score
Review 1	1	1	0	1	3:Pos
Review 2	−1	−1	−1	1	−2:Neg

Given a covered rule set, the algorithm to get MTW can be described as follows:

```
Algorithm 1: MTW metric generation
Input: single term set (TS) in descending order
        For each term $T_i$ in $TS$:
        If covered rule set is not empty
          For each rule in covered rule set:
               If($Rule_j$ contains $T_{i)}$
               Add $Rule_j$ to rule cluster($RC_i$);
               Set: $weight(RC_i) = GetTermWeight(T_i)$;
               Delete $Rule_j$ in covered rule set;
       end If
         end For
     end If
       end For
Return: $Average(\sum_i^k Weight\,(RC_i))$
```

Table 1 illustrates two examples of sentiment prediction on test reviews. The assigned class label depends on whether vote score of metrics is a positive number or negative number.

4. Experiments

4.1. Dataset

Multi-Domain Sentiment Dataset [39] contains reviews of several product types (domains) taken from

Table 2
Number of extracted frequent term sets

Dataset	Number of double-term sets	Number of (>2) term sets	Total number
Book	1875	7448	9323
DVD	1934	9764	11698
Kitchen	3214	9652	12866
Electronic	2076	9841	11917

Amazon.com. Each review consists of a rating (0–5 stars), a reviewer name, a product name, a review title, and the review text. Reviews with rating >3 were labeled positive, those with rating <3 were labeled negative, and the rest discarded because of their ambiguous polarity. Four domains of this dataset: DVD, Book, Kitchen and Electronic are selected in this experiment. Each domain contains 1000 positive and 1000 negative reviews.

In this paper, all the above datasets are pre-processed by stemming and stop-word elimination. The evaluation is conducted through 3-folds cross validation on the training set, which equally splits the training corpus into 3 folds and, for each time, uses two folds as the training set and the left fold as the test set.

4.2. *Frequent term sets extraction*

The frequent term sets extraction starts from the selected single terms by IG. The first scan generates frequent term sets with two terms and these double term sets are used as the input term sets for the candidate-generating algorithm. Then the iteration starts until no candidate is selected to be frequent term set. During the extraction procedure, two parameters are very important to obtain high quality frequent features for classification: the number of single terms and *min-sup*. Single terms are the fundamental source for frequent term sets; therefore, a proper number of single terms must be set to control the quantity of input single terms. Because too few terms are not sufficient for extracting enough frequent features, while too many terms will bring low weight terms and redundant information, as well as a too large collection of frequent term sets. Similarly, *min-sup* is the threshold for the iterating extraction steps to guarantee all the selected term sets are frequent enough to be statistically meaningful. In this paper, input number of single terms is set to be 600 and *min-sup* is 2%.

Table 2 shows the extraction results on the four datasets by *min-sup*, including number of double or 2-term sets, number of (>2) term sets, and total number of all the frequent term sets.

Tables 3 and 4 report more details of selected terms and extraction results. The top single terms are ranked by its IG value, followed by the frequent term sets by support count and refined term sets by AD-Sup. Note that the original text have been processed by stemming algorithm to bring variant forms of words together. Simultaneously, it also changes the form of the words and makes the stemmed terms appear to be different from the real words. Comparing the top frequent term sets ranked by support count and refined sets by AD-Sup restraint, most of the term sets are different. On DVD reviews dataset, all the top 5 frequent term sets by *min-sup* contain word "i". Apparently, they are selected due to their high occurrences, but "i" is a common pronoun without discriminating value for judging

Table 3
Top single term and frequent term sets in book and dvd

Top 5 terms	Frequent term sets	Support count	Refined term sets	AD-Sup	Top 5 terms	Frequent term sets	Support count	Refined term sets	AD-Sup
Bore	i/book	1273	bore/i/so	0.922	wast	i/work	678	terribl/i/all	0.849
Disappoint	so/book	635	bore/i/book	0.896	worst	so/i	612	terribl/do	0.842
Wast	i/so	619	wast/i	0.875	bad	i/all	509	i/worthless	0.839
Bad	book/more	568	disappoint/i/out	0.872	great	great/i	440	worst/work	0.807
Excel	book/what	559	wast/book/i	0.860	bore	would/i	386	wast/i/monei/work	0.795

Table 4
Top single term and frequent term sets in electronic and kitchen

Top 5 terms	Frequent term sets	Support count	Refined term sets	AD-Sup	Top 5 terms	Frequent term sets	Support count	Refined term sets	AD-Sup
Great	us/i	734	terribl/i/	1.0	easi	num/my	373	return/time/first/	1.0
Return	work/i	678	refund/i	1.0	return	my/so	357	worst/ever	0.962
Excel	i/get	518	worst/i	0.963	great	num/so	305	wast/monei/do	0.938
Price	i/all	509	return/i/bui	0.959	love	my/time	275	wast/monei/even	0.935
Wast	i/when	488	wast/work	0.955	disappoint	my/get	254	wast/monei/what	0.870

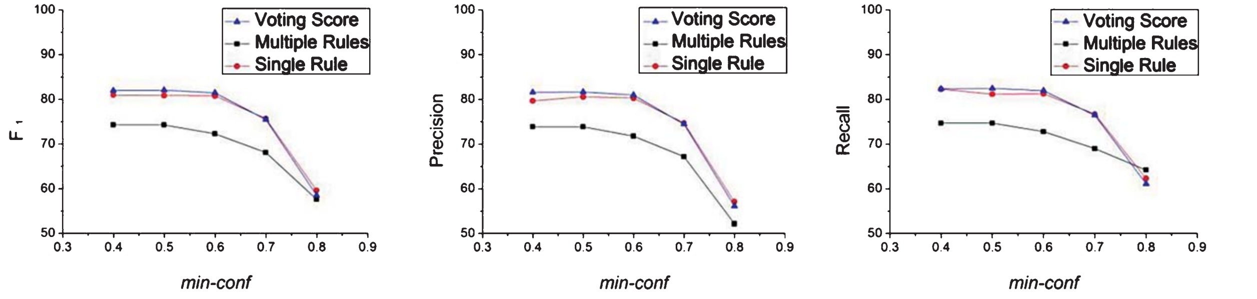

Fig. 2. Classification results of different strategies on Book: F_1, Precision and Recall (%).

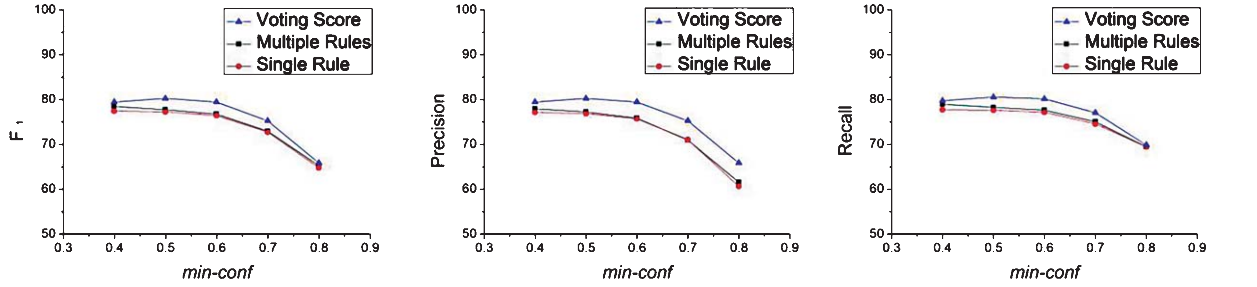

Fig. 3. Classification results of different strategies on DVD: F_1, Precision and Recall (%).

a review's sentiment. We can observe similar results in all the four datasets. It can be concluded that as a kind of personal opinion expression, both positive and negative web reviews contains similar language structures. Support count is not sufficient to extract effective frequent term sets for classification. Contrarily, as is show in Tables 3 and 4, term sets selected by AD-Sup contain more sentiment oriented words and also have better consistency with the most important single terms. A *min-AD-Sup* threshold is then used to prune these un-discriminative term sets before mining classification rules.

We set *min-conf* as 50% in this experiment to extract classification rules. In binary classification problem, 50% is the minimum value to avoid the same antecedent occurs in two different rules in positive rule set and negative rule set. Tables 5 and 6 list the top classification rules of the four datasets.

Table 5
Top positive (pos) and negative (neg) rules for book and dvd dataset

Pos rules	Confidence	Neg rules	Confidence	Pos rules	Confidence	Neg rules	Confidence
Recommend/highli	0.880	bore/i/so	0.951	great/i/perfect	0.906	terribl/i	1.0
Great/book/best	0.822	bore/i	0.941	excel/i/recommend	0.903	work/junk	1.0
Great/book/i/love	0.817	bore/so	0.924	excel/would	0.903	terribl/all	1.0
Great/best	0.808	bore/book	0.909	excel/well	0.90	worst/work	1.0
Excel/book	0.803	wast/i	0.899	excel/bui	0.90	terribl/do	1.0

Table 6
Top positive (pos) and negative (neg) rules for electronics and kitchen dataset

Pos rules	Confidence	Neg rules	Confidence	Pos rules	Confidence	Neg rules	Confidence
excel/price	0.932	terribl/i	1.0	easi/so/love	0.970	wast/monei	1.0
perfect/i/us	0.930	refund/i	0.980	easi/num/love	0.969	return/product	1.0
perfect/us	0.930	return/bui	0.958	easi/great/love	0.968	return/bui	1.0
price/good/well	0.927	worst/i/	0.957	love/dishwash	0.967	return/again	1.0
great/perfect	0.925	support/call	0.95	best/so/can	0.967	worst/ever	1.0

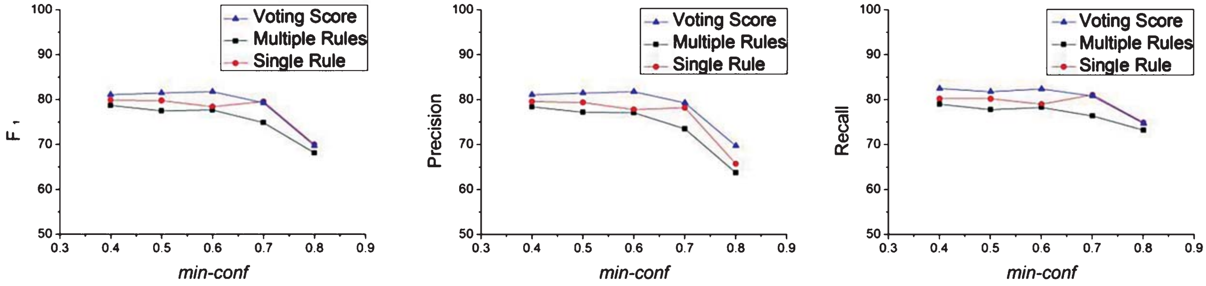

Fig. 4. Classification results of different strategies on Electronic: F_1, Precision and recall (%).

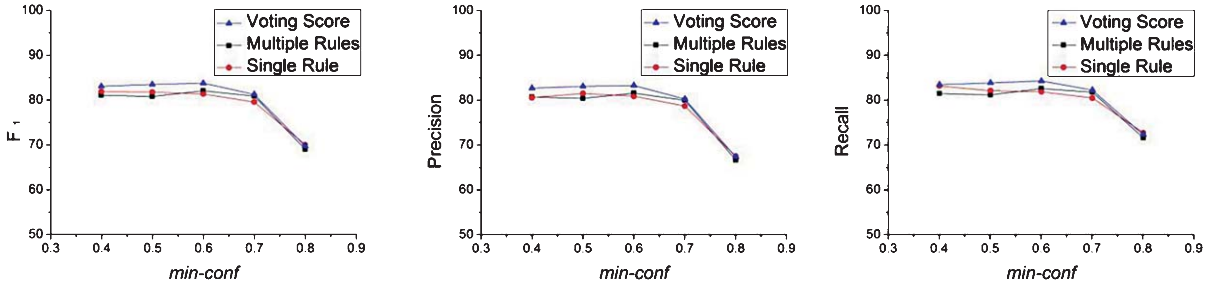

Fig. 5. Classification results of different strategies on Kitchen: F_1, Precision and recall (%).

4.3. *Sentiment prediction using multiple metric votes*

For each test review, four metrics are collected to vote for whether it is positive or negative based on the classification rules that the test review covers. The range of judging score is −4, 4. When the score is a positive number, the review will be labeled as positive, and when the score is a negative number, respectively, the review will be labeled as negative. Figures 2 to 5 are the F_1 value comparison of classification result on the four dataset. To demonstrate the effectiveness of the metric voting method, we select two baseline algorithms: the Single Rule tries to match the maximum confidence rule of the covered rules set, and the Multi-Rules compares both maximum confidence and average confidence value. The results show that in all the four datasets, our strategy outperforms the other two baselines. The maximum improvement of 2.8% was obtained on DVD. On each dataset, all the algorithms were implemented with different *min-conf*. The results also prove that 50% is the best *min-conf* to generate promising and sufficient classification rules. When the *min-conf* increases to 70% and 80%, the performance declined rapidly because the number of rules decreases to a very little scale and makes too many test documents covered by none of the rules.

Table 7
Classification results vs. Other classifiers : F_1 (%)

Dataset	SMO	LibSVM	C4.5	NB	kNN	Voting score
Book	77.1	81.8	78.5	77.7	66.6	**82.1**
DVD	78.7	76.0	75.2	77.1	66.4	**80.3**
Electronic	**82.5**	75.6	79.3	73.5	67.4	81.9
Kitchen	82.1	81.9	80.7	75.6	58.5	**83.9**

Table 7 summarizes the classification results of the proposed method comparing with other popular machine learning classifiers. SVM, Naïve Bayes (NB), kNN and C4.5 are well studied text document classifiers with very good performance track in many previous researches. SVM was implemented with two algorithms, LibSVM and SMO [37]. In three of the four datasets, the F_1 value of multiple metric voting strategy surpassed the other four benchmark algorithms, except for Electronic where the result of our method is very close to SMO. (The bold value in Table 7 are to help to illustrate the last sentence of this paragraph, since the bold value is the maximum one for each Dataset.)

5. Conclusions

This paper has presented a novel approach using association rules for sentiment classification of web reviews. To extract discriminative frequent term sets, a new restraint measure AD-Sup was used which considers more on the term set distribution on different sentiment classes. The extraction was implemented with the Apriori strategy, and the experiment results on multiple domain reviews from real web sites demonstrated that AD-Sup was an effective metric to eliminate terms with no sentiment orientation. An optimal classification rule set was generated which abandons the redundant general rule with lower confidence than the specific one. In the class label prediction procedure, we proposed a new metric voting scheme to solve the problem when the covered rules are not adequately confident or not applicable. The final score of a test review depends on the overall contributions of four metrics. To demonstrate the effectiveness of the voting strategy, we compared the classification performance with other strategies of using confidence. The result shows 50% is the best *min-conf* to guarantee classification rules both abundant and persuasive, and the voting method improves the classification results on all the four datasets. We also compared the proposed method with popular machine learning algorithms including SVM, Naïve Bayes and kNN. The result also shows that our strategy is effective and outperforms the other strong benchmarks. Since this research focused on binary classification, our future work will concentrate on a further optimization and the extension to multi-label sentiment classification problem.

References

[1] B. Pang and L. Lee, Opinion mining and sentiment analysis, *Foundations and Trends in Information Retrieval* **2**(1-2) (2008), 1–135.

[2] T. Huifeng, S. Tan and X. Cheng, A survey on sentiment detection of reviews, *Expert Systems with Applications* **36**(7) (2009), 10760–10773.

[3] W. Zhang, G. Ding and L. Chen, Augmenting chinese online video recommendations by using virtual ratings predicted by review sentiment classification, *Data Mining Workshops (ICDMW), 2010 IEEE International Conference on, IEEE*, 2010, pp. 1143–1150.

[4] D. Lee, O. Jeong and S. Lee, Opinion mining of customer feedback data on the web, *Proceedings of the 2nd International Conference on Ubiquitous Information Management and Communication, ACM*, 2008, pp. 230–235.

[5] M. Rushdi Saleh, M.T. Martín-Valdivia, A. Montejo-Ráez and L.A. Ureña-López, Experiments with SVM to classify opinions in different domains, *Expert Systems with Applications* **38**(12) (2011), 14799–14804.

[6] F. Sebastiani, Machine learning in automated text categorization, *ACM Computing Surveys* **34**(1) (2002), 1–47.

[7] B. Pang, L. Lee and S. Vaithyanathan, Thumbs up? Sentiment classification using machine learning techniques, *Proc ACL-02 conference on Empirical methods in natural language processing-Volume 10, Association for Computational Linguistics*, 2002, pp. 79–86.

[8] H. Mahgoub, Mining Association Rules from Unstructured Documents, *Proc 3rd International Conference on Knowledge Mining*, 2006, pp. 167–172.

[9] F. Thabtah, A review of associative classification mining, *The Knowledge Engineering Review* **22**(1) (2007), 37–65.

[10] J. Han, H. Cheng, D. Xin and X. Yan, Frequent pattern mining: Current status and future directions, *Data Mining and Knowledge Discovery* **15**(1) (2007), 55–86.

[11] B. Pang and L. Lee, A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts, *Proc the 42nd Annual Meeting of the Association for Computational Linguistics*, 2004, pp. 271–278.

[12] M. Thomas, B. Pang and L. Lee, Get out the vote: Determining support or opposition from Congressional floor-debate transcripts, *Proc Conf on Empirical Methods in Natural Language Processing*, 2006, pp. 327–335.

[13] F. Sebastiani, et al., Multi-Facet Rating of Product Reviews, *Proc 31st European Conference on Information Retrieval*, 2009, pp. 461–472.

[14] S. Kiran, P. Pingali and V. Varma, Published supervised learning approaches for rating customer reviews, *Journal of Intelligent Systems* **19**(1) (2010), 79–94.

[15] R. Moraes, F. Valiati and P. GaviãO, Document-level sentiment classification: An empirical comparison between SVM and ANN, *Expert Systems with Applications* **40**(2) (2012), 621–633.

[16] S. Liu, F. Li and F. Li, Adaptive co-training SVM for sentiment classification on tweets, *Proceedings of the 22nd ACM International Conference on Conference on Information & Knowledge Management, ACM*, 2013, pp. 2079–2088.

[17] B. Agarwal and N. Mittal, Sentiment Classification using Rough Set based Hybrid Feature Selection, *Proceedings of the 4th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, Atlanta, Georgia, pp. 115–119.

[18] R.M. Tong, An Operational System for Detecting and Tracking Opinions in on-line discussion, *SIGIR Workshop on Operational Text Classification*, 2001, pp. 1–6.

[19] Y. Hu and W. Li, Document sentiment classification by exploring description model of topical terms, *Computer Speech and Languate* **25**(2) (2001), 386–403.

[20] W. Theresa, J. Wiebe and P. Hoffmann, Recognizing contextual polarity in phrase-level sentiment analysis, *Proc Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, 2005, pp. 347–354.

[21] T. Peter and M.L. Littman, Measuring praise and criticism: Inference of semantic orientation from association, *ACM Transactions on Information Systems* (2003), 315–346.

[22] V. Hatzivassiloglou and K.R. McKeown, Predicting the semantic orientation of adjectives, *Proc 35th Annual Meeting of the Association for Computational Linguistics*, 1997, pp. 174–181.

[23] J. Wiebe, T. Wilson and C. Cardie, Annotating expressions of opinions and emotions in language, *Language Resources and Evaluation* **39**(2) (2005), 165–210.

[24] S. Kim and E. Hovy, Determining the sentiment of opinions, *Proc COLING* **2004** (2004), 1267–1373.

[25] J. Wiebe, Learning Subjective Adjectives from Corpora, *Proc 17th National Conference on Artificial Intelligence*, 2000, pp. 735–740.

[26] H. Cheng, X. Yan, J. Han and C. Hsu, Discriminative frequent pattern analysis for effective classification, *Proc the*

23rd International Conference on Data Engineering, 2007, pp. 716–725.

[27] H. Ahonen-Myka, Discovery of frequent word sequences in text, *Pattern Detection and Discovery, LNAI* **2447** (2002), 180–189.

[28] E. Hernández-Reyes, R.A. García-Hernández, J.A. Carrasco-Ochoa and J. Fco, Martínez-Trinidad, Document Clustering Based on Maximal Frequent Sequences, *Lecture Notes in Computer Science* **4139** (2006), 257–267.

[29] W. Zhang, T. Yoshida, X. Tang and Q. Wang, Text clustering using frequent itemsets, *Knowledge-Based Systems* **23**(5) (2010), 379–388.

[30] Y. Li, S.M. Chung and J.D. Holt, Text document clustering based on frequent word meaning sequences, *Data & Knowledge Engineering* **64**(1) (2008), 381–404.

[31] F. Beil, M. Ester and X.W. Xu, Frequent term-based text clustering, *Proc 8th ACM SIGKDD*, 2002, pp. 436–442.

[32] B. Liu, W. Hsu and Y. Ma, Integrating Classification and Association Rule Mining, *Proc KDD'98*, 1998, pp. 80–86.

[33] M. Antonie and O.R. Zaïane, Text Document Categorization by Term Association, *Proc IEEE International Conference on Data Mining (ICDM'02)*, 2002, pp. 19–26.

[34] G. Forman, An extensive empirical study of feature selection metrics for text classification, *Journal of Machine Learning Research* **3**(1) (2003), 1289–1305.

[35] R. Agrawal and R. Srikant, Fast algorithms for mining association rules in large databases, *Proc 20th International Conference on Very Large Data Bases*, 1994, pp. 487–499.

[36] M. Yuan, Y. Ouyang and Z. Xiong, A text categorization method using extended vector space model by frequent term sets, *Journal of Information Science and Engineering* **29**(1) (2013), 99–114.

[37] J.C. Platt, Sequential minimal optimization: A fast algorithm for training support vector machines, *Technical Report MSR-TR-98-14*, Microsoft Research, 1998.

[38] J. Li, On optimal rule discovery, *IEEE Transactions on Knowledge and Data Engineering* **18**(4) (2006), 460–471.

[39] http://www.cs.jhu.edu/ mdredze/datasets/sentiment/.