PAPER

Unsupervised Sentiment-Bearing Feature Selection for Document-Level Sentiment Classification

Yan LI[†a)], *Nonmember*, **Zhen QIN**[†], *Student Member*, **Weiran XU**[†], **Heng JI**[††], *and* **Jun GUO**[†], *Nonmembers*

SUMMARY Text sentiment classification aims to automatically classify subjective documents into different sentiment-oriented categories (e.g. positive/negative). Given the high dimensionality of features describing documents, how to effectively select the most useful ones, referred to as sentiment-bearing features, with a lack of sentiment class labels is crucial for improving the classification performance. This paper proposes an unsupervised sentiment-bearing feature selection method (USFS), which incorporates sentiment discriminant analysis (SDA) into sentiment strength calculation (SSC). SDA applies traditional linear discriminant analysis (LDA) in an unsupervised manner without losing local sentiment information between documents. We use SSC to calculate the overall sentiment strength for each single feature based on its affinities with some sentiment priors. Experiments, performed using benchmark movie reviews, demonstrated the superior performance of USFS.

key words: *feature selection, sentiment discriminant analysis, sentiment strength calculation, sentiment classification*

1. Introduction

With the rapid development of web technology, huge amount of documents containing opinions and emotions have emerged on the Internet. They provide a large volume of opinionated data about consumer preferences, stored in online review websites, web forums, blogs, etc. For mining knowledge contained in these documents, sentiment analysis has been developed. Document-level sentiment classification aims to automatically judge the type of sentiment orientation, positive ('thumbs up') or negative ('thumbs down') of a subjective document, by mining and analyzing the subjective information. It has been applied in review mining, product reputation analysis, multi-document summarization, multi-perspective question answering, etc. [1]–[4].

One major challenge in document-level sentiment classification is how to deal with the high dimensionality of features used to describe documents. Feature selection is thus regarded as a crucial technique. Unfortunately, to our knowledge, state-of-the-art feature selection techniques for sentiment classification are much less mature than those for topic-oriented classification. As a main reason, topics are always represented by keywords objectively and explicitly, while the sentiments are expressed in an implicit manner. In addition, sentiments are usually hidden in a large amount of subjective information in the documents, which reflects the author's standpoint, view, attitude, mood and so on. Therefore, document-level sentiment classification requires deeper analysis and understanding of textual statement information and thus sentiment-bearing feature selection is more challenging.

Furthermore, unsupervised feature selection for sentiment analysis has received more and more attentions recently [5]–[9]. It is even more difficult to select sentiment-bearing features in an unsupervised manner. As a key reason, it is intractable to assess the relevance of a feature without resorting to class labels. Although there are some materials annotating sentiment labels on movie reviews, product reviews and news articles [10]–[13], the annotation is quite domain dependent and it is time-consuming and costly to obtain labeled data for new resources.

In view of questions mentioned above, we devise an unsupervised sentiment-bearing feature selection algorithm (USFS) in the hope of improving the accuracy of document-level sentiment classification, which is implemented by the support of the following innovations:

- Applying sentiment discriminant analysis (SDA), the unsupervised variation of Linear Discriminant Analysis, to select the most sentiment-bearing features without losing local sentiment information between documents.
- Conducting sentiment strength calculation (SSC) for each single feature. A unique link-weighting scheme, which can preserve overall structural data information, is utilized to measure features' sentiment affinities with some pieces of sentiment prior knowledge.

The rest of this paper is organized as follows. We introduce the related work in Sect. 2 and detail the USFS algorithm in Sect. 3. The evaluation results are shown and analyzed in Sect. 4, and finally, Sect. 5 concludes the paper.

2. Related Work

2.1 Feature Reduction

In general, feature dimension reduction techniques are broadly classified into two types: feature selection and feature extraction [14].

Feature selection algorithms reduce the dimension of

Manuscript received April 8, 2013.
Manuscript revised July 18, 2013.

[†]The authors are with the School of Information and Communication Engineering, Beijing University of Posts and Telecommunications, Beijing 100876, China.
[††]The author is with Rensselaer Polytechnic Institute (RPI), Troy, NY 12180, USA.
a) E-mail: buptliyan@gmail.com

DOI: 10.1587/transinf.E96.D.2805

the feature space by selecting a subset of the most effective features from the original set. Existing feature selection techniques mainly fall into one of the three categories: filter, wrapper and embedded. Filter methods select the best features according to some feature evaluation metrics and use the selected features directly in the classifier. Information gain (IG), mutual information (MI), Chi-square statistic (CHI), etc. are widely-used evaluation metrics [15]. Wang et al. [16] utilized an improved Fisher discriminant ratio to realize filter feature selection for text sentiment classification. Wrapper methods evaluate feature subsets in a real classifier and select features according to their classification performance. Abbasi et al. [17] proposed a wrapper selection method for opinion classification in Web forums by incorporating IG into genetic algorithm. In embedded methods, the search process of the optimal feature subset is built into the classifier construction.

Feature extraction transforms an existing feature space to a lower dimensional space. Principal component analysis (PCA) [18] and linear discriminant analysis (LDA) [19] are the most commonly used techniques for feature extraction. PCA transforms data into a new space by combining original features into a group of uncorrelated variables and selecting some of them which can reflect original information as much as possible. LDA uses the class information to perform a projection of the features which best separate two or more classes. Sugiyama [20] improved traditional LDA through retaining local structure of data. Yang et al. [21] further modified their work to an unsupervised variation.

Different from the previous work, our proposed USFS algorithm focuses on selecting the most sentiment-bearing features from the original ones by analyzing local sentiment information between documents and calculating feature sentiment strength simultaneously. It belongs to the scope of unsupervised filter feature selection method.

2.2 Document-Level Sentiment Classification

Document-level sentiment classification researches have fallen into two categories, i.e., score-based approaches and machine learning techniques.

Score-based sentiment classification generally classify message sentiments based on the total sum of comprised positive or negative sentiment features. [22], [23] and [24] conducted pattern phrase matching to assign positive phrases a "+1" while negative ones a "−1". Turney [25] predicted semantic orientation of a phrase according to its differential PMI (pointwise mutual information) value with two seeds "excellent" and "poor". User reviews were then classified by the average semantic orientation of phrases. On the basis of Turney's method, Gamon and Aue [26] further mined sentiment terms by using an additional assumption that sentiment terms of opposite orientation tend not to co-occur at the sentence level, which yielded a higher classification recall. Subasic and Huettner [27] have also applied score-based methods to affect analysis, where the affect features are scored based on their degree of intensity for a particular emotion class.

Machine learning techniques train a sentiment classifier based on features learned in the training documents. Great bulk of work has been focused on the problem of document-level sentiment classification using machine learning techniques. Pang et al. [10] used various features such as N-grams and Part-of-Speech tags to examine whether it suffices to treat sentiment classification simply as a special case of topic-based classification. As a result, they found that sentiment classification requires deeper understanding. They further improved the classification accuracy by extracting the subjective sentences of the movie reviews using minimum cuts [28]. Abbasi et al. [9] constructed a feature relation network to efficiently enable the inclusion of extended sets of heterogeneous N-gram features to enhance sentiment classification. Lin and He [7] proposed a fully unsupervised probabilistic modeling framework, called joint sentiment/topic model, based on Latent Dirichlet Allocation (LDA) for movie review sentiment classification. Li et al. [8] extended their work and designed a framework of dependency-sentiment-LDA on the assumption that sentiments are related to the topic in the document and are dependent on local context.

In the study of document-level sentiment classification problems, SVMs have been extensively used for movie reviews [5], [6], [9], [10], [28], [29]. Moreover, SVMs have outperformed other classification methods such as Naïve Bayes, centroid classifier, K-nearest neighbor, Winnow classifier [10], [30]. Therefore, in our experiment we also use SVMs as our main classification approach.

3. Unsupervised Sentiment-Bearing Feature Selection

The proposed algorithm starts from capturing sentiment priors according to an existing sentiment lexicon. Then it conducts sentiment discriminant analysis (SDA) and sentiment strength calculation (SSC) simultaneously. Finally, it integrates SDA and SSC by a linear combination, and ranks and selects top sentiment-bearing features in light of their final sentiment scores. The algorithm performs in an unsupervised manner such that no sentiment polarity labels are needed.

3.1 Defining Sentiment Priors

Sentiment priors can be obtained from a sentiment lexicon. We utilized the MPQA subjectivity lexicon[†], which is a widely used knowledge base in the field of sentiment analysis, to generate the sentiment priors in the experiments. MPQA consists of 2,718 positive and 4,911 negative entries. The lexicon is domain-independent and thus does not bear any supervision to a specific context which may influence feature sentiment orientations. To avoid such ambiguous sentiment information, we kept only the entries which have attributes of strong subjectivity strength in the lexicon.

[†]http://www.cs.pitt.edu/mpqa/

Then we matched the remaining entries with the vocabulary of our dataset and removed the ones occurred fewer than 10 times. As a result, 731 subjective (383 positive and 348 negative) words were retained, which finally formed the sentiment prior knowledge set S. The detailed information of the dataset will be given in Sect. 4.1.

3.2 Sentiment Discriminant Analysis

Linear discriminant analysis (LDA) [19] is a popular method for linear dimensionality reduction, which maximizes between-class scatter matrix $\mathbf{S^b}$ and minimizes within-class scatter matrix $\mathbf{S^w}$. It works well for topic-based document classification problem. However, traditional LDA does not consider sentiment information expressed in documents. Moreover, LDA needs class labels to compute the scatter matrices, which violates the goal of our unsupervised algorithm. To alleviate these bottlenecks, we present sentiment discriminant analysis (SDA), which is an unsupervised variant of LDA suitable for selecting sentiment-bearing features.

Let $\mathbf{X} = [\mathbf{x_1}, \mathbf{x_2}, \ldots, \mathbf{x_n}] \in \mathbb{R}^{d*n}$ be the data matrix, where d is the original feature dimensionality. In spite of the lack of class labels, we assume a linear classifier $\mathbf{W} \in \mathbb{R}^{d*c}$ such that $\mathbf{Y} = \mathbf{W}^{\mathrm{T}}\mathbf{X} \in \{0,1\}^{c*n}$, where c is the class number and $\|\mathbf{Y}(:,\mathbf{j})\|_0 = 1$. For the ease of computation, SDA replaces $\mathbf{S^w}$ with the mixture scatter matrix $\mathbf{S^m}$ according to the equality $\mathbf{S^m} = \mathbf{S^w} + \mathbf{S^b}$. That is to say, SDA aims at maximizing $\mathbf{S^b}$ while minimizing $\mathbf{S^m}$. The definitions of $\mathbf{S^b}$ and $\mathbf{S^m}$ are shown as follows:

$$\mathbf{S^b} = \sum_{i=1}^{c}(\sum_{j=1}^{n} y_{ij})(\boldsymbol{\mu_i} - \boldsymbol{\mu})(\boldsymbol{\mu_i} - \boldsymbol{\mu})^{\mathrm{T}} = \tilde{\mathbf{X}}\mathbf{R}^{\mathrm{T}}\mathbf{R}\tilde{\mathbf{X}}^{\mathrm{T}}, \quad (1)$$

$$\mathbf{S^m} = \sum_{j=1}^{n}(\mathbf{x_j} - \boldsymbol{\mu})(\mathbf{x_j} - \boldsymbol{\mu})^{\mathrm{T}} = \tilde{\mathbf{X}}\tilde{\mathbf{X}}^{\mathrm{T}}, \quad (2)$$

where $\boldsymbol{\mu}$ and $\boldsymbol{\mu_i}$ are the means of all samples and the i-th class samples respectively. $\tilde{\mathbf{X}}$ is the centered data matrix which can be realized as $\tilde{\mathbf{X}} = \mathbf{X}\mathbf{H_n}$ where $\mathbf{H_n} = \mathbf{I_n} - \frac{1}{n}\mathbf{1_n}\mathbf{1_n^{\mathrm{T}}}$. And $\mathbf{R} = \mathbf{Y}(\mathbf{Y}^{\mathrm{T}}\mathbf{Y})^{-1/2}$ is the scaled label matrix.

Recently, Sugiyama [20] and Yang et al. [21] demonstrated that local discriminative information is more important than the global one. Inspired by this, SDA tries to keep nearby sample pairs close in the reducing feature space without losing local sentiment information for each document. We denote $\mathbf{x_i^{(s)}}$ as the i-th sample represented in the sentiment prior space S. Then the sentiment similarity between $\mathbf{x_i}$ and $\mathbf{x_j}$ can be computed by their cosine similarity. For each sample $\mathbf{x_i}$, we gather its k nearest sentiment neighbors $\mathbf{x_{i_1}}, \ldots, \mathbf{x_{i_k}}$ and $\mathbf{x_i}$ itself to form $\mathbf{X_{(i)}} = [\mathbf{x_i}, \mathbf{x_{i_1}}, \ldots, \mathbf{x_{i_k}}]$, which is called as the local sentiment region of $\mathbf{x_i}$. Now the local sentiment scatter matrix $\mathbf{S^b_{(i)}}$ and $\mathbf{S^m_{(i)}}$ are defined as follows:

$$\mathbf{S^b_{(i)}} = \tilde{\mathbf{X}}_{\mathbf{(i)}}\mathbf{R}^{\mathrm{T}}_{\mathbf{(i)}}\mathbf{R}_{\mathbf{(i)}}\tilde{\mathbf{X}}^{\mathrm{T}}_{\mathbf{(i)}}, \quad (3)$$

$$\mathbf{S^m_{(i)}} = \tilde{\mathbf{X}}_{\mathbf{(i)}}\tilde{\mathbf{X}}^{\mathrm{T}}_{\mathbf{(i)}}, \quad (4)$$

where $\tilde{\mathbf{X}}_{\mathbf{(i)}} = \mathbf{X_{(i)}}\mathbf{H_{k+1}}$ and $\mathbf{R_{(i)}} = [\mathbf{R_i}, \mathbf{R_{i_1}}, \ldots, \mathbf{R_{i_k}}]$. We define the sentiment discriminative score SDS_i of $\mathbf{x_i}$ as:

$$SDS_i = Tr[(\mathbf{S^m_{(i)}} + \sigma\mathbf{I})^{-1}\mathbf{S^b_{(i)}}], \quad (5)$$

where $\sigma\mathbf{I}$ is added to make the matrix invertible. SDA intends to obtain an optimal linear classifier $\mathbf{W}$ with the highest sentiment discriminative scores for all samples, which leads to the following objective function:

$$\begin{aligned} \underset{\mathbf{W}}{argmin} & \sum_{i=1}^{n}\{Tr[\mathbf{R_{(i)}}\mathbf{H_{k+1}}\mathbf{R^{\mathrm{T}}_{(i)}}] - SDS_i\} + \alpha\|\mathbf{W}\|_{2,1} \\ & s.t. \|\mathbf{Y}(:,\mathbf{j})\|_0 = 1, \quad 0 \leqslant j \leqslant n, \end{aligned} \quad (6)$$

where $\mathbf{R_{(i)}}\mathbf{H_{k+1}}\mathbf{R^{\mathrm{T}}_{(i)}}$ is added to avoid overfitting. $\|\mathbf{W}\|_{2,1}$, the $l_{2,1}$-norm of $\mathbf{W}$, controls the capacity of $\mathbf{W}$ and also ensures that $\mathbf{W}$ is sparse in rows, making it particularly suitable for feature selection. For the ease of representation, we define a selection matrix $\mathbf{L_{(i)}} \in \{0,1\}^{n*(k+1)}$ as follows:

$$(\mathbf{L_{(i)}})_{pq} = \begin{cases} 1, & if\ p = \mathcal{N}_i\{q\}; \\ 0, & otherwise, \end{cases}$$

where $\mathcal{N}_i = \{i, i_1, \ldots i_k\}$ is the indicator set of the sentiment neighbors of sample $\mathbf{x_i}$. Then we have $\mathbf{R_{(i)}} = \mathbf{W}^{\mathrm{T}}\mathbf{X}\mathbf{L_{(i)}}$. Now we can rewrite the first term in our objective function as follows:

$$\begin{aligned} & \sum_{i=1}^{n}\{Tr[\mathbf{R_{(i)}}\mathbf{H_{k+1}}\mathbf{R^{\mathrm{T}}_{(i)}}] - SDS_i\} \\ & = \sum_{i=1}^{n} Tr\{\mathbf{W}^{\mathrm{T}}\mathbf{X}\mathbf{L_{(i)}}\mathbf{H_{k+1}}\mathbf{L^{\mathrm{T}}_{(i)}}\mathbf{X}^{\mathrm{T}}\mathbf{W} - \\ & \quad \mathbf{W}^{\mathrm{T}}\mathbf{X}\mathbf{L_{(i)}}\tilde{\mathbf{X}}^{\mathrm{T}}_{\mathbf{(i)}}(\tilde{\mathbf{X}}_{\mathbf{(i)}}\tilde{\mathbf{X}}^{\mathrm{T}}_{\mathbf{(i)}} + \sigma\mathbf{I})^{-1}\tilde{\mathbf{X}}_{\mathbf{(i)}}\mathbf{L^{\mathrm{T}}_{(i)}}\mathbf{X}^{\mathrm{T}}\mathbf{W}\} \\ & = Tr\{\mathbf{W}^{\mathrm{T}}\mathbf{X}\{\sum_{i=1}^{n}[\mathbf{L_{(i)}}(\mathbf{H_{k+1}} - \\ & \quad \tilde{\mathbf{X}}^{\mathrm{T}}_{\mathbf{(i)}}(\tilde{\mathbf{X}}_{\mathbf{(i)}}\tilde{\mathbf{X}}^{\mathrm{T}}_{\mathbf{(i)}} + \sigma\mathbf{I})^{-1}\tilde{\mathbf{X}}_{\mathbf{(i)}})\mathbf{L^{\mathrm{T}}_{(i)}}]\}\mathbf{X}^{\mathrm{T}}\mathbf{W}\} \\ & = Tr(\mathbf{W}^{\mathrm{T}}\mathbf{A}\mathbf{W}), \end{aligned} \quad (7)$$

where $\mathbf{A} = \mathbf{X}\{\sum_{i=1}^{n}[\mathbf{L_{(i)}}\mathbf{H_{k+1}}(\tilde{\mathbf{X}}_{\mathbf{(i)}}\tilde{\mathbf{X}}^{\mathrm{T}}_{\mathbf{(i)}} + \sigma\mathbf{I})^{-1}\mathbf{H_{k+1}}\mathbf{L^{\mathrm{T}}_{(i)}}]\}\mathbf{X}^{\mathrm{T}}$. The constraint in Eq. (6) makes the objective function difficult to solve. According to common relaxation for label indicator matrix [31], the constraint on $\mathbf{Y}$ is relaxed to orthogonality, i.e., $\mathbf{Y}\mathbf{Y}^{\mathrm{T}} = \mathbf{I_c}$. With this relaxation, Eq. (6) can be reformulated:

$$\begin{aligned} \underset{\mathbf{W}}{argmin} \ & Tr(\mathbf{W}^{\mathrm{T}}\mathbf{A}\mathbf{W}) + \alpha\|\mathbf{W}\|_{2,1} \\ & s.t. \mathbf{W}^{\mathrm{T}}(\mathbf{X}\mathbf{X}^{\mathrm{T}} + \sigma\mathbf{I})\mathbf{W} = \mathbf{I_c}, \end{aligned} \quad (8)$$

where $\sigma\mathbf{I}$ is added to make $(\mathbf{X}\mathbf{X}^{\mathrm{T}} + \sigma\mathbf{I})$ nonsingular.

We describe the detailed procedure of our SDA algorithm in Algorithm 1. Line 1 and line 2 construct the local sentiment regions based on the sentiment prior knowledge S. From line 3 to line 7, SDA computes the matrices defined

Algorithm 1 SDA

Input: $\mathbf{X}$, S, α, σ, c, k.
Output: $\mathbf{SCORE_{SDA}} \in \mathbb{R}^d$.

1: Calculate similarities in the sentiment prior subspace defined by S
2: Select k nearest sentiment neighbors for each sample $\mathbf{x_i}$ and construct local sentiment region $\mathbf{X_{(i)}}$
3: **for** $i = 1$ to n **do**
4: $\quad \mathbf{A_{(i)}} = \mathbf{L_{(i)}}\mathbf{H_{k+1}}(\tilde{\mathbf{X}}_{\mathbf{(i)}}\tilde{\mathbf{X}}_{\mathbf{(i)}}^{\mathrm{T}} + \sigma\mathbf{I})^{-1}\mathbf{H_{k+1}}\mathbf{L}_{\mathbf{(i)}}^{\mathrm{T}}$
5: **end for**
6: $\mathbf{A} = \mathbf{X}(\sum_{i=1}^{n} \mathbf{A_{(i)}})\mathbf{X}^{\mathrm{T}}$
7: $\mathbf{B} = \mathbf{X}\mathbf{X}^{\mathrm{T}} + \sigma\mathbf{I}$
8: set $t = 0$ and initialize $\mathbf{Q_0}$ as an identity matrix
9: **while** not convergent **do**
10: $\quad \mathbf{P_t} = \mathbf{B}^{-1}(\mathbf{A} + \alpha\mathbf{Q_t})$
11: $\quad \mathbf{W_t} = [\mathbf{e_1}, \ldots, \mathbf{e_c}]$ where $\mathbf{e_1}, \ldots, \mathbf{e_c}$ are the eigenvectors of $\mathbf{P_t}$ corresponding to the first c smallest eigenvalues
12: $\quad$ Update the diagonal matrix $\mathbf{Q_{t+1}}$, where the i-th diagonal element is $\frac{1}{2\|\mathbf{W_t}(i,:)\|_2}$
13: $\quad t = t + 1$
14: **end while**
15: **for** each original feature f_i **do**
16: $\quad SCORE_{SDA}(i) = \|\mathbf{W}(i,:)\|_2$
17: **end for**

in Eq. (8). The iterative procedure from line 8 to line 14 is an effective way to solve the $l_{2,1}$-norm minimization problem, which monotonically decreases the objective function and converges to the optimal $\mathbf{W}$. The detailed proof is referred to [32]. SDA outputs a d-dimensional vector $\mathbf{SCORE_{SDA}}$, whose elements equal to the l_2-norm of the corresponding rows in $\mathbf{W}$. In fact, $\mathbf{SCORE_{SDA}}$ records the sentiment discriminative score for each original feature.

3.3 Sentiment Strength Calculation

Guo et al. [33] presented a powerful tool to model complex relationships in real world and build large-scale networks from source data. In the context of their modeling method, we design a weighting scheme to measure activation forces between features in our source corpus and then calculate the sentiment strength for each feature based on its summing affinities with the sentiment priors in S. We name this model SSC for short.

3.3.1 Feature Activation Force

Guo et al. pointed out that features associate with each other in a manner of intricate clusters. The activation strength from one feature to another forges and accounts for the latent structures of the feature network. Specifically, for a given pair of features f_i and f_j, the strength of the link from f_i to f_j is called feature activation force (FAF) and defined as follows:

$$FAF_{ij} = (c_{ij}/c_i) \cdot (c_{ij}/c_j)/d_{ij}^2 \,, \tag{9}$$

where c_i and c_j are the occurrence count of f_i and f_j respectively, c_{ij} is the co-occurrence count of f_i coming before f_j, and d_{ij} is the average distance between the two features.

Algorithm 2 USFS

1: Construct sentiment prior knowledge S based on an existing sentiment lexicon
2: Calculate $\mathbf{SCORE_{SDA}}$ using Algorithm 1
3: Calculate $\mathbf{SCORE_{SSC}}$ using Eq. (12)
4: **for** each feature f_i **do**
5: $\quad SCORE(i) = w * SCORE_{SDA}(i) + (1 - w) * SCORE_{SSC}(i)$
6: **end for**
7: Sort each feature according to $SCORE(i)$ in descending order and select the top ranked ones on the ratio of r.

Obviously, FAF models a directed and weighted feature network and defines a unique link-weighting scheme which determines the strength of the links according to the conditions of the feature occurrences in the given training data set.

3.3.2 Feature Sentiment Strength

With FAF described above, we can formulate the sentiment affinity, SA_{ij}, between each feature f_i and a sentiment prior s_j in S as follows:

$$SA_{ij} = [\frac{1}{|K_{ij}|}\sum_{k \in K_{ij}} OR(FAF_{ki}, FAF_{kj}) \cdot \frac{1}{|L_{ij}|}\sum_{l \in L_{ij}} OR(FAF_{il}, FAF_{jl})]^{\frac{1}{2}}, \tag{10}$$

$$OR(x, y) = min(x, y)/max(x, y), \tag{11}$$

where $K_{ij} = \{k | FAF_{ki} > 0 \; or \; FAF_{kj} > 0\}$, $L_{ij} = \{l | FAF_{il} > 0 \; or \; FAF_{jl} > 0\}$. We can see that K_{ij} and L_{ij} are the out-links and in-links of f_i and s_j respectively. $OR(x, y)$ is an overlap rate function of x and y. So SA_{ij} is the geometric average of the mean overlap rates of the in-links and out-links of the inquired two features. It is deemed that SA measures the semantic sentiment affinity between a feature and a sentiment prior without losing their overall structural information in the data set. Then the whole sentiment strength for a feature is summed over its affinities with all sentiment priors:

$$SCORE_{SSC}(i) = \sum_{s_j \in S} SA_{ij} \,. \tag{12}$$

Similar to $\mathbf{SCORE_{SDA}}$, we denote $\mathbf{SCORE_{SSC}} \in \mathbb{R}^d$ to record the sentiment strength for each feature.

3.4 USFS Algorithm

SDA aims to maximize the between-class scatter and minimize mixture scatter for documents via keeping their local sentiment structure. On the other hand, SSC focuses on capturing sentiment strength for each feature by summing up sentiment affinities. We suggest that SDA and SSC are complementary to each other and thus incorporate them into USFS. As shown in line 5 of Algorithm 2, we combine two scores, $\mathbf{SCORE_{SDA}}$ and $\mathbf{SCORE_{SSC}}$, to calculate the final sentiment score for each feature. The parameter w controls

the importance of the two. Then the original features can be ranked in descending order. The parameter r is the ratio of the selected sentiment-bearing features to the original ones.

4. Experiments

4.1 Experimental Setup

Since extensive work [5]–[7], [9], [17], [28], [29] has used a benchmark movie review dataset[†] developed by Pang et al. [10]. For the ease of comparison, our experiments were also conducted over this dataset, which consists of 1,000 positive and 1,000 negative reviews taken from the IMDb movie review archives. The detailed information is shown in Table 1.

To prepare the documents, we automatically removed the rating indicators and extracted the textual information from the original HTML document format, treating punctuation as separate lexical items. Each document was further decapitalized. After these preprocessing steps, there are 50,920 terms in total. Then we discarded the punctuation and stop words. Considering the effect of negation (i.e. "good" and "not very good" indicate opposite sentiment orientations), we added the tag "NOT_" to the first noun, verb or adjective following a negation term ("not", "no", "never", etc.). The final feature set contains 16,453 individual terms.

We treat document-level sentiment analysis as a binary classification problem (classifying each document as either positive or negative). For the purpose of investigating whether feature selection methods can improve classification performance, we used the LibSVM toolset[††] with linear kernel as our classifier and used the classification accuracy as the evaluation metric. The results reported in the following sections were averaged on 10-fold-cross-validation. The parameter values for USFS are: $\{\alpha = 0.1, \sigma = 0.01, c = 2, k = 5, w = 10\}$. The investigation on the last two parameters will be provided later.

4.2 Evaluations on Baselines and USFS

We explored several baselines consisting of different feature sets to demonstrate the viability of our proposed USFS algorithm.

- B_{ori}: the original feature set described in the above section, which is comprised of 16,453 features.
- B_{sp}: the 731 sentiment priors defined in S.
- B_{SDA}: the feature set selected by SDA ($w = 1$).
- B_{SSC}: the feature set selected by SSC ($w = 0$).

Figure 1 shows the evaluation results. B_{ori} and B_{sp} were evaluated on all of their corresponding features while B_{SDA}, B_{SSC} and USFS were tested on the ratio of selected features. We can conclude the following observations from Fig. 1:

- The results of B_{ori} and B_{sp} are 84.4% and 78.0% respectively, so B_{sp} falls far behind B_{ori}. This may be attributed to the fact that documents express their sentiment orientations in a complex and implicit way. The sentiment prior knowledge cannot hold enough sentiment information.
- Neither B_{SDA} nor B_{SSC} improves classification accuracy drastically. But they achieve their best accuracy scores of 84.3% and 84.5%, which are close to the one of B_{ori}, when only using 50 and 60 percent of the original features respectively. This demonstrates that SDA and SSC are effective to select sentiment-bearing features, but it is not desirable to utilize any of them alone.
- Our proposed USFS algorithm obtains 86.8% accuracy and performs the best on most of the selected feature ratios. This is consistent with our expectation that combining SDA and SSC is promising to select the most sentiment-bearing features. The key to success in USFS is to leverage both the analysis of local sentiment information between documents and the calculation of overall sentiment strength for individual features.
- The USFS algorithm improves sentiment classification performance when r is as small as 0.3, peaks at $r = 0.4$ and keeps performing well when r continues to increase. This observation reveals that USFS is truly suitable for sentiment-bearing feature selection regardless of the number of selected features.

Table 1 Basic information of the movie review dataset.

# of docs		size (MB)		# of	# of
positive	negative	positive	negative	terms	features
1,000	1,000	3.93	3.49	50,920	16,453

[†]http://www.cs.cornell.edu/Pepole/pabo/movie-review-data
[††]http://www.csie.ntu.edu.tw/ cjlin/libsvm

4.3 Comparison with Unsupervised Feature Reduction Methods

In this section, we compare the USFS algorithm with several classical unsupervised feature extraction methods:

- Principal component analysis (PCA) [18] is a mathematical procedure that uses an orthogonal transformation to project a set of samples with possibly correlated

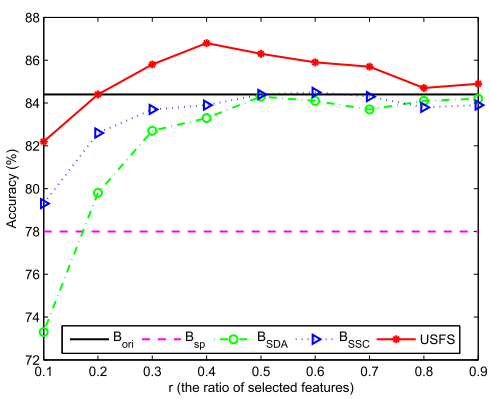

Fig. 1 Results of baselines and USFS.

features into a new space of linearly uncorrelated variables called principal components.
- Latent semantic indexing (LSI) [34] uses a mathematical technique called Singular Value Decomposition (SVD) to correlate semantically related terms that are latent in a text collection.
- Probabilistic latent semantic indexing (PLSI) [35], evolved from LSI, is a novel approach to automated document indexing which is based on a statistical latent class model for factor analysis of count data. PLSI can derive a low dimensional representation of the observed samples in terms of their affinity to certain hidden topics.
- Latent Dirichlet allocation (LDA) [36], another widely used topic model, is a generative probabilistic model which models each sample as a finite mixture over an underlying set of topics. So samples can be represented in a low-dimensional topic-level space.

The comparison results between USFS and these four approaches are shown in Fig. 2, where LSI and PCA achieved the best accuracies when the dimensionality of the embedded space were set to 300 and 1997 (the rank of the data matrix) respectively, and the topic numbers in PLSI and LDA were fixed on 8 and 90 respectively.

We can clearly see that USFS is significantly better than the other unsupervised methods. This is because that USFS considers the latent factors affecting sentiment orientations in both document and feature levels while the other four conduct feature reduction on basis of document content or latent topics. It is notable that all the four unsupervised feature extraction methods even worsen the classification accuracy of B_{ori}. In addition, topic models (PLSI and LDA) perform worse than LSI and PCA, because they lay emphasis on mining the latent topics generating samples and may not be effective for the sentiment classification problem. Although PLSI falls far behind the rest, it acts on a very low feature dimensionality due to its small number of topics.

In the previous section, we have demonstrated that the sentiment prior set holds insufficient sentiment information. We now expand the set using two common methods, pointwise mutual information (PMI) and Markov random walks (MRW). Turney [25] applied a specific unsupervised learning technique based on PMI between document phrases and the words "excellent" and "poor". Hassan and Radev [37] utilized MRW to a word relatedness graph, producing a polarity estimate for any given word. In our experiments, the entire positive and negative sentiment priors were regarded as two seed sets. The weights of other features were computed by the average differential PMI values (for PMI) or first-passage times (for MRW) between the seed sets. These two feature selection approaches can be seen as either unsupervised or semi-supervised.

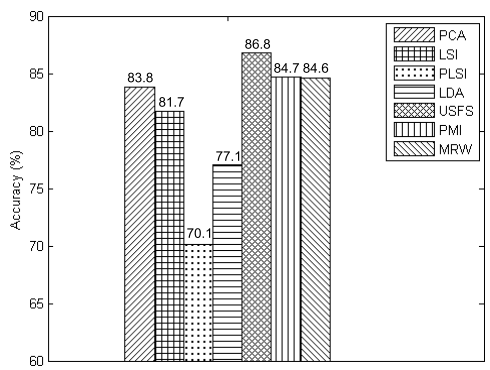

Fig. 2 Results of USFS and unsupervised feature extraction methods.

The best performances of PMI and MRW are also shown in Fig. 2, where the selected feature ratios were fixed to 0.7 and 0.6 respectively. We notice that, comparing to B_{sp} in Fig. 1, PMI and MRW greatly improve the classification accuracy from 78% to nearly 85%, indicating that expanding sentiment priors is indeed helpful to document-level sentiment classification. However, both PMI and MRW perform worse than USFS in spite of their larger feature sets. This further demonstrates the effectiveness of USFS for sentiment-bearing feature selection. In fact, similar to SSC, PMI and MRW only calculate global sentiment strength for each feature, and neither of them capture local sentiment information between documents. Therefore, PMI, MRW and SSC yield comparable results (84.7%, 84.6% and 84.5% respectively).

4.4 Comparison with Supervised Feature Selection Methods

In order to further illustrate the effectiveness of the proposed USFS algorithm, we also investigate the following popular supervised feature selection methods:

- Information gain (IG) [38], a commonly used metric for feature ranking, takes into account the belongingness of a feature in a category as well as its absence in the category.
- Mutual information (MI) [39] is a statistical concept in information theory, which measures dependency between a feature and a particular category.
- Chi-square statistic (CHI) [15] measures the lack of independence between feature and class and can be compared to χ^2 distribution with one degree of freedom to judge extremeness.
- Expected cross entropy (ECE) [40] reflects the probabilistic distributions of categories and their distances given a feature. Different to IG, ECE does not consider the absence of features.
- Contextual merit function (CMF) [41] captures relative importance of features in distinguishing the classes in the context of other features. The main idea is to assign the contextual merit based on the component distance of a feature weighted by the degree of similarity between examples in different classes.

Figure 3 shows the evaluation results. For comparison, the line of B_{ori} is also figured. It can be seen that the presented USFS algorithm still performs the best. Although

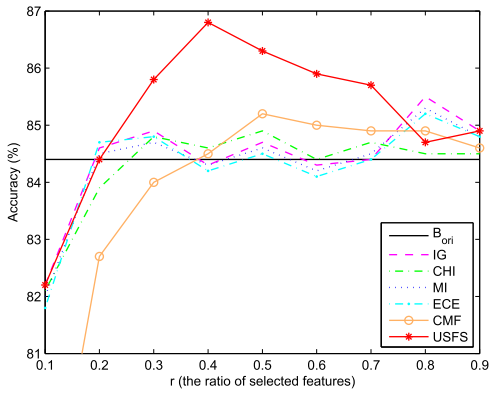

Fig. 3 Results of USFS and supervised feature selection methods.

CMF is rather unstable to r, it performs better in its middle range than the first four methods, which are on par with each other. This further verifies the theory that ranking features by their own correlations to the classes is generally not effective when there is a strong feature interaction in discriminating the classes. Nevertheless, CMF is inferior to USFS on almost all the selection ratios, which may be ascribed to the lack of analysis for feature's global sentiment strength. In addition, CHI seems to be marginally better than IG, MI and ECE and is less sensitive to r. It can be also concluded that the supervised methods are generally better than the unsupervised ones reported in the previous section and can be applied to improve the sentiment classification performance, which is consistent with the observations in [10].

4.5 Comparison with Existing Approaches

For comparison, document-level sentiment classification results on the movie review data set from previous studies are listed in Table 2.

The first three works conducted feature learning in a supervised manner and the following three can be considered as unsupervised approaches. On the basis of the study in [10], Pang and Lee [28] improved classification accuracy to 87.2% by applying SVMs on the subjective portions of the movie reviews which were extracted using a subjectivity detector. Abbasi et al. [17] incorporated IG with an entropy weighted genetic algorithm. By wrapping the SVM accuracy as the fitness function, the best accuracy can achieve 91.7%. Maas et al. [29] presented a model that uses a mix of unsupervised and supervised techniques to learn word vectors capturing semantic term-document information as well as rich sentiment content.

As for the unsupervised ones, Whitelaw et al. [5] used SVMs to train on the combination of different types of appraisal group features and bag-of-words features for sentiment analysis. The reported best accuracy is 90.2% based on 48,314 features. Riloff et al. [6] devised a method to automatically identify features that are subsumed by a simpler feature but that are better opinion indicators. By applying SVMs as the classifier, the reported accuracy is 82.7%. Based on the traditional topic model LDA described in Sect. 4.3, Lin and He [7] proposed a fully unsupervised probabilistic modeling framework called joint sentiment/topic model (JST), which detects sentiment and topic simultaneously. The document sentiment is classified according to the probability of sentiment label given document. Incorporating a prior subjectivity lexicon, the classification accuracy is 82.8% which is much higher than the performance of LDA reported in Sect. 4.3.

Table 2 Results of USFS and previous studies.

Studies	Feature learning methods	Accuracy
Pang and Lee [28]	Subjectivity detector	87.2%
Abbasi et al. [17]	Incorporating IG with GA	91.7%
Maas et al. [29]	Learning word vectors	88.9%
Whitelaw et al. [5]	Appraisal groups	90.2%
Riloff et al. [6]	Subsumption hierarchy	82.7%
Lin and He [7]	JST	82.8%
USFS	Combining SDA and SSC	86.8%

USFS outperforms two unsupervised approaches in [6] and [7]. The good performance reported in [5] may be attributed to the fact that the appraisal groups were constructed semi-automatically by generating candidate features with the help of WordNet and online thesauri. We believe that the additional resources make this research perform in supervision and yield a high accuracy comparable to supervised learning. As for the supervised methods, the results of USFS is only 0.4% and 2.1% lower than the results reported in [28] and [29]. Even for the state-of-the-art result in [17], the accuracy achieved by USFS is only 4.9% lower.

4.6 Parameter Selection

The parameters involved in USFS include: α (controlling $l_{2,1}$-norm regularization), σ (ensuring matrices nonsingular), c (the number of sentiment class labels), k (the number of sample sentiment neighbors), w (the coefficient controlling the weights of SDA and SSC) and r (the ratio of selected features to original ones). Actually, α effects little on the evaluation results and will not be investigated due to the limited space. σ can be set to a small positive number empirically. c was fixed on 2 since the document-level sentiment classification is binary. We set r to 0.4, because it is reasonable that the number of selected features is neither too small nor too large. In a word, we will study the effect of the remaining two important parameters k and w while setting $\{\alpha = 0.1, \sigma = 0.01, c = 2, r = 0.4\}$. The results are shown in Fig. 4.

Most of the time, with the increasing value of both k and w, the performance first increases, reaches its peak value and degrades. The peak accuracy value is achieved when k = 5 and w = 0.4, indicating the importance of retaining local sentiment structure and combining SDA and SSC. Moreover, the low performance on the two edges of w suggests further that utilizing SDA or SSC separately is not enough for sentiment-bearing feature selection.

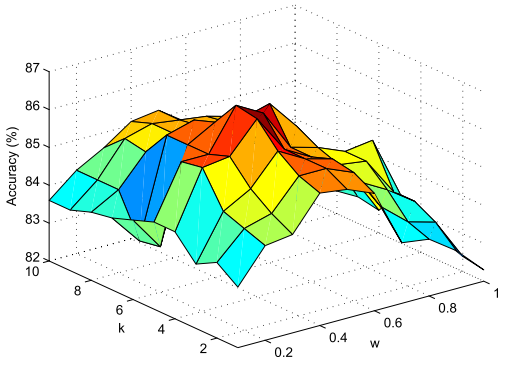

Fig. 4 Evaluation results on k and w.

5. Conclusions and Future Work

In this paper, we presented an unsupervised sentiment-bearing feature selection method (USFS) which conducts sentiment discriminant analysis (SDA) and sentiment strength calculation (SSC) simultaneously. It is believed that USFS is able to preserve both local sentiment information between documents and overall sentiment strength of features. In the experiments conducted on the benchmark movie review dataset, USFS significantly outperformed the no feature selection baseline and some classical feature selection and extraction methods.

Despite the good performance, it is necessary to further demonstrate the efficiency of the proposed USFS algorithm on much more and larger datasets. One of the limitations of our model is that it represents each document as a bag of words and thus ignores the word ordering. In the future, we plan to extend our work to include higher-order information (e.g., N-grams). In addition, how to determine automatically the important parameters involved in our algorithm is another important issue.

Acknowledgments

This work was supported by 111 Project of China under Grant No. B08004 and key project of ministry of science and technology of China under Grant No. 2011ZX03002-005-01. And it was also supported by National Natural Science Foundation of China (60905017, 61072061 and 61273217).

References

[1] L. Zhuang, F. Jing, and X. Zhu, "Movie review mining and summarization," Proc. 15th ACM Int'l Conf. on Inf. and Knowl. Manage., pp.43–50, 2006.
[2] S. Morinaga, K. Yamanishi, K. Tateishi, and T. Fukushima, "Mining product reputations on the web," Proc. 8th ACM SIGKDD Int'l Conf. on Knowl. Disc. and Data Min., pp.341–349, 2002.
[3] R. Ng and A. Pauls, "Multi-document summarization of evaluative text," Proc. European Chapter of the Assoc. for Comput. Linguist., pp.305–312, 2006.
[4] C. Cardie, J. Wiebe, T. Wilson, and D. Litman, "Combining low-level and summary representations of opinions for multi-perspective question answering," Proc. AAAI Spring Symposium on New Directions in Question Answering, pp.20–27, 2003.
[5] C. Whitelaw, N. Garg, and S. Argamon, "Using appraisal groups for sentiment analysis," Proc. 14th ACM Int'l Conf. on Inf. and Knowl. Manage., pp.625–631, 2005.
[6] E. Riloff, S. Patwardhan, and J. Wiebe, "Feature subsumption for opinion analysis," Proc. ACL-06 Conf. on Empir. Methods in Nat. Lang. Proces., pp.440–448, 2006.
[7] C. Lin and Y. He, "Joint sentiment/topic model for sentiment analysis," Proc. 18th ACM Int'l Conf. on Inf. and Knowl. Manage., pp.375–384, 2009.
[8] F. Li, M. Huang, and X. Zhu, "Sentiment analysis with global topics and local dependency," Proc. 24th AAAI Conf. on Artif. Intel., pp.1371–1376, 2010.
[9] A. Abbasi, S. France, Z. Zhang, and H. Chen, "Selecting attributes for sentiment classification using feature relation networks," IEEE Trans. Knowl. Data Eng., vol.23, no.3, pp.447–462, 2011.
[10] B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up? sentiment classification using machine learning techniques," Proc. ACL-02 Conf. on Empir. Methods in Nat. Lang. Proces.-Volume 10, pp.79–86, 2002.
[11] M. Hu and B. Liu, "Mining and summarizing customer reviews," Proc. 10th ACM SIGKDD Int'l Conf. on Knowl. Disc. and Data Min., pp.168–177, 2004.
[12] X. Ding, B. Liu, and P.S. Yu, "A holistic lexicon-based approach to opinion mining," Proc. Int'l Conf. on Web Search and Web Data Mining, pp.231–240, 2008.
[13] J. Wiebe, T. Wilson, and C. Cardie, "Annotating expressions of opinions and emotions in language," Lang. Resour. Eval., vol.39, pp.165–210, 2005.
[14] R.O. Duda, P.E. Hart, and D.G. Stork, Pattern Classification, Second ed., John Wiley & Sons, 1999.
[15] Y. Yang and J. Pedersen, "A comparative study on feature selection in text categorization," Proc. 14th Int'l Conf. on Mach. Learn., pp.412–420, 1997.
[16] S. Wang, D. Li, X. Song, Y. Wei, and H. Li, "A feature selection method based on improved Fisher's discriminant ratio for text sentiment classification," Expert Syst. Appl., vol.38, no.7, pp.8696–8702, 2011.
[17] A. Abbasi, H. Chen, and A. Salem, "Sentiment analysis in multiple languages: feature selection for opinion classification in web forums," ACM Trans. Inf. Syst., vol.26, no.3, 2008.
[18] I.T. Jolliffe, Principal Component Analysis, Springer-Verlag, Berlin, 1986.
[19] R.A. Fisher, "The use of multiple measurements in taxonomic problems," Annals of Eugenics, vol.7, no.2, pp.179–188, 1936.
[20] M. Sugiyama, "Local Fisher discriminant analysis for supervised dimensionality reduction," Proc. 23rd Int'l Conf. on Mach. Learn., pp.905–912, 2006.
[21] Y. Yang, H.T. Shen, Z. Ma, Z. Huang, and X. Zhou, "$l_{2,1}$-norm regularized discriminative feature selection for unsupervised learning," Proc. 22nd Int'l Joint Conf. on Artif. Intel., pp.1589–1594, 2011.
[22] J. Yi, T. Nasukawa, R. Bunescu, and W. Niblack, "Sentiment analyzer: Extracting sentiments about a given topic using natural language processing techniques," Proc. 3rd IEEE Int'l Conf. on Data Min., pp.427–434, 2003.
[23] T. Nasukawa and J. Yi, "Sentiment analysis: capturing favorability using natural language processing," Proc. 2nd Int'l Conf. on Knowl. Capture, pp.70–77, 2003.
[24] Z. Fei, J. Liu, and G. Wu, "Sentiment classification using phrase patterns," Proc. 4th IEEE Int'l Conf. on Comput. Inf. Techn., pp.1147–1152, 2004.
[25] P.D. Turney, "Thumbs up or thumbs down? semantic orientation applied to unsupervised classification of reviews," Proc. 40th Annual Meeting of the Assoc. for Comput. Linguist., pp.417–424, 2002.
[26] M. Gamon and A. Aue, "Automatic identification of sentiment vocabulary: exploiting low association with known sentiment terms,"

Proc. ACL Workshop on Feature En. for Mach. Learn. in NLP, pp.57–64, 2005.
[27] P. Subasic and A. Huettner, "Affect analysis of text using fuzzy semantic typing," IEEE Trans. Fuzzy Syst., vol.9, no.4, pp.483–496, 2001.
[28] B. Pang and L. Lee, "A sentimental education: sentiment analysis using subjectivity summarization based on minimum cuts," Proc. 42nd Annual Meeting of the Assoc. for Comput. Linguist., pp.271–278, 2004.
[29] A.L. Maas, R.E. Daly, P.T. Pham, D. Huang, A.Y. Ng, and C. Potts, "Learning word vectors for sentiment analysis," Proc. 49th Annual Meeting of the Assoc. for Comput. Linguist., pp.142–150, 2011.
[30] S. Tan and J. Zhang, "An empirical study of sentiment analysis for Chinese documents," Expert Syst. Appl., vol.34, no.4, pp.2622–2629, 2008.
[31] U. Luxburg, "A tutorial on spectral clustering," Stat. Comput., vol.17, no.4, pp.395–416, 2007.
[32] F. Nie, H. Huang, X. Cai, and C. Ding, "Efficient and robust feature selection via joint $l_{2,1}$-norms minimization," Adv. in Neural Inf. Proces. Syst., pp.1813–1821, 2010.
[33] J. Guo, H. Guo, and Z. Wang, "An activation force-based affinity measure for analyzing complex networks," Sci. Rep., vol.1, no.113, 2011.
[34] S. Deerwester, S.T. Dumais, G.W. Furnas, T.K. Landauer, and R. Harshman, "Indexing by latent semantic analysis," J. AM. Soc. Inf. Sci., vol.41, no.6, pp.391–407, 1990.
[35] T. Hofmann, "Probabilistic latent semantic indexing," Proc. 22nd Annual Int'l SIGIR Conf. on Res. and Dev. in Inf. Retrieval, pp.50–57, 1999.
[36] D.M. Blei, A.Y. Ng, and M.I. Jordan, "Latent Dirichlet allocation," J. Mach. Learn. Res, vol.3, pp.993–1022, 2003.
[37] A. Hassan and D. Radev, "Identifying text polarity using random walks," Proc. 48th Annual Meeting of the Assoc. for Comput. Linguist., pp.395–403, 2010.
[38] J.R. Quinlan, "Introduction of decision trees," Mach. Learn., vol.1, no.1, pp.81–106, 1986.
[39] R. Fano, Transmission of Information, MIT Press, Cambridge, MA, 1961.
[40] D. Koller and M. Sahami, "Hierarchically classifying documents using very few words," Proc. 11th Int'l Conf. on Mach. Learn., pp.121–129, 1994.
[41] S.J. Hong, "Use of contextual information for feature ranking and discretization," IEEE Trans. Knowl. Data Eng., vol.9, no.5, pp.718–730, 1997.

Yan Li received a M.S. degree from Beijing University of Posts and Telecommunications in 2009. He is currently a Ph.D. student in Beijing University of Posts and Telecommunications. Currently, his main research interests cover opinion mining and sentiment analysis.

Zhen Qin received her M.E. and B.E. degrees in automation from University of Science and Technology Beijing, China in 2009 and 2012, respectively. She is currently a Ph.D. student in Beijing University of Posts and Telecommunications.

Weiran Xu received his Ph.D. degree from Beijing University of Posts and Telecommunications in 2003. He is currently an associate professor in Web Searching Teaching and Research Center, Beijing University of Posts and Telecommunications. His current research fields include information retrieval, pattern recognition and machine learning.

Heng Ji received her Ph.D. in Computer Science from New York University in 2007. Her research interests focus on natural language processing, especially on cross-source information extraction and knowledge base population. She has published over 90 papers. Her recent work on uncertainty reduction for information extraction was invited for publication in the Centennial Year Celebration of IEEE Proceedings. She received a Google Research Award in 2009, NSF CAREER award in 2010, Sloan Junior Faculty award and IBM Watson Faculty award in 2012. She served as the coordinator of the NIST TAC Knowledge Base Population task in 2010 and 2011, the Information Extraction area chair of NAACL-HLT2012 and ACL2013 and the co-leader of the information fusion task of ARL NS-CTA program in 2011 and 2012. Her research has been funded by NSF, ARL, DARPA, Google and IBM.

Jun Guo received B.E. and M.E. degrees from Beijing University of Posts and Telecommunications (BUPT), China in 1982 and 1985, respectively, Ph.D. degree from the Tohuku-Gakuin University, Japan in 1993. At present he is a professor and a vice president of BUPT. His research interests include pattern recognition theory and application, information retrieval, content based information security, and network management. He has published over 200 papers, some of them are on world-wide famous journals or conferences including SCIENCE, Nature Scientific Reports, IEEE Trans. on PAMI, IEICE, ICPR, ICCV, SIGIR, etc. His book "Network management" was awarded by the government of Beijing city as a finest textbook for higher education in 2004.