Orientation Mining-Driven Approach to Analyze Web Public Sentiment

Feng Zhao, Qianqiao Hu
School of Computer Science and Technology, Huazhong University of Science and Technology, Wuhan, China
Email: zhaof@mail.hust.edu.cn

Xiaolin Xu, Runxi Zeng
College of Public Administration, Huazhong University of Science and Technology, Wuhan, China
Email: xiaolin@mail.hust.edu.cn

Yi Lin
Department of Mathematics, Nanjing University, Nanjin, China

Abstract— **In recent years, Internet provides a unique opportunity to express and spread public sentiment, which makes the web contents becoming the largest information source of public sentiment. Since web public sentiment reflects people's attitude to society and politics, the public opinion's orientation is significant to decision-makers. In this paper, we utilize VSM (vector space model) to present the text orientation of web information and offer data-mining approaches to analyze public opinion's orientation, which can assist decision-makers to steer social information and guide the web public sentiment. To achieve the goal of text orientation analysis, two ways are proposed. Firstly, a novel text orientation analysis method is described to analyze the orientation of original web postings and their replies. Secondly, an improved single-pass clustering algorithm is introduced to cluster the subject of web discussion and discover the hot topics. We also construct a prototype system, named WPSAS (web public sentiment analysis system), as experimental platform to validate the presented methodology. The experimental results show that our methods are effective and efficient.**

Index Terms— **web public sentiment, orientation analysis, opinion mining, clustering, VSM**

I. Introduction

Public sentiment is the complex synthesis of people's emotions, willing, attitudes and views. When it expresses the consistent view of the majority of people, it becomes public opinion [1]. Public opinion may be partial or impartial, also may be of fairness or prejudice. Once public sentiment becomes public opinion, it will do powerful impact to society. Web public sentiment is the public sentiment's extension on network/Internet. The goal of web public sentiment analysis is to discover the orientation of opinion before it evolves to negative public opinion or public emergency, which greatly influence real-world society. With the development of Internet, web public sentiment analysis becomes one of the hot topics in the fields of social computing, data mining and public administration.

Web public sentiment activity can be divided as Internet news and Netizen's discussion. Since most of Internet news is the extension of tradition medium, which is strictly checked by administrators, it cannot completely reflect the public opinion, which is not appropriate to take it as the collecting source of web public sentiment. On the contrary, Netizen's discussion is freely published and spread on network/Internet in real-time, which embodies the public opinion properly. As the most popular discussion platform, Bulletin Board System (BBS) becomes one of the most important approaches for Netizen to express their sentiment on Internet/network. Comparing with other sources, such as blog, instant messenger, etc., BBS has obvious advantages on public sentiment expressing. Considering the increased focus on such type of web contents, in this study, we take BBS as the data source to analyze web public sentiment.

The development on web public sentiment analysis attracts researchers' attention. Some international conferences have been held to display the newest research, for example, Text Retrieval Conference (TREC), Special Interest Group on Information Retrieval (SIGIR), Topic Detection and Tracking (TDT). The most well-known study on web public sentiment analysis system is TDT project [2], whose original intention is to design algorithms to detect and summarize useful information from web data flow. It has five basic study missions: story segmentation, topic tracking, topic detection, first-story detection and link detection. In China, many achievements about Chinese Natural Language Processing (CNLP) can be obtained as an open platform to assist public sentiment analysis system's design and development [3].

Text orientation analysis is one of the most popular approaches to realize web public sentiment analysis [4], [5], [6], which can be divided into three kinds: semantic-based text orientation analysis, machine learning-based text orientation analysis, and emotional word statistic value-based orientation analysis.

There are two methods to analyze semantic-based text orientation. One method is to build a semantic orientation pattern library, match semantic pattern referencing to

doi:10.4304/jsw.6.8.1417-1428

the library, and accumulate all orientation value of the matched pattern to get the whole text's orientation value. Yi and his partners proposed a sentiment analyzer *SA* (Sentiment Analyzer) [4]. *SA* uses the grammar parser to parse sentence grammar, uses the emotional vocabulary and emotional semantic pattern library to mine semantic relationship of the sentence, and extract views and opinions on the event from the event reviews. The other method is simplified from the above one, which takes words from text reflecting the subjective feelings, then determines the emotional words' orientation and gives an orientation value, and finally puts the cumulative value of all words' orientation to be the text's orientation value. Turney et al use *Pointwise Mutual Information and Information Retrieval* (PMI_IR) method to estimate the similarity between the word and the basic words of different emotion position (such as *good* and *bad*) [5]. Yanlan Zhu et al make use of semantic similarity and semantic field calculation function provided by *HowNet* to calculate the relevance between the assessed words and the pre-selected based-words, which gives result of words' orientation value [7].

Machine learning-based text orientation analysis considers the semantic orientation of texts as two types, which build classifier to help classifying text by artificially marked training samples. Pang et al uses Naive *Bayes*, maximal entropy, SVM methods o verify the effect of the semantic orientation of the text [8], [9]. The aim in their work is to examine whether it suffices to treat sentiment classification simply as a special case of topic-based categorization (with the two "topics" being positive sentiment and negative sentiment), or whether special sentiment-categorization methods need to be developed. And their experimental results show that SVM has the highest classification accuracy. But machine learning-based text orientation analysis method does not perform as well on sentiment classification as on traditional topic-based categorization since sentiment should be expressed in a subtle manner instead of in an identifiable manner by keywords.

Emotional word statistic value-based orientation analysis treats text as a collection of words in isolation to decrease the computational complexity, which does not take relationship between these terms into account [10], [11]. For example, there is a document liking "a shameless female university student made the posting, a very good gentleman's reply". Literally, the original poster's attitude is to support the gentleman and oppose the female student. From replying postings, we can see that a considerable number of people praise the gentleman and other people abuse the female student. Their views are the same as the original posting. However, emotional word statistic value-based orientation analysis classifies the replying postings into supporting the gentleman in favor of the original posting and abusing the female student against the original posting, because this method cannot distinguish the target of the emotional words.

These three methods have their own advantages and disadvantages, of which the semantical pattern matching-based method and machine learning-based approach have better accuracy, recall rate and less generality than emotional word statistic value-based method. Comparing with emotional word statistic value-based method, semantical patterns matching-based method requires semantical patterns' extraction in advance bring great workload and machine learning-based approach needs to artificially mark training samples for every topic in addition.

However, traditional forms of text orientation analysis mentioned above may not be effective for web contents in BBS forums. Firstly, only a few sentences contained topical information, it is difficult to excavate public opinion from numerous web contents according to topical analysis. Secondly, web contents of BBS are in natural language format. In addition to containing Netizen's discussion, web contents of BBS also imply a certainly behavioral orientation of Internet users and include Netizen's feelings and emotions. Thus, public sentiment analysis requires more understanding than the usual text orientation analysis technology. Being so, it is a challenge to do effective analysis on the BBS forums and provide more credible results for the decision of Internet management department.

To achieve this goal of perfect sentiment analysis and gain a better result on identifying and analyzing opinions and emotion, in this paper we propose the application of text orientation analysis techniques combing data mining technology to handle BBS postings. Our work emphasizes on the following approach: Firstly, VSM (vector space model) is utilized to present the text orientation of web information. The orientation of the text can be expressed by a vector, rather than support, opposition or neutrality. Items in vector are noun extracted from text keywords and item's weight is the keyword's semantic orientation value in text, which is cumulatively related to the emotional words orientation value. Then, a novel text orientation analysis method is described to analyze the orientation of original web postings and their replies. Thirdly, an improved single-pass cluster algorithm is introduced to cluster the subject of web discussion and discover the hot topics. We also construct a web public sentiment analysis system as experimental platform to validate the presented methodology.

The main contributions of this study are as follows: Firstly, an improved weighted VSM-based model is presented for the web public opinion activities, and it can describe both the state and the characteristic of the web public opinion more effectively. Based on the information, the statistics can easily obtain the number of replying postings with the main views expressing the fully support, partially support, partially opposite, fully opposite and neutral. Secondly, a novel text orientation analysis algorithm for BBS forums is proposed, which can achieve better performance by combining semantic mining and emotional analysis. It takes the relationship of network activities in account when executing mining. Thirdly, an improve single-pass clustering algorithm is proposed

TABLE I.
NOTATIONS AND THEIR DESCRIPTIONS

Notation	Description
T	Topic
$\|T\|$	Number of original posting which contained in the topic T
D	n-dimensional weighted vector space
S, P	Variable of temp set
d	Domain
w	Weight
Q	Web posting
sim	Cosine-correlational similarity
tf	Term frequency
M	Number of collected topic
m	Number of topics which contain d
e	Parameter of emotional strength
i, j, n	Numerical variable

based multi-comparing, it can be applied for clustering hot topics from uneven distribution of categories and avoiding the dependency on the postings order in BBS forums.

The rest of the paper is organized as follows: Section 2 introduces the data mining framework of web public sentiment analysis including basic concepts, the improved Vector Space Model and the processing flow; Section 3 describes a novel text orientation analysis method based on emotional strength to mine web public sentiment; Section 4 presents an improved single-pass clustering algorithm to cluster the subject of web discussion and discover the hot topics; In section 5, we develop a prototype system-WPSAS (Web Public Sentiment Analysis System) as experimental platform and show the experimental results; Section 6 surveys the related work and section 7 details the conclusion.

II. FRAMEWORK OF WEB PUBLIC SENTIMENTAL ANALYSIS BASED ON DATA MINING

To describe the framework and make our study more clearly, in this section, we first introduce some basic concepts and give all notations used in this paper as shown in Table I.

Definition 1: Document is a separated message unit that appeared on the Web pages, BBS, Blog, etc.

Definition 2: Text is the literal information in a document.

Definition 3: Topic, which includes document, is the discussion focus for groups of person.

Definition 4: Web posting is a document which is found on the web, but is not part of an online journal or book. The first web posting about a certain topic in BBS forums is original posting, other web postings focus on the same topic replied by Netizen are replying postings.

Definition 5: Web public opinion is a set of topics on the Internet which share a similar essential focus.

Definition 6: Orientation of the web public opinion describes how web posting is used to adopt or express an attitude of some kind towards some topics, such as positive or negative.

A. A Brief of VSM

VSM is firstly brought forward by Sahon G. in 1975, and applied to the text index and text representation recently. The vector approach allows for a mathematical and a physical representation using a vector space model. Each processing token (word) can be considered another dimension in an item representation space [12], [13]. There are two approaches to the domain of values in the vector: binary and weighted. Under the binary approach, the domain contains the value of one or zero, with one representing the existence of the processing token in the item. In the weighted approach, the domain is typically the set of all real positive numbers. The value for each processing token (term) represents the relative importance of that processing token in representing the semantics of the item. In this paper, we choose weighted-VSM to describe the web posting contents.

There are numerous reasons for our preference for weighted-VSM, and we shall here explore only a few of most important ones. One chief reason is that this model is extremely convenient to calculate similarity between two texts, which is very important to web public sentiment analysis. The method used to calculate similarity between document vectors includes inner product, cosine similarity, correlation distance, spearman distance, *Euclidean* distance and *City Block* distance [14], [15], [16], [17]. In this paper, we use the common *cosine-correlation* similarity to calculate the similarity between original posting and topic corresponding to the weighted-VSM.

Based on weighted-VSM, the content of a web posting is composed by a set of topic-specific feature terms and the association between topic, features and sentiment. However, the feature term exaction should satisfy some relationships, such as *part-of* relationship, *attribute-of* relationship etc. Let D be n-dimensional weighted vector space represented as D=<(d_1, w_l); (d_2, w_2); ..; (d_n, w_n)>, terms are axes of the space, documents are points or vectors in this space. w is a weighted value fixed and represented a particular domain d, w=0 if a term is absent from vector space. Thus, web posting can be structured as ¡<topic| feature, ..., feature>, where features are limited within the scope of D. Consequently, for a given topic, the web posting is an example of D.

Suppose the original posting is Q=¡<w_{q1}, w_{q2}, ..., w_{qn} > the reply posting is D_j=¡<w_{j1}, w_{j2}, ..., w_{jn} >, where D_j is the vector for a given topic T_j, Q is the vector of original posting needing to be compared, w_{ji} is the weight of vector item i of topic T_j, w_{qi} is the weight of vector item i of original posting. If term weights are normalized, the *Cosine-correlation* similarity between Q and D_j is described as [18], [19]:

$$sim(Q, D_j) = \sum_{i=1}^{n} w_{qi} * w_{ji} \ . \quad (1)$$

Otherwise, similarity is defined liking [13], [18], [19]:

$$sim(Q, D_j) = \frac{\sum_{i=1}^{n} w_{qi} * w_{ji}}{\sqrt{\sum_{i=1}^{n} (w_{qi})^2 \times \sum_{i=1}^{n} (w_{ji})^2}} \quad . (2)$$

When using weighted-VSM to describe the sentiment topics, the first step is to extract keywords and emotional words to generate axes of vector space, the second step is to calculate weight of every domain. In traditional document mining, VSM is combined with term frequency (TF) as weight to compute similarity [12]. Assume that tf_{qi} is the term frequency of term d_i in original posting Q, term weight of d_i in D is determined by w_{qi}=$tf_{qi} \times log(M/m_i)$, where M is the number of collective topics, $m_i$is the number of topics which contain d_i. The w_{qi} assigns high weights to terms that appear frequently in a small number of documents in the document set. However, to analyze web public sentiment, TF is too simple to calculate the weight of Internet opinion. Firstly, TF only denotes how often a term occurs in the document and never takes the document collecting process into account; Secondly, the appearance of a term cannot wholly represent the importance of the keywords in the document

So, we improve the TF computation by taking into account the Netizen's emotion in web postings. Thus, the weight of a term is determined by three factors: how often the term d_i occurs in the topic T_j (the term frequency tf_{ji}) how often it occurs in the whole document collection and how strong it presents the emotion. Precisely, the weight of a term d_i of topic T_j for a given topic set T is:

$$w_{ji} = \sum_{T} (tf_{ji} \times \log \frac{M}{m_i}) / |T| \times e_{ji} \quad , (3)$$

where $e_{ji} \in (0, 1)$ is the parameter describing emotional strength, $|T|$ is the number of original posting which contained in the topic T. Comparing with [13] and [19], the most significant advantage of Eq.(3)is that it describes the textual information as natural language, and the important terms are automatically calculated from documents specifying Netizen's emotional expressions. Comparing with [18], Eq.(3)enriches the Cosine-correlation similarity as a metric due to that it satisfies the discrete and uncertainty of textual information.

B. Flow of Web Public Sentiment Analysis

Based on weighted-VSM, to analyze orientation of web public sentiment, two data mining technologies are introduced: text orientation mining and clustering. In this paper, text orientation mining is used to design a novel text orientation analysis method based on a matching container to analyze the orientation of original web postings and their replies. The matching container, like a sandbox, encapsulates the web data, data mining algorithms and assistant programs for providing a tightly-controlled set of resources for analysis programs to run in. A matching container only associates with one given topic. Using matching container, a topic is encapsulated, which simplifies the recognition and statistic process. Clustering is used to get hot topics from numerous web

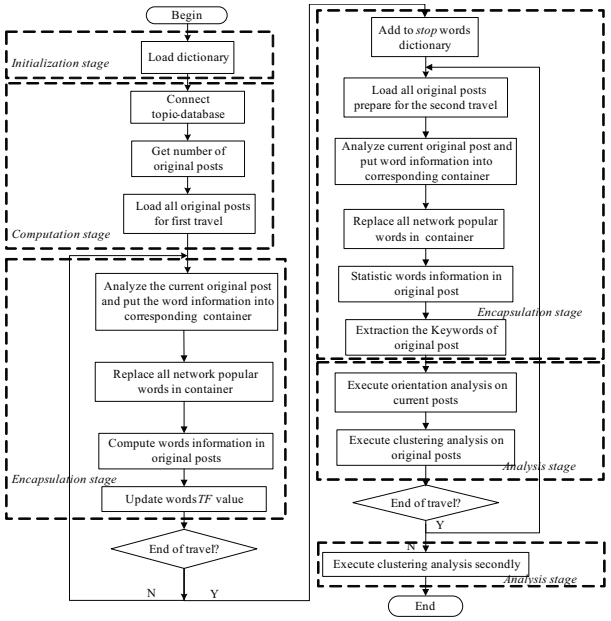

Figure 1. Flow of Web Public Sentiment Analysis

postings. Figure 1 shows the workflow of web public sentiment analysis based on data mining approaches.

The whole flow can be classified as the following stages:

Initialization stage: Load the necessary dictionary for analysis process, such as emotional words dictionary, negative words dictionary, stop words dictionary, network popular words dictionary, et al. These preparing dictionaries are to limit the analysis process and improve the accuracy. For example, some words emerging in many topic texts are not sensitive to distinguish topics, but their weights are relatively large. These words are easily extracted as vector items and clustered into a big category. The stop words dictionary is used to prevent these words from being described as vector items. Usually, if the topic words' TF is greater than 20% of the total num of original postings, they are added in the stop vocabulary to forbid them to be described as the vector items.

Computation stage: Connect the database, which stores the web postings and their statistical data, and travel all original postings to get words' term frequency. These data are used to calculate initial weight of original postings. In this step, there are two kinds of words needed to be paid attention. One is those words whose TFs are greater than a specified value. They are added into *stop* vocabulary used to filter keywords from texts. The other is those words which appear most frequently in topic texts. They are against to the topic distinguish and need not to be treated as keywords, since these words may be modal words or idioms. These words are garbage to extract valuable features in the process of constructing VSM model.

Encapsulation stage: Construct container according to the specified topics, which includes all necessary data and module for data mining on these topics, such as keyword dictionary, analysis module, et al. In the container, it is easy to travel all original postings secondly and execute

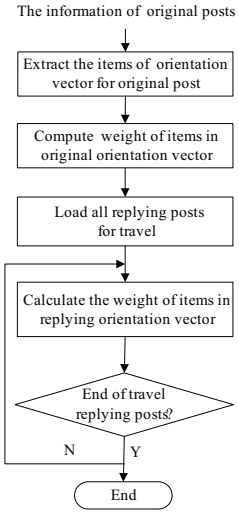

Figure 2. Process of Text Orientation Mining

orientation analysis on current original postings and their replying postings.

Analysis stage: Execute text orientation mining and clustering to discover orientation of web public sentiment. In this stage, the data mining engineer may exchange information with other three stages. For example, send the original posting into topic clustering module to get hot topics. In this step, some original postings with middle threshold are re-clustered, which are missed to be clustered since their maximum similar thresholds are not up to the classified value and greater than the value of establishing new topics.

III. Text Orientational Mining based on VSM

Text orientation mining is utilized to get public opinion of original posting and its replying postings. Comparing traditional text orientation mining methods, our approach is outstanding in two ways.

Firstly, traditional statistic-based mining method [19] treats text as a collection of isolated words, which does not take the involved event into account. In this paper, we combine semantic mining and emotional analysis to discover web public opinion. We use weighted-VSM to represent the text's orientation, such as neutral, support, against, et al. Items in vector come from web keywords which is extracted according to their TF, and the weight is calculated according to formula (3) which takes the text emotion into account. In this study, those words whose TF are within a user-defined bound are decided as keywords.

Secondly, in traditionally semantic pattern, <*body*> <*behavior*> <*receptor*> is the basic pattern [7]. For web public sentiment, the authors of different positions have various orientation, they often use emotional words for ¡*body*¿ or ¡*receptor*¿ to express their opinions instead of real actions. So, we use emotional words of <*body*> or <*receptor*> to distinguish semantic orientation without concrete action in <*behavior*>.

Suppose D=<d_1, w_l; d_2, w_2; ..; d_n, w_n> indicates the text orientation vector, where d_i is the selected term considered as keyword, w_i is the emotional weight of text term summarized from bias terms, n is the vector length describing the feature numbers. Figure 2 is the process of text orientation mining for web public sentiment analysis. Our methodology of text orientation mining is to build up weighted-VSM and compute emotional orientation value for web postings.

A. Orientation Vector Generation

The vector generation process can be stepped as feature extraction and weight computation.

To improve efficiency, we consider feature extraction from two aspects. On one hand, title nouns of postings are added into the vector items in priority, since the title usually contains the text's general idea and the original posting's attitude, that these nouns are likely to represent the focus of discussions. On the other hand, highly frequent terms in text are also treated as the vector items to prevent lacking information in headings. We utilize a sorting method to extract features from numerous words. For specified original postings Q and its replying postings, count the item appearance probability and select those items with top greatest probability as a preference for vector terms.

When domain $d$is ascertained, for original posting, the initial emotion is classified into three types: *neutral*, *support* and *against*. In order to facilitate the calculation, *neutral* is represented as 0, *support* is described as 1 and *against* is expressed as -1. Thus, original posting can be described as Q=$\{d_i, w_i | \ w_i \in \{1, 0, -1\}$ and $1 \le i \le n\}$.

Generally speaking, to get the emotional weight of the items in orientation vector, the steps are following: Firstly, intercept sentences from web postings orderly, and sum up orientation value of all emotional words in sentence. In this step, we can get the orientation values of sentences. Then, intercept separate segments from sentences treating comma as segmental note, and cumulate orientation value of all emotional words in segments. In this step, we can get the orientation values of segments. Finally, after completing orientation analysis of a sentence, add its orientation value to the whole text's orientation value.

After establishing orientation vector, the attitude of replying postings to original posting can be easily obtained by matching the item weights of vector. Instead of *neutral*, *support* and *against*, the attitude is extended to the following five types according to the matching result:

Fully support: Orientation vector of replying postings is fully consistent to orientation vector of original posting;

Partially support: Orientation vector of replying postings is partially consistent to orientation vector of original posting, and there is no contrary weight on the same item, such as -1 to 1;

Partially object: there is contrary weight on the same item between orientation vector of replying postings and orientation vector of original posting;

Fully object: Every item weight of orientation vector of replying postings is opposite to orientation vector of original posting;

Neutral: All item weight of orientation vector of replying postings is 0.

B. Orientation Mining Algorithm

Based on text orientation vector, we propose a novel algorithm to analyze public opinion and emotion in this paper. The following is the implemental description of our algorithm.

Step 1: Generate orientation vector of web posting, we implement it by function GetVSM (D).

Input: Original posting and its replying postings text ranged by textual words sequence.

Output: Orientation vector of web posting.

num= the total number of words in postings;
fen= 0; // weight of sentence orientation
all_fen=0; // weight of text orientation
i = 0; $S = \emptyset$
; // initialization
while ($i \leq$ *num*)
{
word = ith word of current posting;
$S = S \cup$ *word*;
if (*word* $\in${"?", "!", ""} or (i= *num*))
{ *fen* = SentenceAnalysis(S); // go to step 2
all_fen = *all_fen* + *fen*;
GetComma (S, *fen*); // go to step 3
} }

Step 2: Generate sentence orientation vector, we implement it by function SentenceAnalysis(S).

Input: Set of words in sentences or segments.

Output: Orientation value of sentence or segments.

num= the total number of words in S;
fen=0 ; i = 0;
while ($i \leq$ *num*)
{ *word* = ith word of current posting;
if (*word* $\in$ emotional words dictionary)
fen= CheckNegativePrefix(S, *word*)$\times$orientation value of *word* in the emotional words dictionary which is labeled by 0, 1 or -1; // go to step 5
return(*fen*);
}

Step 3: Compute orientation value of segment identified by comma, we implement it by function GetComma (S, w).

Input: Set of words in sentence S and the sentence orientation value w.

Output: Orientation value of segemt.

num= the total number of words in set S;
fen=0 ; i = 0; $P = \emptyset$
; // initialization
while ($i \leq$ *num*)
{ *word* = ith word of current posting;
$P = P \cup$ *word*;
if ((*word*= ",") or (i=*num*))
{ *fen* = SentenceAnalysis(P);
if (*fen*=0) *fen* = w; }
}
SetNounWordWeight(P, *fen*);
}

Step 4: Compute weight of noun, we implement it by function SetNounWordWeight(P, *fen*).

Input: Set of the words in segments P and the orientation value of segment.

Output: Weight of noun

if (the item exists in P)
sum up *fen* to the weight of item;

Step 5: Check negative prefix of word, we implement it by function CheckNegativePrefix(S, i).

Input: Set of the words in sentence or segment S and the word's location i.

Output: if it has negative prefix return -1, else return 1.

j = 0;
while ($j \leq i$)
{ *word* = the jth word;
if ($j \in$ negtive words dictionary) return -1;
else return 1;
}

Step 6: Match orientation vector of original posting and its replying postings to get public attitude on a specified topic.

IV. Hot Topic clustering

Clustering in web public sentiment analysis is to discover hot topics from numerous postings in web forums. Although approaches to cluster unspecified topics are related, the most important method is single-pass clustering [20], a heuristic cluster method, in that it is under the constraint of limited main memory meeting the requirement of fast calculation and small storage space for hot topic clustering. According to a definite order, single-pass clustering method uses the first text as the basis to cluster, the rest texts are sequentially compared with their similarities. If the similarity reaches the specified threshold, the texts are clustered into a group, and their features are recalculated as the basis for matching with other texts; Or else, textual information is directly extracted as basis of a new kind of topic. All texts are clustered abiding by this rule sequentially. The computation complexity of single-pass clustering is O (nk), where k is the number of the class. However, for web public sentiment analysis, there are two shortcomings on single-pass clustering. First, this method is greatly dependent on the order, that clustering on the same object by different order will get different result. The other is that it is easy to be tendentious to some big category where clustering, because of the uneven distribution of categories.

To avoid these disadvantages, we improve single-pass clustering algorithm and propose a new clustering algorithm based multi-comparing. The clustering steps are as follows:

Step 1: Extract keywords from the original postings and get the word-document frequent information. Then, construct original posting vector and calculate vector

weight according to formula (3). Finally, return those words that have higher weights.

Step 2: Load and travel existed topic database, if it is empty, establish new topic only containing the current original posting and construct the VSM for topic. If it is not empty, travel topics and continue to next step.

Step 3: For every current topic i, compute its similarity to original posting according to formula (1) or (2). According to the calculation result, it is easy to get the most similar topic to the original posting with maximal similarity probability.

Step 4: This stage is a multi-clustering step. Firstly, compare similarity probability of topic i with a specified limit. If the similarity is less than the lower limit, build a new topic based on original posting and give its VSM description; if the similarity is greater than the higher limit, group original posting into this topic and update the topic vector. Otherwise, estimate whether or not it is the first time to cluster, if it is not, compare similarity probability of topic i with a specified value and do the same work with above cluster. In other words, if the similarity is less than the value, build a new topic based on original posting and give its VSM description; if the similarity is greater than the threshold value, group original posting into this topic and update the topic vector. When all topics are travel, go to next step, otherwise, go to step 3.

Step 5: Group topic into a class, set the topic identify and refresh the topic vector.

An example is shown in Figure 3, whose first limit extent is set as [0.4, .07], the second threshold value is equal to 0.55.

Comparing with single-pass cluster method, our approach improves accuracy by secondary clustering. Since single-pass clustering is obviously dependent on the data order, some original postings with middle threshold are re-clustered in second time instead of being clustered in first time. In theory, the time of clustering is infinite, but the more times, the lower efficiency of computation. Thus, in this paper, we only cluster topics in two times, that the original postings not reaching corresponding threshold is compulsively clustered in second time.

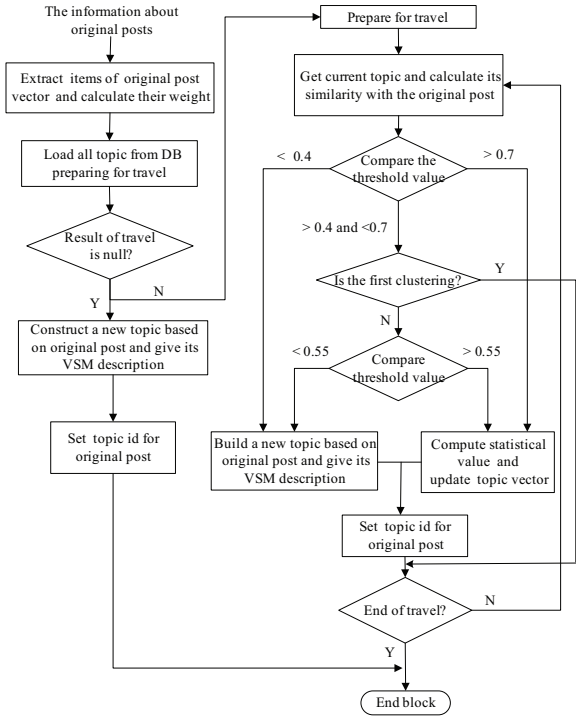

Figure 3. Example of Hot Topic Clustering

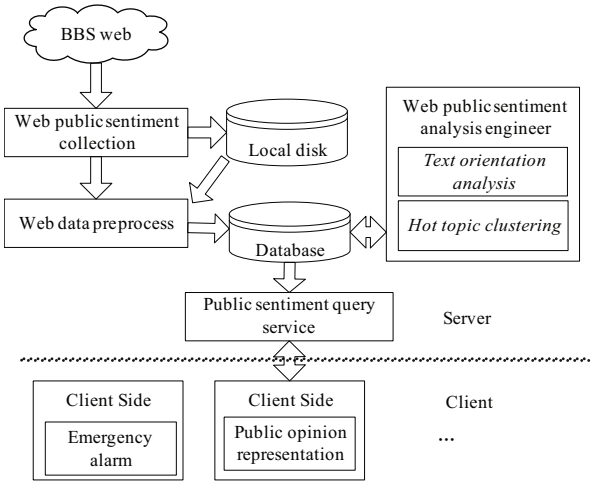

Figure 4. System Architecture of WPSAS

V. DESIGN OF WPSAS AND EXPERIMENTAL RESULTS

A. System Arichitecture

The architecture of WPSAS is shown in figure 4. Generally speaking, the component of WPSAS is mainly including public sentiment collection, web data preprocess, emotional orientation analysis, hot topic clustering and public opinion representation, et al. WPSAS is composed by server-side and client-side, where data collection and data analysis are in server and public opinion representation is in client.

Web public sentiment collection: The primary function of this module is to use the web crawler to download web pages from BBS. Most of them are html file. In WPSAS, we use powerful *Heritrix Crawler* [21], which is an Internet Archive's open-source, extensible, web-scale, archival-quality web crawler, to get original data. For a specified task, some condition is set to limit the search scope. This module stores downloaded HTML file in local disc on server to prepare for preprocessing on web postings.

Web data pre-process: This module is to analyze download HTML file, extract useful information such as information of the original posting and replying postings, and save this data into database. This module uses *nekohtml* and *xerces* to change HTML file to a DOM tree. So, we can use Node, Document, Element, Text and other interface provided by package *org.w3c.dom* to extract title, author, text, date, URL and other information and save them into database. The information of texts and segments is also saved into database and prepared for

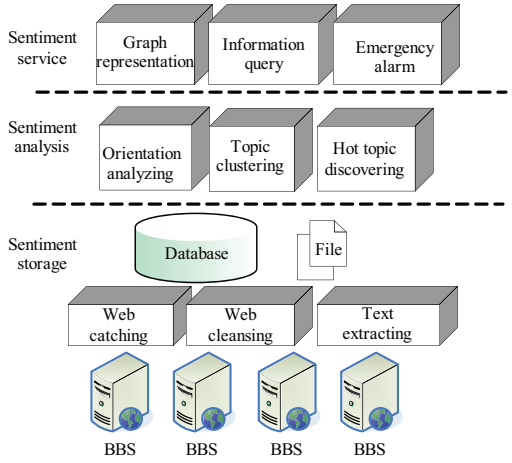

Figure 5. Functional Layers of data in WPSAS

public sentiment analysis module, which is developed by system *ICTCLAS* [22].

Web public sentiment analysis engineer: This module includes two functions: text orientation analysis and hot topic clustering. The function of this module is to analyze web postings as follows: Firstly, emotional orientation analysis is used to analyze the original posting's orientation and its replying postings' orientation; Then, cluster the original postings and classify them into the corresponding topic, which describes the same matter or activity; Finally, deposit the analysis results into database prepared for query and analysis of public sentiment service module.

Web public sentiment query service: The main function of this module is to obtain relative information through query in the database according to different network events.

Public opinion representation: This module is an interactive interface for different types of users, which describes the results of analysis engineer feeding back to user in graphs or tables.

Emergency alarm: This module is to forecast and give alarms of the emergency, which should threat real-world's public security.

B. Data flow of WPSAS

From functional view, data in WPSAS can be classified into three layers shown in Figure 5. In sentiment storage layer, raw web pages are obtained, cleansed and extracted into database; In sentiment analysis layer, the original data is analyzed by data mining approaches including emotional orientation mining, topic clustering and hot topic discovering; In sentiment service, application-faced data is represented to users according to their requests. Figure 6 is data flow of WPSAS. *Web public sentiment collection* collects web source pages from designated web forum and transforms them into HTML files, then *web data pre-process* filters needless pages and extracts the unstructured data into text orientation vector, thirdly *web public sentiment analysis engineer* excavates interested

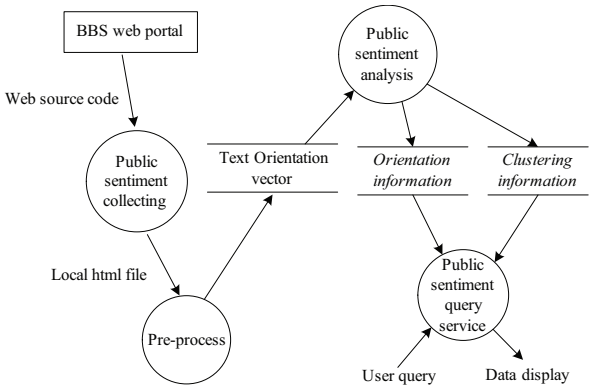

Figure 6. Data Flow of WPSAS

orientation and clustering information from text orientation vector space, *public sentiment query service* obtain these useful information and respond users' request, finally *public sentiment query service* gives different types of query result to users.

C. Chinese word segmentation

A primary hypothesis of WPSAS is that Chinese words in web pages of web forum can be easily segmented and extracted into VSM. Therefore, WPSAS can get structured and statistical information as data sources for analysis engineer. WPSAS adopts famous *ICTCLAS* (Institute of Computing Technology, Chinese Lexical Analysis System) to achieve goal of Chinese word segmentation, which is developed by Institute of Computing Technology of Chinese Academy of Sciences and its jar package can be download from their web site.

The main features of *ICTCLAS* include Chinese word segmentation, part-of-speech tagging, named entity recognition, new words identification and user dictionary. The latest version of *ICTCLAS* is *ICTCLAS3.0* and *ICTCLAS* 1.0 is open source and freely downloaded. Using *ICTCLAS*, the speed of Chinese word segmentation is up to 996KB/s, accuracy of segmentation is up to 98.45%, API library dose not exceed 200KB, all kinds of dictionary is less than 3M after compression [22]. *ICTCLA* is the best Chinese lexical analyzer and this is why we choose it in WPSAS.

Another focus of Chinese word segmentation is named entity, which refers to text with a specific entity including human being's name, place names, institutions, and special nouns, et al. Under normal circumstances, named entity plays essential role on different event distinction, such as swine flu, *Wen chuan* earthquake, Tibet terrorism, Sinkiang violence et al. General speaking, task of named entity recognition is to identify text to be processed in three categories (entity class, time class and number class), and in seven categories (name, affiliation, address, time, date, currency and percentage). In this paper, named entity recognition is also implemented by *ICTCLAS*, which is used to extract words from web texts in order to construct VSM to describe the characteristics of web pages.

TABLE II.
WEB FORUM COLLECTION STATISTICS

Forum	Authors	Topics	Replies	Time
Tencent	173	1874	15,457	2009/5
Tianya	77	284	30866	2009/5

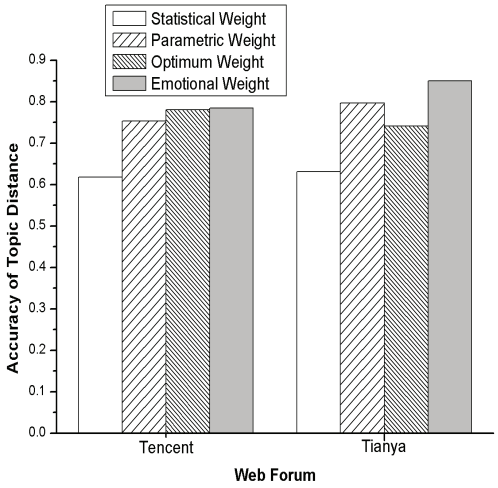

Figure 7. Comparison on Accuracy of Topic Distance

D. Experimental Results

To validate our methods, we implement WPSAS on DAWNING3000A and 2 personal computers (*Intel Core™2 Duo CPU T8100* with *2.10GHz*, and 2GB memory). The programming environment is JAVA and MySQL, and the tools of system development including Eclipse3.1 (jdk1.6) and Tomcat 6.0.10. Operating system of client is *Windows XP*. DAWNING3000A is used as web public sentiment analysis engineer and PCs are used as clients to submit request and represent excavated results. In this paper, two Web forums are selected for our public sentiment analysis, *Tencent* forum (http://bbs.news.qq.com/b-1000083746) and *Tianya* forum (http://www.tianya.cn). A summary of the collection statistics is presented in Table II.

We firstly compare the accuracy of topic distance using the weight in Eq.(3) with statistical weight [23], optimum weight [18] and parametric weight [19] to validate the performance of proposed method. In this experiment, the distance is small when the two vectors are from the same topic and larger when they are each from a different topic. We split web postings of *Tencent* forum and *Tianya* into separate training and test sets. 70% of the documents are added to training set and the remaining are in testing set. Figure 7 shows the comparison results on accuracy of topic distance for *Tencent* forum and *Tianya* forum. The experimental results show that our method is more efficient than statistical weight and parametric weight to analyze topic distance on public opinion. The accuracy of emotional weight method is much higher in that it takes Netizen's emotional strength into account, which is much helpful to distinguish two different topics. However, the overall accuracy of WPSAS is shown in Figure 8.

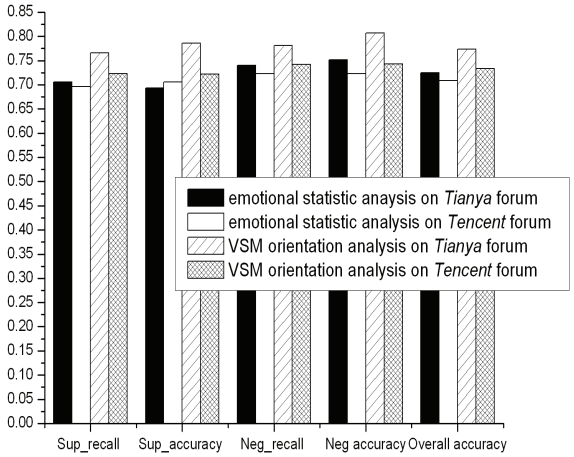

Figure 8. Comparison on Accuracy

Considering affect analysis in WPSAS, we compare the accuracy on our method and emotional words statistic-based analysis method [11] to validate the performance of distinguishing the target of the emotional words. Here, we use accuracy, recall rate as evaluation metrics. The accuracy denotes the ratio of the correctly analyzed number and analyzed number for web postings, which is the precision rate. The recall rate denotes the ratio of the number of relatively analyzed web postings and total number of web postings, which is the fully analyzing rate. As described above, we use 1 presenting positive view, -1 presenting negative view, 0 presenting neutral view in this experiment. Experimental comparison on accuracy between emotional words statistic-based analysis and VSM-based orientation analysis method, including recall rate of supporting postings (described as *Sup_recall*), accuracy of supporting postings (described as *Sup_accuracy*), recall rate of negative postings (described as *Neg_recall*), accuracy of negative postings (described as *Neg_accuracy*) and overall accuracy, is shown in figure 8. The experimental results show that efficiency of our method is better than emotional words statistic-based analysis method by improving accuracy, about 3% higher, on different types of postings.

Considering topic clustering in WPSAS, we compare the semantic similarity on our method and machine learning-based analysis method, as an example to *Bayes* probability-based clustering [20], to validate the performance on clustering the topics. In this experiment, we use *ful_support* presenting fully support, *part_support* presenting partially support, *ful_object* presenting fully object, and *part_object* present partially object. Figure 9 shows the comparison results on error rate of clustering similarity on *Tencent* forum and *Tianya* forum. The experimental results show that our method is more efficient than *Bayes*-based method to analyze public opinion according to two aspects. Firstly, error rate of VSM-based method is much more stable than *Bayes*-based method on different

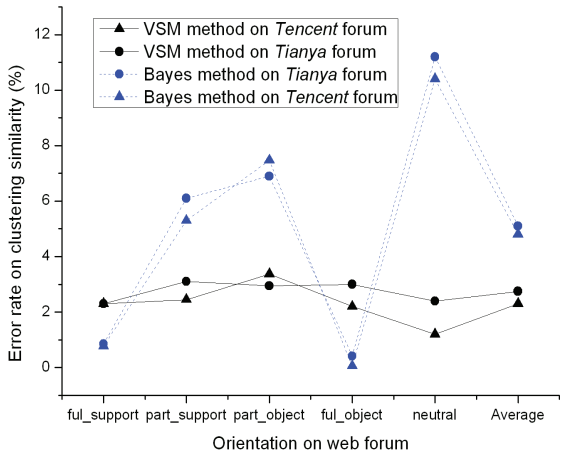

Figure 9. Comparison of Error Rate on Clustering

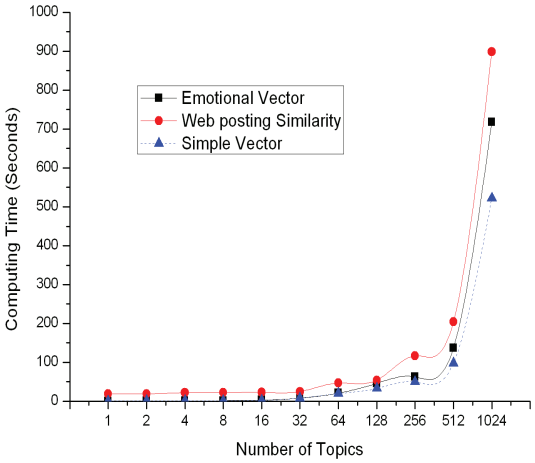

Figure 10. Time Cost for Vector Construction and Computing the Web Posting Similarity

text orientation. Secondly, the average error rate of VSM-based method is lower than *Bayes*-based method, about 2.5%.

Figure 10 shows time cost of manipulating the emotional vector to compute the web postings similarities, which exhibits the time costs of building the emotional vector from 1 topic to 1024 topics in *Tencent* forum, the time costs of computing the text similarities based on the emotional vector and time cost of computing simple text vector are also illustrated in Figure 10 as well. The experimental results show that the time cost of building an emotional vector is close to construct a simple text vector. Actually, the time cost on similarity computation is also very close to emotional vector construction, in that the hot topic clustering algorithm presented in this paper is a linear time complexity.

VI. Related Work

Within the abundant literature that exists in the context of public sentiment analysis and public opinion mining, several lines of related work have been addressed.

Sentiment classification attracts increasing attentions from researchers in recent years. Peter D. Turney proposes a simple unsupervised learning algorithm for classifying reviews as recommended (thumbs up) or not recommended (thumbs down) [5]. Dong Li etc. presents the problem of extracting contextual opposite sentiments in classified free format text reviews. They adapt the sequence data model to text mining with Part-of-Speech tags, and propose a belief-driven approach for extracting contextual opposite sentiments as unexpected sequences with respect to the opinion polarity of reviews [6]. Bo Pang et al take classification of a document or a sentence as expressing a negative or positive sentiment [8]. These works aim at mining useful information, for example topic/sentiment models, sentiment dynamics etc., from web text and aim at improving the classification accuracy. However, these works do not take the correlation between web postings and Netizen's emotion into account, which greatly affects the orientation accuracy.

Some researchers have been aware of this limitation and use another method, opinion clustering, to discover network public opinion. For sentiment clustering, Bo Pang et al use rating scales to exploit class relationships for sentiment categorization [9]; Huaizhong Kou etc. propose a new mathematical model to estimate the association between terms and define a ε-similarity model of documents, which greatly improves the effectiveness of clustering [14]. For document clustering, an efficient phrase-based document similarity for clustering is proposed in [15]; some other similarity-based clustering is also presented, such as term semantic analysis in XML [16], Wikipedia-based conceptual contexts [17], and discriminate analysis [18] etc.. However, performance of these clustering methods depends on the training data of public topics.

Feature extraction technology is another line of research in excavating public sentiment, which is made up by two successive steps: features identification and opinion detection [4], [6], [24]. Ahmed Abbasi etc. developed the entropy weighted genetic algorithm for efficient feature selection in order to improve accuracy and identify key features for each sentiment class, and they also apply their methodologies to English and Arabic Web forum postings [25]. Songbo Tan presents an empirical study of sentiment categorization on Chinese documents, which is based on four feature selection methods [26].

To the best of our knowledge, emotional orientation mining on public sentiment analysis has not been addressed in existing work. However, there are some related works on identifying emotions in text mining. Carlo Strapparava etc. concern with the automatic analysis of emotions in text, which is annotated for six basic emotions: anger, disgust, fear, joy, sadness and surprise [11]. Xia Mao etc. give a rough set and SVM based approach is adopted to categorize text into four emotional classes, including happy, sad, anger and surprise [27]. Abbasi Ahmed etc. propose a support vector regression correlation ensemble method, named SVRCE, for

enhanced classification of affect intensities, which also compare several feature representations for affect analysis including learned n-grams and various automatically and manually crafted affect lexicons [28]. These methods are concerned with the analysis of text containing emotions and are helpful to our work.

VII. Conclusion

Network provides unique opportunities to express and spread public sentiment, whose orientation greatly affects the real-world society. In this study, we present data mining approaches to discover network public opinion. The contributions include the following aspects: First, a novel text orientation analysis method is proposed to analyze the orientation of original web postings and their replies; Secondly, an improved single-pass clustering algorithm is introduced to cluster the subject of web discussion and discover hot topics; Thirdly, a web public sentiment analysis system, named WPSAS, is designed and implemented as running platform to analyze public opinion of web forum and validate the presented methodology.

However, there is still some limitation in our experimental system. First, WPSAS does not take topic tracking into account. With the growth of web postings, the computation complexity of topic clustering algorithm increases so rapidly that the performance is relatively low. Furthermore, orientation analysis module does not deal with semantic fragmentation in text automatically. In future, we will focus on the following aspects: First, perfect relevant knowledge which uses semantic pattern matching-based methods to find emotional words and changeable items. Secondly, design parallel clustering algorithm to improve performance of topic clustering. Finally, take other part of speech into account, such as verbs et al, and add these terms into characteristic items to describe text orientation.

References

[1] Qiaozhu Mei, Xu Ling, Matthew Wondra, Hang Su, ChengXiang Zhai. "Topic Sentiment Mixture: Modeling Facets and Opinions in Weblogs". *In: Proceedings of 2007 International World Wide Web Conference* (WWW 2007), Banff, Alberta, Canada, May 2007. 171-180

[2] James Allan (eds.). *Topic Detection and Tracking: Event-based Information Organization*, Kluwer Academic Publishers, 2002

[3] Haibin M. Web public sentiment and analyzing technology. GuangMing Daily, 2007-1-21, 006 (In Chinese)

[4] Jeonghee Yi, Tetsuya Nasukawa, Razvan Bunescu, Wayne Niblack. "Sentiment Analyzer: Extracting Sentiments about a Given Topic using Natural Language Processing Techniques". *In: Proceedings of the third IEEE International Conference on Data Mining* (ICDM 2003), IEEE Computer Society Press, Los Alamitos, November 2003. 427-434

[5] Peter D. Turney. "Thumbs UP or Thumbs Down? Semantic orientation Applied to UnsuPervised Classification of Reviews". *In: Procedings of the 40th Annual Meetion of the Association for Computational Liguistic* (ACL 2002), Philadelphia, PA, USA, July 2002. 417-424

[6] Dong Li, Anne Laurent, Mathieu Roche, Pascal Poncelet. "Extraction of Opposite Sentiments in Classified Free Format Text Reviews". *In: Proceedings of 19th International Conference on Database and Expert Systems Applications* (DEXA 2008), Turin, Italy, September 2008. 710-717

[7] Zhu Yanlan, Min Jin, Zhou Yaqian, Huang Xuanjing,Wu Li-de. "Semantic Orientation Computing Based on HowNet". *Journal of Chinese Information Processing* , 2006, 20(1): 14-20 (In Chinese)

[8] Bo Pang, Lillian Lee, Shivakumar Vaithyanathan. "Thumbs up? Sentiment Classification Using machine Learning Techniques". *In: Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing* (EMNLP 2002), Philadelphia, USA, July 2002. 79-86

[9] Bo Pang, Lillian Lee. "Seeing Stars: Exploiting Class Relationships for Sentiment Categorization with Respect to Rating Scales". *In: Proceedings of the Association for Computational Linguistics* (ACL 2005), Michigan, USA, June 2005. 115-124

[10] P. Turney, M. Littman. "Measuring praise and criticism: Inference of semantic orientation from association". *ACM Transactions on Information Systems*, 2003, 21(4):315-346

[11] Carlo Strapparava, Rada Mihalcea. "Learning to Identify Emotions in Text". *In: Proceedings of 23rd Annual ACM Symposium on Applied Computing* (SAC 2008), Fortaleza, Cear'a, Brazil, March 2008. 1556-1560

[12] Dikl. Lee, Huei Chuang, Kent Seamons. "Document Ranking and the Vector-Space Model". *IEEE Software*, 1997, 14(2): 67-75

[13] Katrin Erk, Sebastian Pad'o. "A Structured Vector Space Model for Word Meaning in Context". *In: Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing* (EMNLP 2008), Waikiki, Honolulu, October 2008.

[14] Huaizhong Kou, Gardarin G.. "Similarity model and term association for document categorization". *Proceedings. 13th International Workshop on Database and Expert Systems Applications* (DEXA 2002), 2002

[15] Chim Hung, Deng Xiaotie. "Efficient Phrase-Based Document Similarity for Clustering", *IEEE Transactions on Knowledge and Data Engineering*, 2008, 20(9): 1217-1229

[16] Jianwu Yang, Cheung W.K., Xiaoou Chen. "Integrating element and term semantics for similarity-based XML document clustering". *In: Proceedings of 2005 IEEE/WIC/ACM International Conference on Web Intelligence*, Compiegne, France, September 2005. 222-228

[17] Kaiser F., Schwarz H., Jakob, M.. "Using Wikipedia-Based Conceptual Contexts to Calculate Document Similarity", *In: Proceedings of third International Conference on Digital Society* (ICDS 2009), Cancun, Mexico, February 2009. 322 -327

[18] Yong Ma, Shihong Lao, Erina Takikawa, Masato Kawade. "Discriminant Analysis in Correlation Similarity Measure Space". *In: Proceedings of the 24^{th} International Conference on Machine Learning* (ICML 2007), Corvallis, OR, 2007. 577-584

[19] Lebanon G.. "Metric learning for text documents". *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2006, 28(4):497-508

[20] Karkkainen I., Franti P.. "Gradual model generator for single-pass clustering". *In: Proceedings of Fifth IEEE International Conference on Data Mining* (ICDM 2005), Houston, Texas, USA, November 2005. 681-684

[21] Heritrix. http://crawler.archive.org/

[22] ICTCLAS. http://ictclas.org/

[23] Katherine A. Heller, Zoubin Ghahramani. "Bayesian hierarchical clustering". *In: Proceedings of 22th ACM International Conference on Machine Learning* (ICML 2005), Bonn, Germany, August 2005. 297-304

[24] J. Tipan Verella, Ahson Wardak. "Modeling Public Opinion and Voting as a Complex System with Agent-Based Simulations". *In: Proceedings of the 2008 IEEE Systems and Information Engineering Design Symposium*, Charlottesville, VA, USA, April 2008. 261-266

[25] Ahmed Abbasi, Hsinchun Chen, Arab Salem. "Sentiment Analysis in Multiple Languages: Feature Selection for Opinion Classification in Web Forums". *ACM Transactions on Information Systems*, 2008, 26(3), Article 12

[26] Songbo Tan, Jin Zhang. "An empirical study of sentiment analysis for Chinese documents". *Expert Systems with Applications*, 2008, 34(4): 2622-2629

[27] Xia Mao, Zheng Li, Haiyan Bao. "A Rough Set and SVM Based Approach to Chinese Textual Affect Sensing". *In: Proceedings of 2008 International Conference on Intelligent Systems Design and Applications* (ISDA 2008), Kaohsiung City, Taiwan, November 2008. 307-311

[28] Abbasi Ahmed, Chen Hsinchun, Thoms Sven, Fu Tianjun. "Affect Analysis of Web Forums and Blogs Using Correlation Ensembles". *IEEE Transactions on Knowledge and Data Engineering*, 2008, 20(9): 1168-1180

Xiaolin Xu is currently a Professor of Public Administration at Huazhong University of Science and Technology, China. He is the Dean of College of Public Administration of Huazhong University of Science and Technology. His current research interests include e-government, online sentiment and public administration etc..

Feng Zhao is currently an associate professor of Computer Science, Huazhong University of Science and Technology, China. He received his PhD degree in computer science from Huazhong University of Science and Technology in 2006. His research interests include data mining, security and distributed computing.