

BREM: A Distributed Blogger Reputation Evaluation Model Based on Opinion Analysis

Yu Weng

National Language Resource Monitoring Research Centre Minority Language Branch
College of Information Engineering
Minzu University of China, 100081, China
E-mail : mr.wengyu@gmail.com

Changjun Hu, Xuechun Zhang and Liyong Zhao

School of Information and Engineering
University of Science and Technology Beijing, 100083, China
E-mail: zlyong1981@163.com

Keywords: blogger reputation evaluation, opinion analysis, distributed computing

Received: November 21, 2009

As a booming virtual community platform, blogosphere has won more and more public attention and preference. For improving the social status analysis ability of blogosphere more effectively, a distributed blogger reputation evaluation model based on opinion analysis is presented (named BREM). The model not only evaluates the reputation level of blogger in the inner-network domain, but also cooperatively schedules the blogger reputation information among the inter-network domains. In the application process, BREM firstly tracks the variation trend of various factors (including the amount of reviews, comments and the published time), identifies the comments opinions of each topic, and evaluates the reputation level of blogger in the single blogosphere periodically. On the other hand, through cooperatively scheduling the local reputation information of bloggers among different blogosphere, the model extends the scope of reputation evaluation and manages the bloggers in the virtual social community more comprehensively. To validate the performance, the experiments on the data corpus about "Unhealthy Campus Culture" demonstrate that BREM has higher application validity and practicality of blogger reputation evaluation in distributed environment.

Povzetek: Razvit je model ocenjevanja ugleda blogov na osnovi mnenj.

1 Introduction

In the real society, people usually are classified into different groups by retrieving the personal information (such as age, sexual, job and etc.). However, due to the limits of user authority and personal privacy, the personal information of users could not be obtained freely and truthfully in virtual community-blogosphere. As a kind of novel analysis method, the reputation evaluation has been successfully applied in finance, insurance and the other domains. Using the reputation evaluation into blogosphere will group the virtual community users more effectively and provide the data support for various complex applications.

Recent years, lots of efforts have been made to the research of reputation evaluation. The common idea is to use the number of page links as the estimation of its reputation ^[1, 2]. S. Brin and L. Page ^[3] (1998) modeled the page links graph for the reputation computing, where vertices represent pages and edges represent the links

between pages. Klessius Berlt and Nivio Ziviani ^[4] (2007) proposed a representation of web pages and improved the page links hypergraph evaluation model by reducing the impact of non-votes links. Combined user's individual activity analysis approach and collaborative activity analysis approach, Fusheng Jin and Zhendong Niu ^[5] (2008) proposed a user reputation model and applied it to the DLDE Learning 2.0 community. Jennifer Golbeck and James Hendler ^[6] (2004) presented a voting based algorithm for aggregating reputation ratings on the Semantic Web. Some business companies ^[7, 8] also proposed the online reputation systems to rate and find the more potential customers.

Different from the traditional online reputation calculation methods which mostly focus on the individual activities, the reputation evaluation of blogosphere should give more emphasis on the social relations analysis of bloggers. By mining the comment opinion attitudes of other bloggers (e.g. positive,

negative or neutral), the blogger reputation status in the whole virtual community would be reflected. According to the scenario above, we present a distributed blogger reputation evaluation model based on opinion analysis (named BREM). The model not only evaluates the local reputation of blogger in the single blogosphere, but also cooperatively schedules the blogger reputation information in the other blogospheres. On one side, BREM analyses the semantic orientation (SO) of blog comments and tracks the opinion relations between bloggers. Two calculation methods for long text and short text are adopted respectively. For the long comment text, BREM calculates the SO weight of each character and the distribution density of opinion characters in target text. Through constructing the text opinion case base for long text, the model reuses the evaluation result of historical case and shortens the execution time effectively. For the short comment text, the text opinion is calculated by summing the SO weight of each character. Then, BREM tracks the supportive degree of blog topics and evaluates the reputation of blogger. On the other hand, the model schedules the reputation information of blogger in the other blogosphere periodically and improves the analytical ability of blogger reputation in distributed environment.

This paper is organized as follows. Section 2 outlines the previous approaches of opinion analysis. In section 3, some problems and the general process are described. In section 4, each part of BREM is presented in details. In section 5, experimental results on the corpus of “Unhealthy Campus Culture” are given. Finally section 6 concludes the work with some possible extensions.

2 Related works

With the rapid development of Web 2.0 technology, text opinion analysis is attracting more and more attention. Hatzivassiloglou and McKeown^[9] (1997) used textual conjunctions such as fair and legitimate or simplistic but well received to separate similarly and oppositely connoted. Pang^[10] (2002) classified the documents by sentiment analysis and showed that machine learning approaches on sentiment classification do not perform as well as that on traditional topic-based categorization at document level. Hu and Cheng^[11] (2005) illustrated an opinion summarization of bar graph style, categorized by product features. Soo-Min Kim and Eduard Hovy^[12] (2006) describe a sentence-level opinion analysis system. The experiment based on MPQA (Wiebe et al.^[13], 2005) and TREC (Soboroff and Harman^[14], 2003) showed that automatic method for obtaining opinion-bearing words can be used effectively to identify opinion-bearing sentences. Lun-Wei Ku, Hsiu-Wei Ho and Hsin-Hsi Chen^[15] (2006) selected TREC, NTCIR^[16], and some web blogs as the opinion information sources and proposes an algorithm for opinion extraction at word, sentence and document level. Ruifeng Wong and et al.^[17] (2008) Proposed an opinion analysis system based on linguistic knowledge which is acquired from small-scale annotated text and raw topic-relevant webpage. The system used a classifier based on

support vector machine to classify the opinion features, identify opinionated sentences and determine their polarities. Veselin and Claire^[18] (2008) presented a novel method for general-purpose opinion topic identification and evaluate the validity of this approach by the MPQA corpus. Table 1 shows the comparison of four methods of text opinion analysis.

These technologies above could be applied in the comments opinion analysis in single blogosphere successfully. However, since neglecting the blogger reputation influences of the other network domains, the applied scope and the precision of reputation evaluation would be affected sharply. Through cooperatively scheduling the blogger reputation information among the inter-network domains, BREM comprehensively considers the impacts of topic opinion in multi-blogosphere, strengthens the analysis ability of blogger reputation evaluation and improves the bloggers management level of the whole virtual social community

Table 1: Text Opinion Analysis Comparison.

Author	Method Description	Testing Results
Hatzivassiloglou ^[9]	Identifying the constraints from conjunctions on the positive or negative SO of the conjoined adjectives (e.g. and, but, either-or, etc.).	21 million words (Wall Street Journal) annotated with part-of-speech tags using the PARTS (Church, 1988). Accuracy: 82%.
Peter D. Turney ^[19]	The classification of a review is predicted by the average SO of the phrases in the review that contain adjectives or adverbs.	410 reviews from Epinions, sampled from domains (including banks, , movies, travel and automobiles). Accuracy: 74%
Lun-Wei Ku ^[16]	A major topic detection method is proposed to capture main concepts of the relevant documents. Then retrieving all the sentences related to the major topic, determining the opinion polarity of each relevant sentence, and summarizing positive and negative sentences.	TREC corpus, NTCIR corpus and articles from web blogs. TREC corpus is in English, the other two are in Chinese. Accuracy 40%
Soo-Minkim ^[12]	An approach of exploiting the semantic structure of a sentence, anchored to an opinion bearing verb or adjective. This model uses	2028 annotated sentences from FrameNet data set. (834 from frames related to opinion verb and 1194 from opinion

semantic role adjectives) and
labeling as an 100 sentences
intermediate step to selected from
label an opinion online news
holder and topic sources (New
using data from York Times and
FrameNet. BBC).
Accuracy: 47.9%

3 Problems description and general process

To evaluate the blogger reputation more reasonably, in the design process of BREM, the following three parts should be considered:

(1) Comment Opinion Analysis. The aim of comment opinion monitoring is to analyse the attitudes of reviewers to topics (e.g. positive, negative or neutral), and evaluate the blogger reputation more precisely. The calculation process, BREM considers the comprehensive influence of the length of text, the SO (Semantic Orientation) of characters, and the distribution status of opinion characters and identifies the SO of blog comments

(2) Blogger Reputation Evaluation. The reputation of blogger is the reflection of blogger social status in virtual

community. Through monitoring the amount of comments, reviews and the semantic opinion of blog comments, BREM could effectively analyse and calculate the supportive degree of the each blog topic and evaluate the reputation of the blogger.

(3) Blogger Reputation Cooperative Scheduling. In the virtual social community-blogsphere, bloggers could publish or comment the topic logs in different blogsphere freely. So the blogger reputation evaluation would be affected by the multi-network domains. BREM simulates the dynamic spreading process, schedules the local blogger information of other network domains and strengthens the blogger reputation analysis ability in the distributed environment.

In Figure 1, the general process of BREM is given. As the part of data preprocessing, firstly BREM analyzes the compositions of blog and represents them by Resource Description Frame (RDF) [20]. Through monitoring the semantic opinion of blog comment, BREM tracks the supportive ratio of the other reviewers to a specific blog topic and evaluates the reputation of bloggers. For improving the practicality of BREM, the model schedules the local blogger reputation information among different blogsphere periodically and manages the bloggers more effectively.

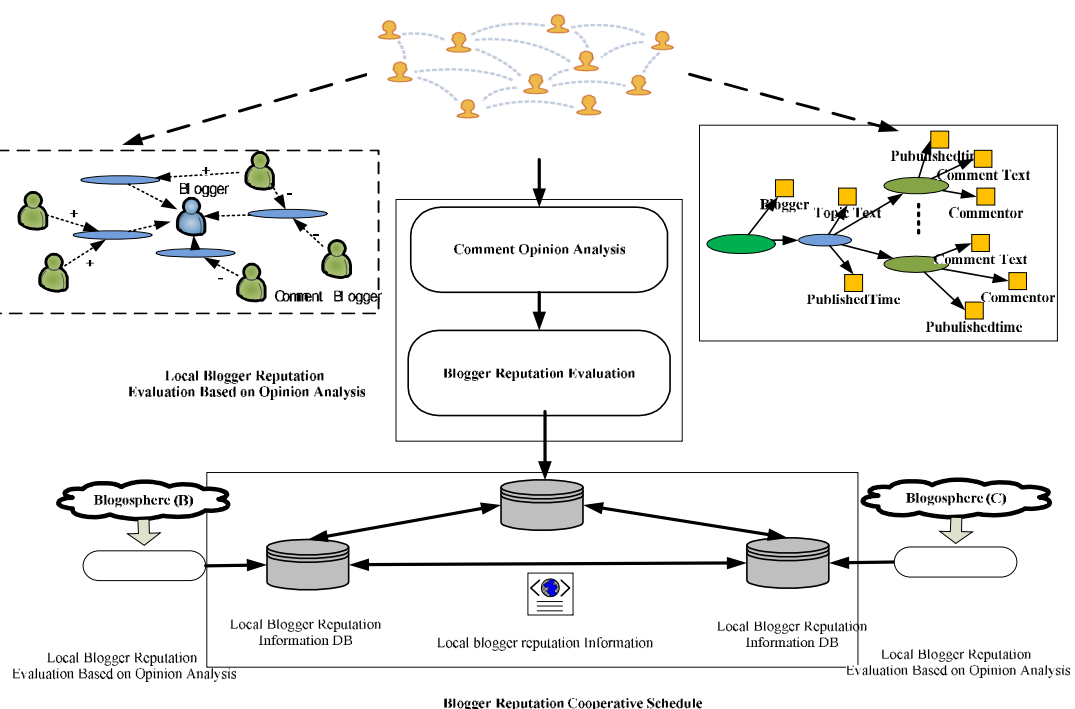


Figure 1: The general process of BREM.

4 Distributed blogger reputation evaluation based on opinion analysis

4.1 Blog knowledge representation

From the perspective of composition, blogosphere is made up of lots of blogs and the related page links [21]. Each blog includes a series of topics which are ordered by the published time. The author of a blog is named as “blogger” who owns the unique blog sphere. As shown in Figure2 (A) and Figure2 (B), BREM extracts some blog information (blogger, topic title, topic text, published time, comment text and the reviewers) and represents them as the format of RDF [22, 23]. In Figure2 (C) and Figure2 (D), for improving the performance of the blog comment opinion analysis, some typical blog comments are abstracted as the opinion cases.

With the excellent knowledge representation ability of RDF, The opinion case is described as the following three-triples:

OpinionCase =

<Case Subject, Case Predicate, CaseObject> (1)

Here, *Subject* represents the case resources which are uniquely identified by a Uniform Resource Identifier (URI). *Object* denotes the specific literals. Predicate is the binary relation between *Subject* and *Object*. Seven kinds of predicate attributes are defined as *Semantic Opinion*, *Positive Distribution Threshold*, *Negative Distribution Threshold*, *Positive Character Frequency*, *Negative Character Frequency*, *Positive Characters Corpus* and *Negative Characters Corpus*.

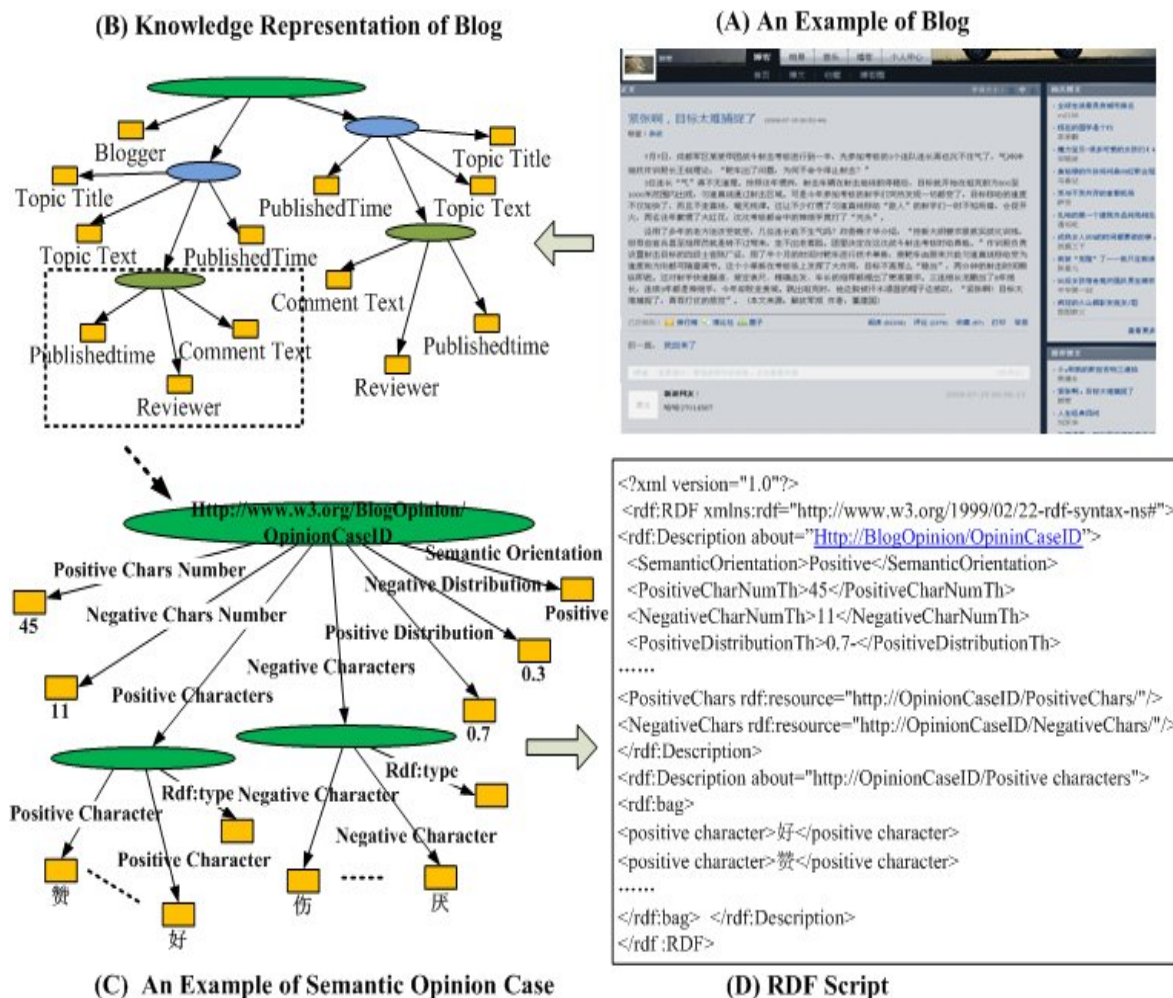


Figure 2: Blog Knowledge Representation Based on RDF.

4.2 Blog comment opinion analysis

When we read English text, people could identify the specific word by the blank character. However, in Chinese text, there is no any label between any two words. It greatly increases the difficulty of Chinese text mining. The traditional Chinese text opinion analysis methods usually split the words by some Chinese dictionaries firstly^[24]. While, due to being limited by the Chinese segmentation technology, the precision of opinion analysis could not meet the actual application requirement.

Raymond W.M. Yuen and Terence Y.W. Chan^[25] (2004) presented a general strategy for inferring SO for Chinese words from their association with some strongly-polarized morphemes. The experimental results proved that using polarized morphemes is more effective than strongly-polarized words. Based on this scenario, BREM improves the calculation model^[15] (Liu-Wei Ku and Yu-Ting Liang 2006) and evaluates the text opinion by analyzing the semantic orientation of Chinese characters. In the Blogosphere, users could publish or comment the topics freely. Some comments maybe consist of hundreds of words. Nevertheless, some ones only have dozens of words. For fitting in with the open environment of blogosphere, BREM adopts the different opinion calculation methods for the long text and short text, respectively.

In table 2, T is a paragraph of comment text, C_i represents the i -th character of T , N_{count} is the amount of words of T . fp_{ci} and fn_{ci} stands for the occurring frequency of C_i in positive and negative corpus. S_{ci} denotes the opinion degree of C_i . $OpDensity(S_{ci})$ is the distribution density of positive characters. $\|S_{ci}\|$ is the amount of positive characters in T . m and n denote the total number of unique characters in positive and negative words corpus. $Th_{LongText}$ is the boundary threshold of long text and short text.

Table 2: Opinion Analysis Algorithm for Blog Comment.

Input:	Comment Text T , C_i , N_{count} , $Th_{LongText}$, m , n
Output:	Semantic Orientation of $T \rightarrow S(T)$
Step 1:	//Initialize Inputs $S(T)=0$; Calculate the SO of each character //transverse all the characters of T For each character C_i
Step 2:	// where P_{ci} and N_{ci} denote the weights of ci as positive and negative characters.

$$P_{ci} = \frac{fp_{ci} / \sum_{j=1}^n fp_{cj}}{fp_{ci} / \sum_{j=1}^n fp_{cj} + fn_{ci} / \sum_{j=1}^m fn_{cj}} \quad (2)$$

$$N_{ci} = \frac{fn_{ci} / \sum_{j=1}^m fn_{cj}}{fp_{ci} / \sum_{j=1}^n fp_{cj} + fn_{ci} / \sum_{j=1}^m fn_{cj}} \quad (3)$$

//SO of character

$$S_{ci} = P_{ci} - N_{ci} \quad (4)$$

if ($\|S_{ci}\| \leq Th_{NeutralChar}$)

$$S_{ci} = 0 \quad (5)$$

Step 3: //Evaluate the Comment Opinion of T

//judge the length of T

if ($N_{count} \leq Th_{LongText}$)

// T is short text.

$$\text{then; } S(T) = \sum_{i=1}^{N_{count}} S_{ci} \quad (6)$$

// T is long text.

else

$S(T) =$

$$\sum_{j=1}^{\|S_c^+\|} S_g^+ \times OpDensity(S_c^+) - \sum_{j=1}^{\|S_c^-\|} S_g^- \times OpDensity(S_c^-) \quad (7)$$

Step 4: Return $S(T)$;

In Step2, BREM traverses all the characters of target text T and calculates the SO value of each one. Considering the quantitative difference of positive and negative words corpora, BREM normalizes the occurring frequencies of C_i and evaluates them respectively. In formula 4, through comparing P_{ci} (the character occurring frequency in positive words corpus) and N_{ci} (the character occurring frequency in negative words corpus), the semantic orientation of C_i is determined. If the certain character appears more times in positive words, then it is a positive value; and vice versa. To shorten the calculation error, in formula 5, BREM sets a threshold for the neutral sentiment character in advanced and returns to zero the absolute value less than $Th_{NeutralChar}$.

In Step3, BREM adopts two calculation methods to solve the different length of blog comments respectively. If the length of T is less than the threshold $Th_{LongText}$, the opinion of target text is determined by the SO sum total of all the characters. Otherwise, the length of T is greater than the threshold. We traverse the opinion cases and reuse the historical evaluation result. If there is not any case matching with the target text, as shown in formula 7, the SO of T is evaluated by comprehensively considering the mutual influence of the semantic orientation of characters and the opinion distribution density. In formula 8, through clustering the subjective characters of T , BREM analyzes the ratio of the sum of cluster radiuses to the whole amount of characters and calculates the opinion distribution density of subjective characters.

$$OpDensity(S_c^+) = \frac{\sum_{i=1}^k R_{Cluster} [Position(S_c^+)]}{N_{count} / 2} \quad (8)$$

Where, $Position(S_c^+)$ represents the position of S_c^+ in T . k denotes the amount of clusters. Some examples are given in Table 3 and 4.

Table 3: Short Comment Opinion Analysis Examples.

Short Text 53 words	Short Text 31 words
只有我们自己发扬助人为乐的精神，与人为善，我们才能得到别人的帮助和尊敬，才能在互动的真诚中感到真正的快乐。 (We will not get help and respect from the others until we are willing to help and be kind to anyone else. Then, the real happy will come in good faith.)	我被骗了！我被误了！他们是罪魁祸首，我开始喊了出来，我要宣泄。 (I was cheated! I was missed! They are the culprit, and I want to cry loudly and abreact.)
Score: +19.89 Classification: Positive	Score: -2.22 Classification: Negative

Table 4: Long Comment Opinion Analysis Examples.

Long Text 201 words	Long Text 145 words
多少年来，人们一直把教师比作红烛，赞颂她默默发光、无怨无悔奉献的精神，教师也一直以蜡烛精神来鞭策自己。“教师”虽说只是一个职业的称谓，而在现实中，教师的职业行为似乎成了他们生活的全部。他们以牺牲自我来换得学生茁壮成长，他们情系学生，倾情讲坛。讲台催人老，粉笔染白头，但教师们却无怨无悔，矢志不改，耕耘不辍。这种甘为人梯、	商家打着学雷锋的旗号，却利用雷锋的名字和图像乃至其精神进行炒作，甚至恶搞，行为极不严肃，这是对雷锋精神的漠视、曲解和颠覆，是对雷锋精神和形象的一种玷污，也是对时代进步的莫大讽刺。对于这种行为应该严肃对待，坚决抵制，否则我们辛苦树立的英雄形象将会被毁灭，真正的雷锋精神也将消失在我们的手

无私奉献精神就如同中。

红烛一般，燃烧自己来照亮别人，用自己的付出换来一批批学生的成长。

(For so many years, people always use the burning candle to analogy the devotion spirit of teachers. Teachers also use this spirit to encourage themselves. Although “teacher” is a kind of job, it becomes the only part of their lives)

Sum(S+) : 56.56
Sum(S-) : -18.52
OpDensity(S+):0.65
OpDensity(S-):0.31
Score: +30.86
Classification: Positive

(Some ones use the name the pictures and the spirit of “LeiFeng” to obtain the business interests. That disregards, distorts and subverts the spirit of “Lei Feng”, and satirizes the progress of our times sharply. We should resist these behaviours seriously. Otherwise, the spirit of “LeiFeng” would disappear for ever.)

Sum(S+) : 37.15
Sum(S-) : -35.50
OpDensity(S+):0.54
OpDensity(S-):0.70
Score: -4.72
Classification: Negative

4.3 Blogger reputation evaluation

Given a blogosphere $CBlogosphere$, A is any blogger of $CBlogosphere$. In the blogosphere, each blogger could publish the topics in the personal space or comment some ones of other blogs. In formula 9, $Reputation(A, t)$ and $Reputation(A, t+1)$ represent the reputation of A at t and $t+1$ respectively. $\Delta Reputation(A, t, t+1)$ is the increment reputation of A within t to $t+1$.

$$Reputation(A, t+1) =$$

$$f(Reputation(A, t), \Delta Reputation(A, t, t+1)) \quad (9)$$

Formula 9 is further expanded. As shown in formula 8, through tracking the supportive ratio of blog topics, the reputation of blogger is evaluated dynamically.

$$Reputation(A, t) =$$

$$\sum_{i=1}^{\|A_{Topic}\|} \left(\frac{\|Comments_i^+\| - \|Comments_i^-\|}{\|Comments_i\|} \right) \bullet \|View_i\| \quad (10)$$

Where, $\|A_{Topic}\|$ denotes the blog topics of A , $\|View_i\|$ and $\|Comments_i\|$ represent the reviewers and comments of the i -th topic. $\|Comments_i^+\|$ and $\|Comments_i^-\|$ are the number of positive and negative comments. The more the positive comments are, the more reliable the blogger is, and the reputation is higher.

On the contrary, with the increment of negative comments, the reputation of blogger is declined.

In formula 11, $P(A, \Delta t)$ is the increment of reputation between t and $t+1$. BREM analyzes the reputation fluctuation of A in deeply and projects the influence into the range of 0 to 1 by the exponent function.

$$\Delta Reputation(A, t, t+1) = e^{-|P(A, \Delta t)|} = e^{-\left| \frac{Reputation(A, t+1) - Reputation(A, t)}{Reputation(A, t)} \right|} \quad (11)$$

Through monitoring the number variation of positive comments, three cases should be discussed as follow:

- (1) With the increment of positive comments (namely $P(A, \Delta t) > 0$), the reliability of blogger is ascended and the reputation is increased.

$$Reputation(A, t+1) = Reputation(A, t) * [1 + Reputation(A, t, t+1)] \quad (12)$$

- (2) If the positive comments of two times are equal (namely $P(A, \Delta t) = 0$), the reputation of blogger keeps invariant.

$$Reputation(A, t+1) = Reputation(A, t) \quad (13)$$

- (3) With the reduction of positive comments (namely $P(A, \Delta t) < 0$), the reputation of blogger is decreased.

$$Reputation(A, t+1) = Reputation(A, t) * [1 - Reputation(A, t, t+1)] \quad (14)$$

4.4 Blogger reputation information cooperative scheduling

To balance the different reputation of the same blogger in multi-blogsphere, BREM further cooperatively schedules the local blogger reputation information and improves the blogger reputation evaluation ability in the global virtual social community.

Given any blogger α . $DomainA$ and $DomainB$ represent two blogsphere. $BRDB_{DomainA}$ and $BRDB_{DomainB}$ denote the local blogger reputation information database of $DomainA$ and $DomainB$, respectively. Δt is the time interval of cooperative scheduling. β is any blogger of $BRDB_{DomainB}$, and Th is the threshold of local reputation variation. The cooperative schedule algorithm of local blogger reputation is as follow:

Table 5: Blogger Reputation Information Cooperative Scheduling Algorithm.

Input:	Blogger α and β , $BRDB_{DomainA}$, $BRDB_{DomainB}$, Th , Δt
Output:	Target reputation information database $BRDB_{DomainB}$
Step 1:	//Initialize Inputs $SendListDomainA, ReceivedListDomainB \leftarrow \phi$
Step 2:	//Local Blogger Reputation Distribution //Traverse all bloggers of $BRDB_{DomainA}$ For any blogger α of $BRDB_{DomainA}$ //Local Blogger Reputation Evaluation. if ($Reputation(\alpha, \Delta t) > Th$) //Prepare to be scheduled by other network domains then $\alpha \rightarrow SendListDomainA$; //Blogger Reputation Scheduling; //Retrieving α from Domain A $\alpha \rightarrow ReceivedListDomainB$; //Traverse all blogger reputation information of $BRDB_{DomainB}$ For blogger β of $BRDB_{DomainB}$ //Analysing whether two bloggers are same or not by comparing with the <i>Email Address</i> if ($\alpha.email = \beta.email$) then // If they are same one, update the reputation and take the bigger one. If ($\beta.reputation < \alpha.reputation$) $\beta.reputation = \alpha.reputation$ Elseif ($\beta.reputation > \alpha.reputation$) //prepare to send β to $DomainA$ and modify the reputation of α $\beta \rightarrow SendListDomainB$; // if they are not the same one, insert new blogger reputation into $BRDB_{DomainB}$ else, Insert α into $BRDB_{DomainB}$;
Step 4:	//Output Return $BRDB_{DomainB}$;

In step2, BREM analyses the blogger reputation fluctuation in single blogsphere, and distributes the ones which have the higher number variation to the other network domain. In step3, the model retrieves the local blogger reputation information and queries whether there is the same one by comparing with the email address which is used as the unique identity of blogger. We do not consider whether two or more email addresses belong to the same blogger. If there exists the same blogger in target network domain, BREM updates the reputation

and takes the bigger one. Otherwise, BREM inserts the new blogger reputation information into target database.

5 Experiment

5.1 Experiment corpus

To validate the performance of BREM, we download over 70,000 blogs (time span from February 4 to May 30, 2009) from the Sina (<http://blog.sina.com>) and Renren (<http://www.renren.com>), construct an experimental corpus about “Unhealthy Campus Culture” (named UCC) and test the validity of the comment opinion analysis algorithm and the blogger reputation evaluation. Table 6 presents the information of UCC (Average increment of topics, reviews, comments, long comments and short comments) at four testing time.

Table 6: Information on UCC corpus.

	Feb.	Mar.	April	May
□Topics	22.2	24.1	18.3	27.4
□Reviews	164.11	126.65	185.5	144.6
□Comments	74.30	74.88	90.4	85.4
□ Comments(L)	30.6	31.01	40.4	35.2
□ Comments (S)	44.2	26.6	40.5	53.3

As the basis of opinion analysis, we collected and revised two sets of opinion words as the testing corpus, including General Positive-Negative Dictionary (abbreviated as GPND) and Chinese Network Sentiment Dictionary (abbreviated as CNSD). Table 7 shows the statistics of the revised testing corpus.

Table 7: Testing Corpus of Opinion Words.

Dictionary	Positive Corpus	Negative Corpus	Total
GPND	5,421	3,514	8,935
CNSD	1,431	1,948	3,379
Total	6,852	5,462	12,314

5.2 Experimental results

Testing 1: The Comparison Testing of Comment Opinion Analysis

To compare the validity of opinion analysis, we took a comparison testing among OSNB^[15], Morpheme^[25] and BREM. We selected 40,000 blogs from BREM as the testing set and divided them into long text corpus and short text corpus, respectively. Through calculating the Precision (P), Recall(R), F-measure (F) and Average

Execution Time (T), the performance of three methods was compared.

From the results of the comparison testing, we noticed that, BREM could adapt the different features of long text and short text, and improve the validity and practicability of opinion analysis.

Table 8: Opinion Analysis Comparison Testing for Long Comments.

	Long Comments Corpus		
	OSNB	Morpheme	BREM
P	59.87%	73.85%	79.49%
R	78.48%	74.11%	82.24%
F	67.92%	73.98%	80.84%
T	1.5s	1.1s	0.5s

For the long comments corpus, BREM reuses the evaluation results of historical case and comprehensively considers the mutual influence (the semantic orientation of Chinese characters, the distribution density of positive and negative characters). The precision of BREM (P 79.49%, R 82.24%, 0.5 second) is much than OSNB (P 59.87%, R 78.48%, 1.5 second) and Morpheme (P 73.85%, R 74.11%, 1.1second).

Table 9: Opinion Analysis Comparison Testing for Short Comments.

	Short Comments Corpus		
	OSNB	Morpheme	BREM
P	54.22%	70.23%	72.55%
R	65.79%	75.03%	74.28%
F	54.55%	72.55%	73.40%
T	1.1s	0.22s	0.15s

For the short comments corpus, BREM adopts the similar method with Morpheme, avoids the limit of Chinese segmentation technology and had better performance than OSNB (P 54.22%, R 65.79%).

Testing 2: The Validity Testing of Blogger Reputation Evaluation

Six blogs were constructed to evaluate the validity of blogger reputation evaluation ability of BREM. As shown in Table 10, six kinds of topics about “Unhealthy Campus Culture” are selected from UCC: Unhealthy Psychology (UP, 311, topics), Bad Habits (BH, 165 topics), Warning Speeches (WS, 264 topics), Corruptible Learning (CL, 242 topics), Campus Violence (CV, 202 topics) and Campus Eroticism (CE, 153 topics).

We further input different kinds of alert topics into three blogs of the two blogosphere respectively (Blogosphere A-> Unhealthy Psychology, Bad Habits and Warning Speeches; Blogosphere B->Corruptible Learning, Campus Violence and Campus Eroticism) at three time spans (t1, t2 and t3). Through comparing the number variation of (including the whole amount of comments, comments (+), comments (-) and the reviews) and the trend of blogger reputation, the validity of blogger reputation evaluation will be validated.

Table 10: the Input Data Statistics for Blog Reputation Validity Testing.

	Av Comments(+)/Page, Av Comments(-)/Page, Av Comments(all)/Page		
	t1	t2	t3
UP	15(+), 8(-),26	35(+),12(-),57	64(+), 21(-),96
BH	14(+), 10(-),33	31(+),16(-),72	43(+), 22(-),85
WS	15(+), 12(-),31	24(+),19(-),46	36(+), 28(-),73
CL	8(+), 5(-),33	19(+),14(-),54	22(+), 16(-),68
CV	12(+), 11(-),33	31(+),17(-),58	35(+), 37(-),79
CE	6(+), 8(-),21	14(+),11(-),32	23(+), 26(-),57

Type	Av Reviews/Page		
	T1	T2	T3
UP	53	84	119
BH	42	93	133
WS	56	121	88
CL	72	163	135
CV	89	86	94
CE	74	78	82

The testing results show that, BREM has the good blogger reputation evaluation ability and practicality.

As shown in figure 3, with the amount fluctuation of topic comments and reviews, BREM analyzes the supportive ratio of the other blogger to the topics in deeply and tracks the variation trend within the time span t1 to t3.

Take “Blog A-> Unhealthy Psychology (UP)” and “Blog E->Campus Violence (CV)” for example. The positive comments of blog A increase in the whole time, so the reputation of blogger A ascends. On the contrary, at t2 the supportive ratio of Blog E begins to reduce, BREM captures this trend and lowers the reputation level.

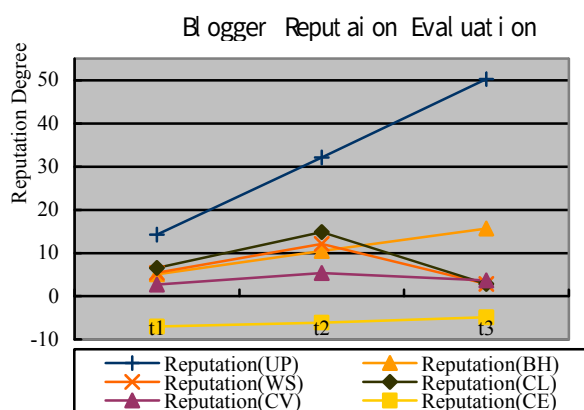


Figure 3. Blogger Reputation Evaluation Validity Testing

6 Conclusion & future work

In this paper, a distributed blogger reputation evaluation model based on opinion analysis (named

BREM) is proposed. Different with traditional reputation computing methods based on the page links, BREM analyzes the SO of each blog comment, tracks the semantic opinion attitudes of the bloggers and evaluates the blogger reputation level dynamically. Oriented to the length of blog comment, BREM designs two kinds of semantic orientation identification methods by calculating the mutual impacts of opinion weight of Chinese characters and the distribution density of opinion characters comprehensively. To balance the different reputation of the same blogger in the different network domains, BREM cooperatively schedules the local blogger reputation information among the multi-blogsphere and strengthens the management and analysis ability of blogosphere effectively.

In the experiment, we constructed a corpus about “Unhealthy Campus Culture” to validate the comment opinion analysis and the blogger reputation evaluation. The statistics results showed that, with increment of testing corpus, the model had higher opinion analysis ability (Long Comment: Precision 79.49%, Recall 82.24%, Average Executive Time 0.5 second; Short Comment: Precision 72.55%, Recall 74.28%, Average Executive Time 0.15 second) and the validity of blogger reputation evaluation. The statistics results of corresponding compared experiments are showed in table 8 and table 9 which also illustrate the advantage of our method.

In the future work, for improving the calculation scalability of BREM, we will transplant and deploy the original system into the distributed environment or cloud computing platform. With the help of the Map/Reduce technology^[26], a blogger reputation evaluation service will be built to strengthen the social status analysis ability of the virtual community - blogosphere

Acknowledgement

We would like to thank Dr. Liyong Zhao for discussing some issues about this paper. The work reported in this paper was supported by the Key Science-Technology Plan of the National ‘Eleventh Five-Year-Plan’ of China under Grant No. 2006BAK11B03 and No. 2008AA01Z109, Natural Science Foundation of China under Grant No. 60373008.

References

- [1] T. Bray. Measuring the web. In Proceedings of the 5th International World Wide Web Conference on Computer Networks and ISDN Systems, Elsevier, pp. 993–1005, Amsterdam, Netherlands, 1996.
- [2] J. M. Kleinberg. Authoritative sources in a hyperlinked environment. In Proceedings of the 9th Annual ACM-SIAM Symposium on Discrete Algorithms, pp.668–677, San Francisco, California, USA, January 1998.
- [3] S. Brin and L. Page. The anatomy of a large-scale hypertextual web search engine. In Proceedings of

- the 7th International World Wide Web Conference, pp. 107–117, April 1998.
- [4] Klessius Berlt, Edleno Silva de Moura, Andr e Carvalho and etc. A Hypergraph Model for Computing Page Reputation on Web Collections, SBBD 2007.
 - [5] Fusheng Jin, Zhendong Niu, Quanxin Zhang, and etc. A User Reputation Model for DLDE Learning 2.0 Community, ICADL 2008, LNCS 5362, pp. 61–70, 2008.
 - [6] Jennifer Golbeck, James Hendler. Inferring Reputation on the Semantic Web, WWW 2004, May 17–22, 2004, New York, NY USA.
 - [7] A Resnick, P, Kuwabara, K, A Zeckhauser, R and etc. Reputation systems, 2000, ACM New York, NY, USA.
 - [8] Yang, M, Feng, Q, Dai, Y and Zhang, Z. A multi-dimensional reputation system combined with trust and incentive mechanisms in P2P file sharing systems. 27th International Conference on Distributed Computing Systems Workshops, 2007. pp. 29–29
 - [9] HATZIVASSILOGLOU, V., AND MCKEOWN. Predicting the Semantic Orientation of Adjectives. Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and the 8th Conference of the European Chapter of the ACL, Madrid, Spain, 1997, pp.174–181.
 - [10] Pang, B., Lee, L., and Vaithyanathan, Sentiment classification using machine learning techniques. Proceedings of the 2002 Conference on EMNLP, 2002, pp. 79–86.
 - [11] Liu B., Hu M. and Cheng, J. Opinion Observer: Analyzing and Comparing Opinions on the Web. the 14th International World Wide Web Conference, 2005, pp.342–351.
 - [12] Kim Soo-Min and Hovy Eduard. Extracting Opinions, Opinion Holders, and Topics Expressed in Online News Media Text. In Proceedings of the ACL Workshop on Sentiment and Subjectivity in Text, 2006, pp. 1–8.
 - [13] Wiebe, J., T. Wilson and C. Cardie. Annotating expressions of opinions and emotions in language. Language Resources and Evaluation, 2005.
 - [14] Soboroff, I. and Harman, D. Overview of the TREC 2003 novelty track. The Twelfth Text REtrieval Conference, National Institute of Standards and Technology, 2003, pp. 38–53.
 - [15] NTCIR Project, <http://research.nii.ac.jp/ntcir/index-en.html>.
 - [16] Lun-Wei Ku, Yu-Ting Liang and Hsin-Hsi Chen. Opinion extraction, summarization and tracking in news and blog Corpora. Proceedings of AAAI-2006 Spring Symposium on Computational Approaches to Analyzing Weblogs, 2006, pp. 100–107.
 - [17] Ruifeng Xu, Kam-Fai Wong, et al. Learning Knowledge from Relevant Webpage for Opinion Analysis. Web Intelligence and Intelligent Agent Technology, 2008, pp. 307–313.
 - [18] Veselin Stoyanov and Claire Cardie. Topic Identification for Fine-Grained Opinion Analysis. Proceedings of the 22nd International Conference on Computational Linguistics, 2008, pp. 817–824.
 - [19] TURNEY, P.D. Thumbs up or Thumbs down? Semantic Orientation Applied to Unsupervised Classification of Reviews. Proceedings of the Association for Computational Linguistics 40th Anniversary.
 - [20] RDF Primer, <http://www.w3.org/TR/rdf-primer>.
 - [21] Blog, <http://en.wikipedia.org/wiki/Blog>
 - [22] Thomas (2008), Recent Developments in the Evaluation of Information Retrieval Systems: Moving Towards Diversity and Practical Relevance. In :Informatica 32, pp.27–38.
 - [23] S Mu oz, J P rez, C Gutierrez. Web Semantics: Science, Services and Agents on the world wide web, 2009, Elsevier.
 - [24] Yohei Seki, David Kirk Evans, Lun-Wei Ku and etc. Overview of opinion analysis pilot task at NTCIR-6, Proceedings of the Workshop Meeting of the National Institute of Informatics (NII) Test Collection for Information Retrieval Systems (NTCIR), pp. 265–278, 2007
 - [25] R W M Yuen, T Y W Chan et al. Morpheme-based Derivation of Bipolar Semantic Orientation of Chinese Words[A]. In: Proceedings of the 20th International Conference on Computational Linguistics, 2004, pp.1008–1014.
 - [26] Hung-chih Yang, Ali Dasdan, Ruey-Lung Hsiao and D. Stott Parker. Map-reduce-merge: simplified relational data processing on large clusters. Proceedings of the 2007 ACM SIGMOD international conference on Management of data, Beijing, pp.1029–1040, 2007.