A Social Stability Analysis System Based on Web Sensitive Information Mining

Wei Wang(✉)

Department of Electronic Technology,
Engineering University of CAPF, Xi'an, China
wjwangwei@pku.edu.cn

Abstract. Researches on domestic social stability analysis mainly focus on construction of social stability theory, architecture and index, while few pay attention on quantitative analysis. In this paper, a social stability supervising framework is proposed based on sensitive Web information mining, semantic pattern matching and quantitative calculating. A sensitive information knowledge base is constructed by analyzing sensitive information about social environment, national harmonious and happy index of people's live in natural language online news texts from Internet, and recognizing hot keywords as well as the event trends led by the keywords. A social stability index theoretic model and a quantitative calculating model are proposed to evaluate social stability quantitatively. Parameters of the calculating model are determined by employing social investigations and an iterative feedback learning method. A prototype system is built on proposed framework and experiments are conducted on six frontier provinces, e.g., Xinjiang and Tibet. The result of an average accurate of 73.29 % shows the effectiveness of the proposed model.

Keywords: Sensitive information · Social stability index · Web text mining

1 Introduction

There are many kinds of information on the Web, e.g., information about gaps between the rich and the poor, bad social security and unemployment which related to the social environment; information about religious convictions, different lifestyles and penetrations of foreign culture which involved in ethnic harmony; as well as information about living environment, social insurance and disposable incomes which related to people's livelihood. With the continuous improving of the popularity of the Internet, the virtual online space has a growing influence on the real world. Facts proved that some real affairs such as parades, meetings and associations in the real world came into being by discussion in online community at first. So, employing some information technologies to perform a comprehensive, accurate and timely supervision on the sensitive information on the Internet, and then to issue early warnings for fast responses in the real word, the social stability and unity, the vigorously development of the economy can be effectively maintained and protected.

At present, the supervision of network information is mainly done by public opinion monitoring systems. These systems are able to monitor web information, trace

W. Chen et al. (Eds.): BDTA 2015, CCIS 590, pp. 46–58, 2016.
DOI: 10.1007/978-981-10-0457-5_6

hot events and carry out correlation analysis and trend analysis, but they cannot give apparent results on the social stability [1, 2]. The existing domestic work on social stability analysis mainly focus on the theory, system, index construction [3–11], few aims to achieve real-time assessment on the situation of social stability using Web information [12]. The deficiency of existing work lies in two aspects. On one hand, a large number of studies have been carried out only for qualitative analysis, but not for more meaningful quantitative results. On the other hand, some work is limited to a single factor, for example, research only focuses on happiness index without taking into account the combined effects of multiple factors.

In this paper, a social stability index theoretic model and a quantitative calculation model are proposed to evaluate social stability quantitatively. The theoretic model comprehensively considered three kinds of factors, the social environment factor, the national harmonious factors and the happiness index factors. Parameters of the calculating model are determined by employing social investigations and an iterative feedback learning on massive natural language text on the Web. Based on these models, a social stability supervising framework is proposed by sensitive Web information mining, semantic pattern matching and quantitative calculating. A prototype system is built on proposed framework. By analyzing sensitive information about social environment, national harmonious and happy index of people live in natural language news of six frontier provinces, e.g., Xinjiang and Tibet on the Internet, and recognizing hot keywords as well as the event trends led by the keywords, the system is able to monitor the social stability in time. The experiment result of an average accurate of 73.29 % shows the effectiveness of the system.

The rest of the paper is organized as follows. We first review related work in Sect. 2. The proposed technical framework is introduced in Sect. 3. Implementation of the prototype system and a case study are described in Sect. 4. In Sect. 5, we evaluate the system and demonstrate its feasibility and applicability. In Sect. 6, we conclude this paper.

2 Related Work

Some social science researchers in China have worked on the social stability situation analysis, index system construction and management system development. Li [2] gives an empirical analysis on the influence factors of social stability in the frontier minority areas from two aspects: the fact evaluation index and the stable confidence index on the basis of structural survey statistics. In documents [4, 6], the economic significance of happiness index, the construction of the index system, as well as the collection and the empirical analysis of happiness index were studied. The psychological factors of social group events were discussed in [7, 8]. The index system of harmonious development of economy society, which is composed of 38 important indicators, is constructed in [9]. From the perspective of economics, documents [10, 11] make a tentative analysis of the social and political stability of our country respectively. In [12], construction of a social stability early warning and management system is described, while the system can only use information input manually but not information grabbed automatically from the Web. So far, we can see that the research mainly focus on the theory, system and index, and Web information mining technology has not been used to analyze the social stability.

The public opinion monitoring applications are directly related to this paper. A public opinion monitoring system is able to discover and extract useful information from semi-structure or non-structure data in Web page content automatically, find and trace hot spots and focus events that newly happened and interested by people in the vast amounts of Web information, form a certain correlation analysis and trend analysis. There are some relatively good public opinion monitoring systems, for example, the Founder's Intellectual Thought public opinion early warning and assistant decision support system[1], the TRS Internet public opinion information monitoring system[2], the People's opinion[3], the Eagle micro blogging and emotion[4]. These systems are based on the information acquisition technology and employ information processing, content management, knowledge management and information classification technologies to achieve network public opinion monitoring, hot news tracking and supervision.

In this paper, we use text mining technology to realize the monitoring of social stability. Different from previous work, we pay more attention to quantitative social stability situation analysis. Namely, according to the proposed theoretical model and calculate model of social stability index, on the basis of acquisition, analysis and processing of social stability related network information, we perform quantitative calculation to get the situation of social stability.

3 Social Stability Analysis Technology Framework

This paper proposes a social stability situation analyzing framework based on sensitive information mining technology. In the framework, an exponential model of social stability is constructed, and an automatic quantitative social stability index calculation process is realized. The overall technical framework is shown in Fig. 1, it comprises three layers:

- ***The Web Mining Layer:*** This layer offers massive Web text mining services. By examining elements involved in the social stability model, this layer first crawls relevant Web pages, then uses TML (Text Mining Language) to extract keywords, understand semantic meaning, recognize associate relation and analyze sensitive information within page content. TML encapsulated complex web crawling and natural language processing technologies, it can easily maps the theoretical model and the extracting rules to the specific text mining process [13].
- ***The Knowledge Discovery Layer:*** This layer performs theoretical modeling, rule extraction and knowledge discovery. According to the social stability index theory model, it first recognizes the key words and the relationships of each factor to construct a sensitive information matching rule base. Then it uses an iterative feedback mechanism to determine the weight of each factor in the social stability index calculate model, to realize the quantitative calculation of social stability.

[1] http://www.founder.com/templates/T_Second/index.aspx.

[2] http://www.trs.com.cn/product/product-om.html.

[3] http://yuqing.people.com.cn.

[4] https://www.eagtek.com.

– ***The Data Presentation Layer:*** This layer supports data visualization and maneuverability. It takes the sensitive information extracted by the social stability index quantitative calculation model as input, represents them in forms of charts or other visualized methods to show the changes of social stability, and provides human-machine interface for further intelligent information analysis and decision making.

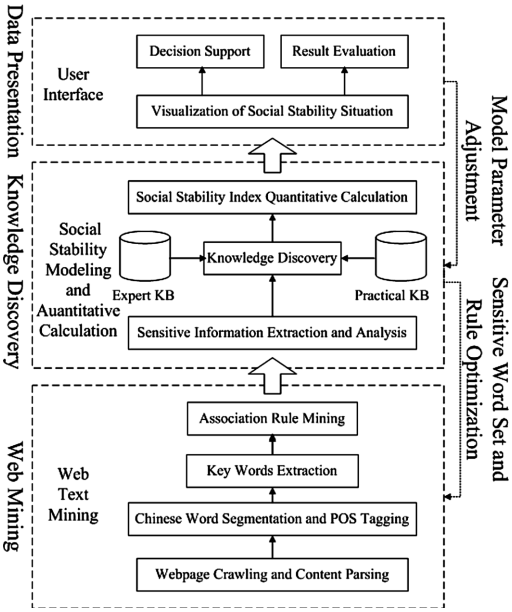

Fig. 1. A quantitative social stability analysis framework based on web sensitive information mining.

3.1 The Web Mining Layer

At the bottom of the framework, the text mining layer extracts sensitive words and matching rules from the massive network text under the guidance of the knowledge discovery layer. The extracted words and rules can be classified into three categories which have direct associations to the social stability, i.e., social environment, ethnic harmony and happiness index.

News is a kind of style that is widely used by newspapers, radio and other media to record facts, transfer information and reflect the times. The openness of the Internet enables the network news accounting the real society more directly and more quickly. So, many factors affecting the social stability situation can be found in online news. Two ways are employed to obtain sensitive information on the Web in this layer:

1. Extract sensitive information manually. This method is executed by reading news page artificially, selecting sensitive information as ‘seed’ according to some public

views on the current situation and the policy. It has a good effect in the initial state, but its efficiency is relatively low.
2. Using TML to capture online information automatically. Initially, some manually obtained sensitive words are feed to TML's web crawler as key words to carry out a directed crawl, and then the returned web pages are analyzed to obtain more sensitive words and all these words are put into a sensitive word set to support an iterative crawling process.

TML is a natural language processing platform, which contains a compiler, a virtual machine and an integrated development environment. Users can use the TML language to write text mining code, and then these codes are compiled into byte-codes to run on the virtual machine. TML implemented and encapsulated most commonly used text mining technologies to provide a simple way for complex text mining.

In the text mining layer, TML functions of web crawler, text extraction, Chinese word segmentation, part of speech tagging and named entity recognition, keyword extraction, concept extraction and relation extraction are used to implement the sensitive information mining; it's the basis for constructing the knowledge base.

According to the social stability theory model, the CONCEPT and PREDICATE directives of TML are used to define the sensitive word sets and rules, directive PAGES is used to determine the range of information acquisition, and the concept and relationship is extracted by directive SELECT. The TML codes are described in Table 1.

Table 1. TML Codes of Web Mining.

CONCEPT x;	/* Define sensitive word set X */
CONCEPT y;	/* Define sensitive word set Y */
PREDICATE x-y;	/* Define relations between sensitive word sets */
PAGES Sample Define website	/* Define the range of Web page for crawler */
SELECT x-y from Sample;	/* Extract the relations */
OUTPUT;	/* Output the results in XML */

For example, in analyzing the 'economic income' factor in 'social environment', we first manually recognize word set CONCEPT (income) as {"收入", "工资", "薪水", "生活费"}, after taking it as a seed to crawl the Web and expand it with synonyms, we get CONCEPT (income) = {"收入", "工资", "薪水", "生活费", "平均收入", "平均生活费", "经济", "物质", "生活必需品", "饮食质量", "伙食费", "平均工资", "平均薪水", "可支配收入", "可支配工资", "可支配薪水", "生活用品"}.

The semi-automatic learning process only completed the identification of sensitive words. In order to observe the changes of social stability, some verbs are needed to describe the trend of sensitive words. For example, in the 'social environment' factors, we need to analyze the changes related to the sensitive information 'economic income'. So the semi-automatic learning method is also used to construct a word set of verbs which can denote the status of 'economic income', namely CONCEPT (income-v) = {"低", "下降", "减少", "滑", "降", "回落", "低落", "低下", "没有", "不够", "拮据"}.

In order to realize the precise semantic matching, and avoid complex analysis of Chinese grammar, the 'co-occurrence' method is used to define the predicate modifier relations between the sensitive words and the trend verbs. Function *PREDICATE* SE_1*(income* n_1*, income-v* v_1*) {dist_15(*n_1*,* v_1*);}* means that at a distance of 15 words (the average length of a sentence), a word from the collection of *income* and a word from *income-v* formed a relation of the subject and predicate, which describes a kind of factor which will affect the social stability. This matching method based on the distance between two sets forms a mapping of $|income| \times |income\text{-}v|$, which will improve the rule's coverage, and will also improve the recall rate just like synonym expansion.

3.2 The Knowledge Discovery Layer

The function of knowledge discovery layer is to construct the theoretic and the quantitative calculation models of social stability. When the models are built, they can be used to guide the text mining layer to execute the rule extraction and knowledge discovery tasks.

Theoretic Model of Social Stability Index. Through empirical analysis, we find the diversified characteristics of factors which have effect on social instability. That is to say, a bunch of factors such as economy, employment, social security, price, interest, political, ethnic, cultural, religious, hostile forces penetration, emergencies, land acquisition and so on are found undermining the social stability.

In this paper, by thorough analysis, discussion and investigation, we believe that the social stability index (*SI*) is a linear combination of the social environment (*SE*) factors, the national harmony (*NH*) factors and the happiness index (*HI*) factors. The definition of *SI* is shown as Eq. (1).

$$SI = \alpha SE + \beta NH + \gamma HI \tag{1}$$

Where $SE = \alpha_1 RP + \beta_1 SS + \gamma_1 ES + \cdots$, it means the social environment is defined as a combination of a variety of factors related to social environment. Here, RP, SP, EQ… respectively represents the element of the rich and the poor, the social security, employment situation and other elements.

$NH = \alpha_2 RC + \beta_2 FP + \gamma_2 LS + \cdots$, it means national harmony is defined as a combination of a variety of factors related to national unity. Here, RC, FP, LS, … respectively represents the element of religion convictions, foreign penetration, lifestyle and so on.

$HI = \alpha_3 DI + \beta_3 SA + \gamma_3 EQ + \cdots$, the happiness index, is defined as a combination of a variety of factors related to happiness. Here, DI, SA, EQ, … respectively represents the element of disposable income, social assurance, environmental quality and other factors.

Factors Affecting Social Stability. As mentioned above, the social stability index is influenced by many factors. In order to determine the importance of each factor, we designed a questionnaire to investigate the factors' influences on social stability. In 2013 March and April, we randomly issued a total of 600 questionnaires in universities,

enterprises and streets to make a survey. The response rate was 91 % and 500 questionnaires were available, where 187 people were ethnic minority and 313 people were Han. The age distribution, occupation distribution and the education level distribution are found in Table 2.

Table 2. Statistics of questionnaire participants.

Age	Num.	%	Professional	Num.	%	Education	Num.	%
<20	119	23.8	Migrant worker	100	31.4	Middle school	54	10.8
20 ~ 30	178	35.6	Student	103	21.6	High school	89	17.8
30 ~ 40	103	20.6	Teacher	89	17.85	University	198	39.6
40 ~ 50	67	13.4	Doctor	39	10.2	Postgraduate	97	19.4
>50	33	6.6	Businessman	79	15.8	Ph.D	62	12.8
			Worker	90	24.2			

The statistics results of the survey showed the influence factors of social stability. We classified the detail factors into social environment, national harmony and happiness index. They are shown in Table 3.

Table 3. The factors that affect social stability.

Categories	Factors	Items
SE	Economic income	(1) per capita income (2) income growth and price growth ratio (3) is stable
	Employment status	(1) employment situation (2) attitude to current occupation (3) is stable
	Career promotion	(1) chance of promotion (2) self fulfillment
	Social status	(1) local or migrant (2) rural or urban (3) regional superiority
	Welfare support	(1) endowment insurance (2) medical insurance (3) city infrastructure (4) environment
	Family life	(1) housing and transportation (2) marriage (3) spouse (4) family relationship network
	Living condition	(1) pollution degree (2) city planning (3) public security level
	Group event	(1) social contradictions (2) illegal assembly activities (3) Riot (4) fury
NH	Economic development	(1) backward economy (2) the gap between rich and poor (3) price rise (4) unemployed persons
	Government duty	(1) unbalanced social development (2) social security (3) social injustice (4) increased crime rate
	Ethnic issues	(1) ethnic separatist activities (2) ethnic conflicts (3) religious issues

(*Continued*)

Table 3. (*Continued*)

Categories	Factors	Items
HI	Quality of life	(1) consumption level (2) environmental quality index (3) per capita disposable income of urban residents (4) Engel coefficient (5) per capita living space
	Social order	(1) incidence of mass incidents (2) duty crime rate of civil servants (3) incidence of major accidents (4) negative political rumors
	Social stability	(1) inflation rate (2) actual unemployment rate of urban (3) social security coverage (4) medical insurance coverage

3.3 Quantitative Calculation of Social Stability Index

In the proceed of questionnaire survey, the respondents were asked to sort the factors from big to small by the factors' influences on social stability according to their personal feelings. The same sorting was made on items in each kind of factors. For the sorting result of a certain class of factors, assuming the number in the first place is x_1, the number in the second place is x_2,…, the number in the m place is x_m, then according to the statistical results, the influence coefficient a_i of the factors on social stability was calculated by Eq. (2).

$$a_i = \frac{x_1 \times \theta_1 + x_2 \times \theta_2 + \cdots x_m \times \theta_m}{500 \times m} \tag{2}$$

Where $\theta_j = \frac{m-j+1}{m}, j = 1, 2, \ldots, m$.

According to Eq. (2), we get a subjective estimated parameters of the model. In order to refine these parameters, we need to select several websites from the frontier provinces to get actual experimental data. Six sites such as Xinjiang and Tibet are chosen as data sources to be crawled by comparing the capacity and the update frequency of the content on the websites, as shown in Table 4.

Table 4. Website list for data sampling and analyzing in the prototype system.

Province	Website URL
Tibet	http://www.chinatibetnews.com
Xinjiang	http://www.ts.cn
Guangxi	http://www.gxnews.com.cn
Inner Mongolia	http://www.nmg.xinhuanet.com
Jilin	http://www.jl.xinhuanet.com/
Yunnan	http://www.yn.xinhuanet.com

By analyzing empirical data, parameters of the model are verified and adjusted, and the calculation formula of the social stability index *SI* is eventually defined in Eq. (3):

$$SI = 0.45 \times SE + 0.35 \times NH + 0.2 \times HI \tag{3}$$

In the equation, the social environment *SE* is:

$$\begin{aligned} SE = {} & 0.25 \times (\text{Economic income}) + 0.09\times \\ & (\text{Employment status}) + 0.05\times \\ & (\text{Career promotion}) + 0.12 \times (\text{Social status}) + 0.13\times \\ & (\text{Welfare support}) + 0.15 \times (\text{Family life}) + 0.08\times \\ & (\text{Living environment}) + 0.13 \times (\text{Group event}) \end{aligned} \tag{4}$$

The national harmony factor *NH* is:

$$\begin{aligned} NH = {} & 0.5 \times (\text{Economic development}) + 0.3 \times (\text{Government duty}) \\ & + 0.2 \times (\text{Ethnic issues}) \end{aligned} \tag{5}$$

The happiness index factor HI is:

$$\begin{aligned} HI = {} & 0.4 \times (\text{Quality of life}) + 0.4 \times (\text{Social order}) \\ & + 0.2 \times (\text{Social stability}) \end{aligned} \tag{6}$$

3.4 The Data Presentation Layer

In the data presentation layer, graphs and tables are employed to display the social stability data. The graphic interface can provide a dynamic and visualized view for users to make better decisions. The optional data display modes include:

1. Line Chart: a social stability index linear chart is drawn according to the quantitative analysis result of stability index, and the line chart can intuitively shows the stability trend of the supervised provinces in a period of time.
2. Situation Map: a China Map is rendered everyday to dynamically monitor the stability index of different provinces in time. To display the stability status intuitively, the Map is colored into green, blue, yellow, orange and red according to the general international security level.

4 Prototype System Demonstration

4.1 System Construction

According to the proposed technical framework, a Browser/Server based frontier province social stability index analysis system is implemented in this paper. Apache, TML and JSP are used to develop the server side program; JavaSript and Ajax are employed to build friendly use interfaces in the client side. Figure 2 shows the architecture of the social stability index analysis system.

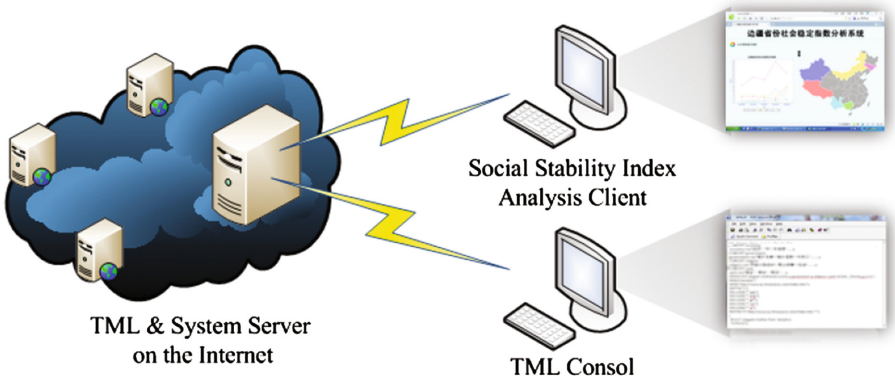

Fig. 2. Structure of the frontier province social stability analysis system.

4.2 Case Study

We ran the system on the Internet, and analyzed the social stabilities of six frontier provinces such as Tibet, Xinjiang and so on. The stability index line charts of these provinces from 2013/5/6/to 2013/9/6 are shown in Fig. 3.

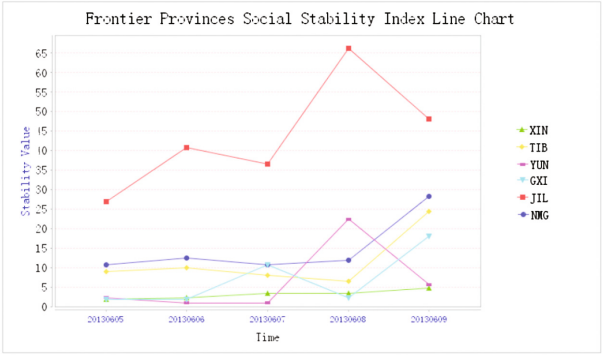

Fig. 3. A social stability line chart of 6 frontier provinces.

In the figure, the stability indexes of Jilin Province are high and the values change significantly. By manually analyzing the content in the grabbed web page, we found that a fire explosion accident had occurred in Jilin province on June 3rd, 2013. So, during those days, many reports appeared about the event. Some extracted sensitive information of the system are listed below:

(1) 3/6–5/6: the 6.3 serious fire explosion event happened, reports appeared to reveal the accident.
(2) 6/6–7/6: the death in the accident is announced one after another.

(3) 7/6–8/6: The explosion accident continuously fermented, it became a hot event and led to a strong reaction in society. News about the accountability and influence control were gathered and published. For example, “当地曾为出事工厂违规开路”, “政府道歉后还需追责”, “液氨高温后易造成流行病与疫病流行”.
(4) 8/6–9/6: Problems were solved, reports about the fire explosion gradually reduced. At the same time, news about the college entrance examination became the headlines and the stability index looked normal.

In addition to the reports about fire and explosion accident, a large number of other news which had impact on the social stability of Jilin in that period were also extracted. For example, “吉林长春市一地铁施工处发生施工事故”, “吉林一法院‘温馨提示’引发公众批评”, “吉林石化乙二醇出厂报价小幅上涨”, “吉林榆树高考乱象娱乐了谁”, “吉林男子行凶 见义勇为者身中多刀”, “韩企白菜价进口中国人参暴利 吉林千亿计划阻击”, and so on.

5 Evaluation and Analysis

To verify the feasibility of the system and the proposed social stability index model, the system results were compared with the fact we manually extracted from the actual news. The statistics results are shown in Fig. 4.

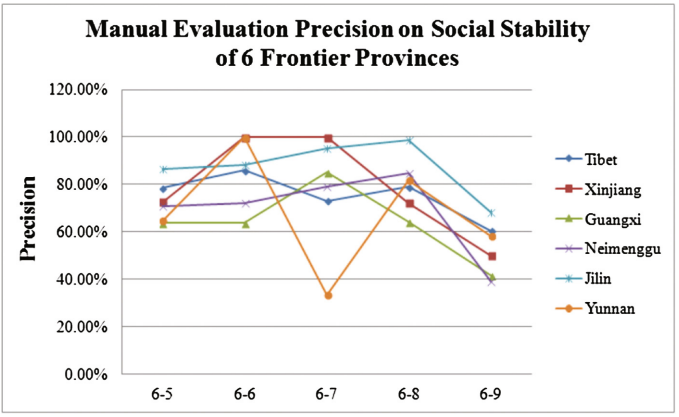

Fig. 4. The manual evaluation accuracy on social stability of 6 frontier provinces in 2013-6-5～2013-6-9.

The precision is defined as:

$$percision = \frac{|sensitive\ words\ identified\ manually|}{|sensitive\ words\ identified\ automatically|} \times 100\,\% \tag{7}$$

The figure shows the precisions of the sensitive information extraction from 6 representative websites in the frontier provinces. In June 7th and June 9th, the

precisions of Yunnan are low. By tracing the calculating procedure of social stability index, the problems in the expansion of sensitive words and the design of the words set structure were found. So we classified the sensitive word sets according to their semantics and optimized the cross correlation among the word sets. After correction and optimization, the average accuracy rate of the system raised to 73.72 %.

The experimental results show that the proposed model and technical framework can better monitor the social stability, and timely reflect the changing trend of social stability. To further improve the practicality of the system, more work should be done in two aspects: (1) When selecting sensitive information, refer to Baidu hot words list and other resources to enhance the authority of the constructed sensitive information knowledge base; (2) Employing text polarity analysis technology to grasp public opinion's trend in a finer granularity.

6 Conclusions and Future Work

Sensitive information extraction and social stability analysis technologies are studied in this paper. A social stability index theoretic model and a quantitative calculation model are proposed to evaluate social stability quantitatively. A B/S based prototype system is implemented based on TML's text mining and exact semantic matching technologies. Practical evaluation were conducted on six frontier provinces such as Xinjiang and Tibet, the results confirmed that the proposed model and the prototype system could better reflect the situation of social stability.

This work is able to provide useful information to the army, government and public security intelligence departments for making better decision, and eventually maintain the national stability and unity.

Acknowledgments. This work is supported by the Young Scientists Fund of the National Natural Science Foundation of China (Grant No. 61309022) and the Military Application Research Project of CAPF (Grant No. WXK2015-13).

References

1. Shou, L., Chen, G., Hu, T., Chen, K., Wang, Y.: A relevance mining method of Internet hot spot topic. Invention patent CN101158957 (2008)
2. Li, Y., Sun, L.: Hot-word detection for Internet public sentiment. J. Chin. Inf. Proc. **25**(1), 48–59 (2011)
3. Li, Y.: Analysis of social stability influence factors in frontier ethnic areas. Heilongjiang Nat. Periodicals **2010**(1), 36–43 (2010)
4. Tang, X., Yang, P.: On evaluation model for Chinese citizens happiness index. J. Anhui Sci. Technol. Univ. **26**(2), 61–65 (2012)
5. Kang, J.: The meaning and measurement of happiness. China Statistics **2006**(9), 18–19 (2006)
6. Gong, C.-Z.: How to build the index system of GNH. J. Eastern Liaoning Univ. (Soc. Sci.) **8**(6), 84–87 (2006)

7. Liao, H., Cao, H.: Social psychological mechanism produced by group events and its countermeasures. Innovation **2009**(1), 83–87 (2009)
8. Qiu, Z.: A social psychological foundation analysis on network public opinion in massive incidents. J. Gui Zhou Province Committee Party's School of C.P.C. 2011(3), 82–85 (2011)
9. Zhu, Q.: A comprehensive evaluation on index system of the harmonious development in economic society. Society of China Analysis and Forecast (2007)
10. Song, L., Appleton, S.: An empirical investigation into social discontent in urban China. China Econ. Q. **6**(4), 1339–1358 (2007)
11. Hu, L., Hu, A., Wang, L.: A empirical analysis on the changing situation in social unstable factors. Discovery **2007**(6), 105–114 (2007)
12. Yan, Y.: The measurement of the social stability and the construction of presentiment management system. Sociol. Stud. **2004**(3), 1–10 (2004)
13. Li, J., Li, X., Meng, T.: A universal and efficient language text mining. In: The 19th China Conference on Information Retrieval, vol. 7 (2013)