International Journal of Smart Home
Vol. 8, No. 6 (2014), pp. 169-178
http://dx.doi.org/10.14257/ijsh.2014.8.6.16

A Study on the Collection Site Profiling and Issue-detection Methodology for Analysis of Customer Feedback on Social Big Data

Eun-Jee Song[1] and Min-Shik Kang[2]

[1]*Department of Computer Science, Namseoul University, Cheonan 331-707, Korea*
[2]*Department of Industrial Management and Engineering, Namseoul University, Cheonan331-707, Korea*
sej@nsu.ac.kr, mskang@nsu.ac.kr

Abstract

As competition among the corporations in the service industry is growing fiercer, efficient management of customer feedback is necessary in order to grasp customer needs, which change day by day. Recently the corporations have been trying to obtain customer feedback using Big Data from social media, which contains the diverse voices of the customers. Therefore, the corporations focus their attention on how to analyze and utilize the Big Data, which is a key resource of the mobile smart revolution.

Firstly, this paper proposes a profiling method that can effectively analyze company reputation in the service industry. To that purpose, the proposed system extracts and lists a set of specialized target sites for each service.

This paper proposes a methodology which detects issues by analyzing diverse data patterns as a method for analyzing the Big Data of social media. The Issue-detected Methodology defines the independent variables as contents and writers which affect the spread of negative public opinions, and the dependent variables as average reaching time and speed of the issues. The influence of the negative public opinions is detected concerning issues based on the numbers of tweets and re-tweets. . The service providing corporations may prepare appropriate measures by the issue detection prior to the spread of the negative public opinions.

Keywords: *Service Industry, Social Big data, Customer Feedback, Collection site profiling, Pattern Analysis, Issue-detection*

1. Introduction

With ever-developing technology of networks, online users can express their opinions in a variety of spaces such as message boards in websites, blogs, cafes and social service networks (SNS). In blogs and SNS there are many consumers to whom the corporations try to sell their products or services, so it is very efficient to use Big Data on social media as a method to understand customers' needs in real time [1]. While the internet has been the key resource during the IT(Information Technology) period, Big Data plays the role of key resource during the recent mobile smart period. In particular, Big Data is the necessary resource in order to get the customer feedback information for efficient management of corporations [2].

The existing techniques for data analysis are not sufficient to analyze Big Data because Big Data has characteristics such as the enormous volume of information, the high speed of generation and circulation of real time updated information, and the fusion of untypical and unstructured data [3].

ISSN: 1975-4094 IJSH

The majority of Big Data are composed of untypical data such as texts and images. So for the approach to an analysis of untypical Big Data, different information retrieval techniques from the existing analysis techniques, and untypical data analysis techniques such as Text Mining, are required. There is Buzz Monitoring as a method for analyzing and investigating the words of customers online. This is a system which analyzes and collects diverse information automatically on the web.

For instance, the results of buzz monitoring of customer reputation were reported in an analysis of clients' responses to the services of some major large hospitals and hotels in Korea [4, 5].

The current paper proposes a profiling method that can effectively analyze customer reputation in the service industry. To that purpose, the proposed system extracts and lists a set of specialized target sites for each service. Then, it extracts information/opinion-sharing message boards and knowledge-sharing sites containing questions and experts' recommendations. A set of major websites for each service industry are selected and listed to collect valid information

Also, this study proposes a method to detect and predict recent issues by analyzing data patterns based on Buzz Monitoring system, which analyzes the social big data. It is possible to detect negative issues through various data pattern analysis methods such as the analysis of the social data themselves related to the target institutions or corporations, and correlation analysis with the related data.

2. Related Works

2.1. Evaluation of Service Quality Elements

Parasuraman (1988) has continuously contributed to understanding the concept of service quality and developed SERVQUAL as a general instrument for evaluating the service quality [6].

Although there are diverse opinions, the analysis of the relation between service quality and satisfaction level using the mentioned measurement indexes resulted in that there is a correlation among all the indexes .Those diverse measurement indexes have correlation with the structural elements of SERVQUAL which is a general tool for measuring satisfaction level. As Figure 1 we select 5 standard Quality Factors (Responsiveness, Tangibles, Assurance, Empathy, and Reliability) and 10 Measurement Factors and determine questionnaire items which are measurement items for each factor.

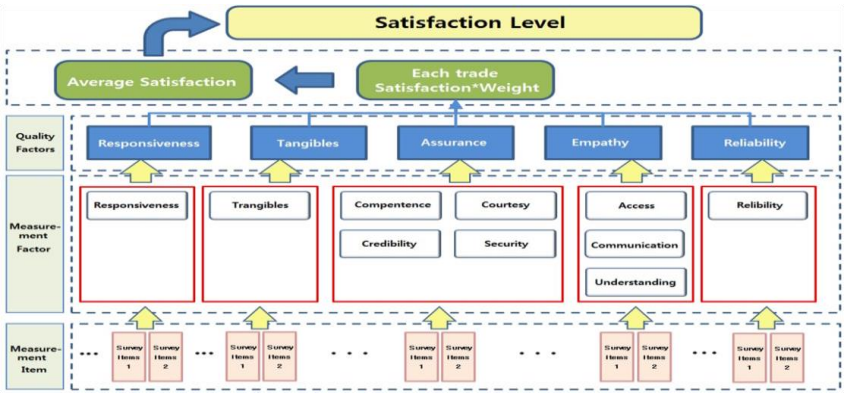

Figure 1. The Satisfaction Analysis Model using SERVQUAL

At present, many buzz monitoring systems of big data from social media sites and other web sites use either positive or negative keywords in analyzing and evaluating customers' responses. However, in order for a more accurate and effective analysis, it is necessary that not only the dichotomy of positive and negative attitudes but also the amount of recommendations and exposure of service quality elements should be analyzed and compared in determining the ranking for each category [4,7].

The four factors below are proposed in analyzing exposure of service quality elements:
(1) Analysis of media: level of exposure depending on different media to extract target media
(2) Analysis of change between periods: used for analysis of service-related trends
(3) Analysis of major keywords: for the change patterns of major keywords
(4) Analysis of major influential figures: analysis of proliferation of issues

First, the present research conducted media analysis to perform profiling of target collection sites. The proposed system extracted and listed target sites specialized in each service sector.

To take an example, for the tourism industry, as shown in Figure 2, we extracted influential figures and powerful bloggers in the area of tours and traveling to use in examining recent trends and recent public opinion. To that end, we extracted knowledge-sharing sites to collect service-related information, opinion-sharing boards and questions and experts' recommendations. Major web sites for each service industry were selected and listed for collection of useful data.

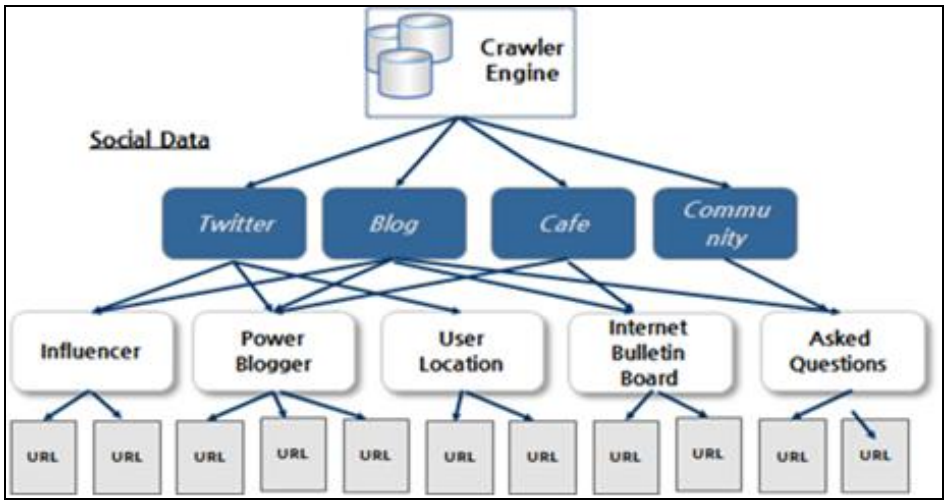

Figure 2. Target Collection Profiling

Also, continuous management of changes in collection sites needs to be done. That is, it is necessary to keep track of newly created, changed and disappearing web sites, to monitor the amount of collected information and tune the rules of collection filtering so that the quality of collected data can be maintained [8].

2.2. Data Pattern Analysis

The technique to detect and predict issues by analyzing social data related to the analysis targets is available by combining various data pattern analysis methods as follows[9]:

(1) Anomaly Detection: To detect unknown anomalous and irregular patterns compared with normal cases. After classifying and patternizing the results of basic analysis or the raw data collected according to given conditions such as subject words and attributes of interest, as shown in Figure 3, it detects anomalous patterns which deviate from the general tendency.

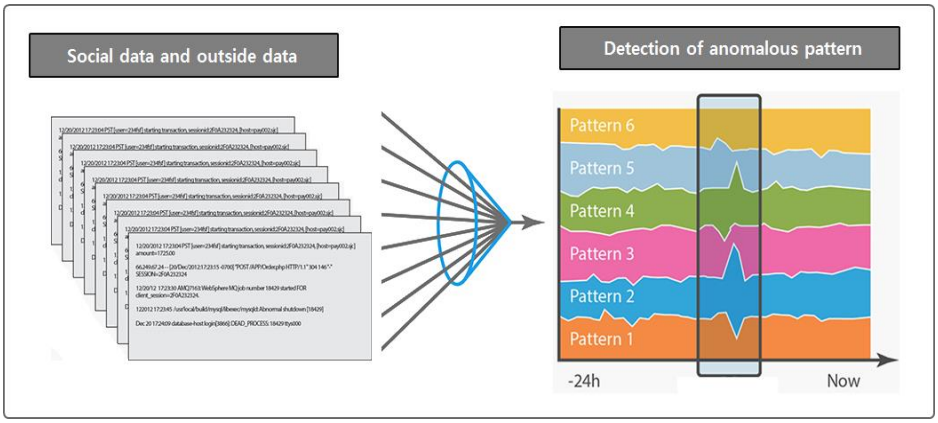

Figure 3. Detection of Anomalous Pattern by Anomaly Detection

(2) Predictive Models: To detect and analyze the issue information by discovering the occurrence patterns of the unknown complicated negative issues through diverse prediction methods such as hypothesis-based verification, prediction evaluation, etc. To predict the future trend based on 'Granger Causality' analysis between the actual issue occurrence trend and the social media data trend.

(3) Network Analysis: To detect unknown negative issues by correlation analysis using data such as social network, location information, etc.

(4) Text Mining: To detect negative issues by analyzing the meaning of untypical text of social media.

The model to detect and predict issues by combining those diverse methods is as Figure 4.

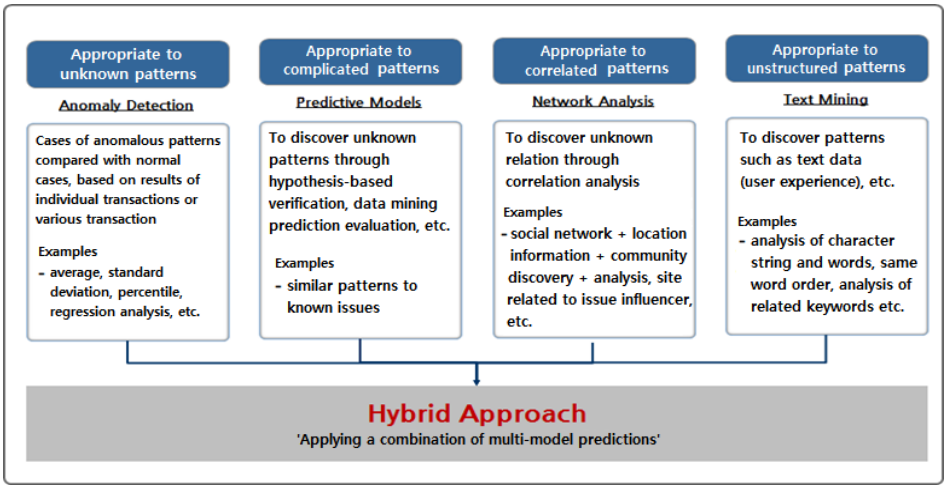

Figure 4. Hybrid Issue Detection and Prediction Model

3. Customer Feedback Analysis on Social Big Data

3.1. Target Collection Site Profiling

The method of profiling target collection sites specialized in a service sector is proposed for a more accurate and effective method of measuring customer satisfaction feedback.

The proposed system extracts a set of collection sites for each industry from a group of sites selected from a preliminary online survey, and those selected by a set of selection criteria. The process of examination and selection of collection sites is illustrated in Figure 5.

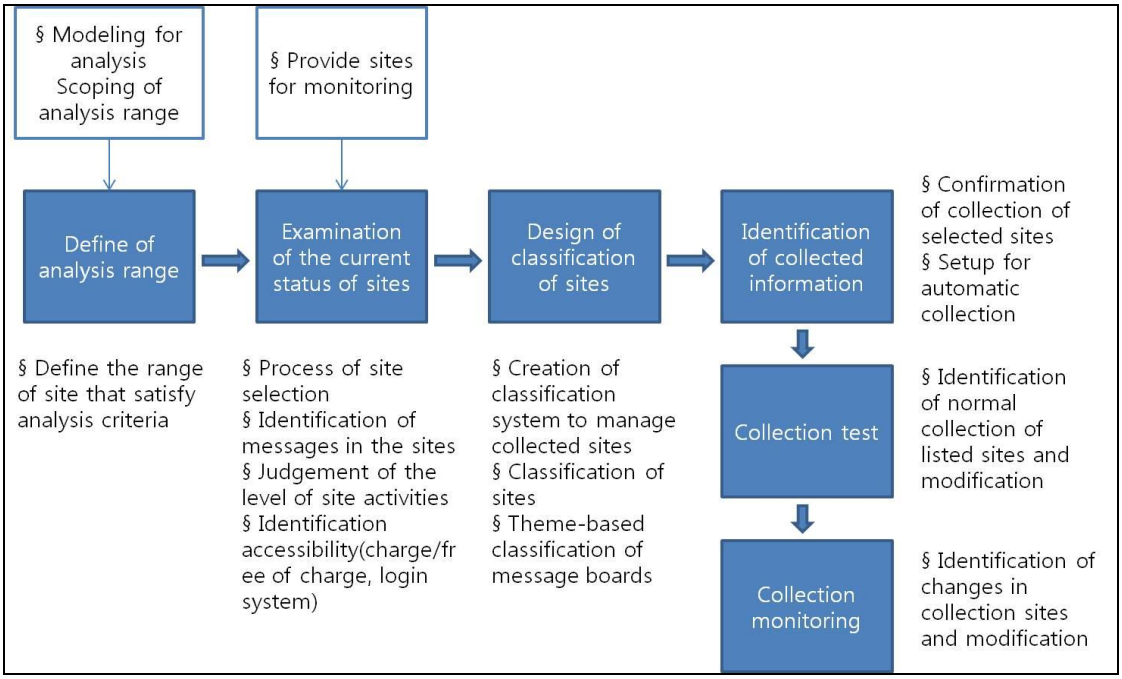

Figure 5. Examination of Collection Sites and the Selection Process

Such a process of examination and setup provides us with more accurate collection in major sites for each service industry, a minimized collection infrastructure (H/W) through extraction of target monitoring sites, and eventually a more effective operation.

The selection criteria for collection sites are given in Table 1.

Table 1. Selection Criteria for Collection Sites

Categories	Contents	Remarks for reference
Posting comments	Website names Board names	Extraction of major web sites for each industry
Information on message authors	Classification of message authors	
Site activities	Selection(articles) Number of members and new messages per day	-Identify the level of activities in the sites

	Number of hits	
	Reply comments	
Validity of site collection	Number of new messages in the past week	Identify the level of activities in the sites
	Existence of membership upgrading	Membership upgrade necessary in cafes
	Existence of login system	ID necessary for login
	Mechanical accessibility	Identify whether collection is impossible due to the complexity of sites

The proposed profiling process was applied to the tourism service industry. Collection intervals were decided based on the analysis of site activities including the characteristics of collection sites of a variety of categories on mass media. Weighted values of media were used on the basis of their influencing power. Power bloggers (experts in tours and restaurants) and outstanding bloggers were extracted and added for collection. The result was used as basis data to measure threshold values for application of weighted values. Approximately 3,000 target collection sites were added and Table 2 illustrates a set of extracted collection sites.

Table 2. Extraction of Target Sites

Categories	Types	# of collection sites
Mass media	Sites and major message boards	971
Blogs on Naver	Power bloggers (tours, restaurants, etc)	356
Blogs on Daum	Good bloggers (2007 ~ 2012)	1,668
Total		2,995

Verification of the results of target collection site profiling between June 1 and Oct. 31, 2013 showed that the number of collected data, compared to pre-profiling, increased by 52,000, which means that the population parameter of customers' comments actually increased on the social media.

Table 3. Increase in Collected Tourism-related Data

Categories	Before adding collection sites	After adding collection site	difference	Average collection per day
Mass Media	4,043	7,503	+3,460	250
Blogs	0	19,856	+19,856	661
Cafes	0	17,245	+17,245	574
Tweetter	0	11,399	+11,399	379
total	4,043	56,003	+51,960	1,866

3.2. Issue-detection Methodology

The process and method of issue detection are as follows:

(1) Selection of domain for development of predictive model: To check standardization, general purpose and availability of data on social media for extension and application to other domains. To check the continuity of issues and collect the social data of the target domain.

(2) To build a domain-specialized sensitivity dictionary: To build a subject-oriented sensitivity dictionary within domain and to define the major sensitivity category and to do sensitivity vocabulary mapping. To extract time series patterns as per sensitivity category.

(3) To analyze data patterns of target domain for issue prediction: To predict and detect anomalous patterns by analyzing data patterns of social media

(4) To detect issues: To define the independent variables as contents and writers which affect the spread of the negative public opinions, and the dependent variables as average reaching time and speed of the issues. To detect the influence of the negative public opinions based on the numbers of tweets and re-tweets.

Process of issue detection is as Figure 6.

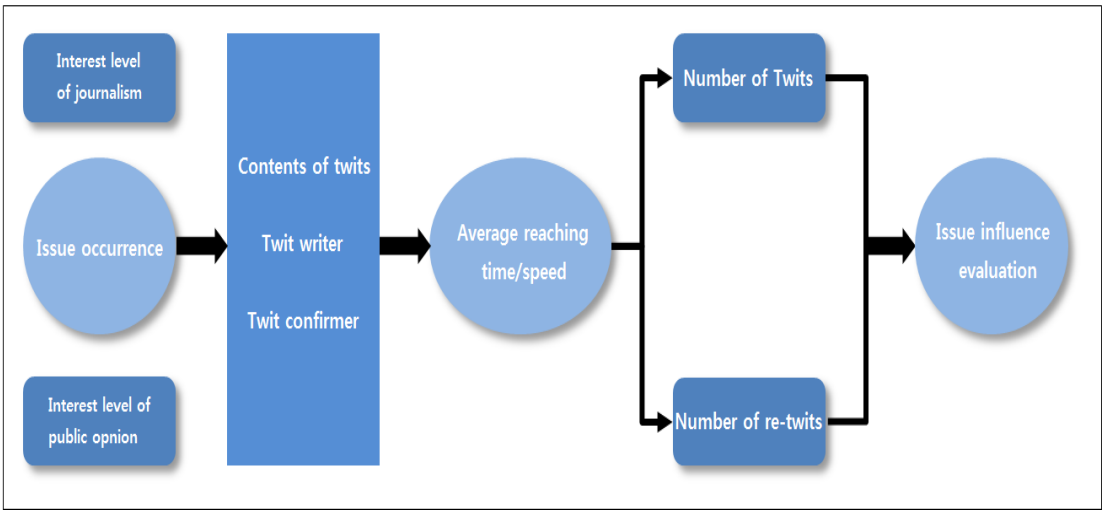

Figure. 6. Issue Detection Process

Actual application result of the proposed method is as follows:

(1) Tweet and media reports on key dates: The points of time when the amount of tweets and media reports increase sharply were similar, but the responsiveness of tweet and media to detailed events showed some differences. Anyway, the media and twitter had a mutual amplification effect.

(2) Ratio of re-tweet to tweet at time of issue amplification: In twitter the re-tweet ratio affects the spread or amplification of issues. Actually, the re-tweet ratio increased by 5~10% compared with the total average re-tweet ratio. The time of issue amplification could be detected as Figure 7.

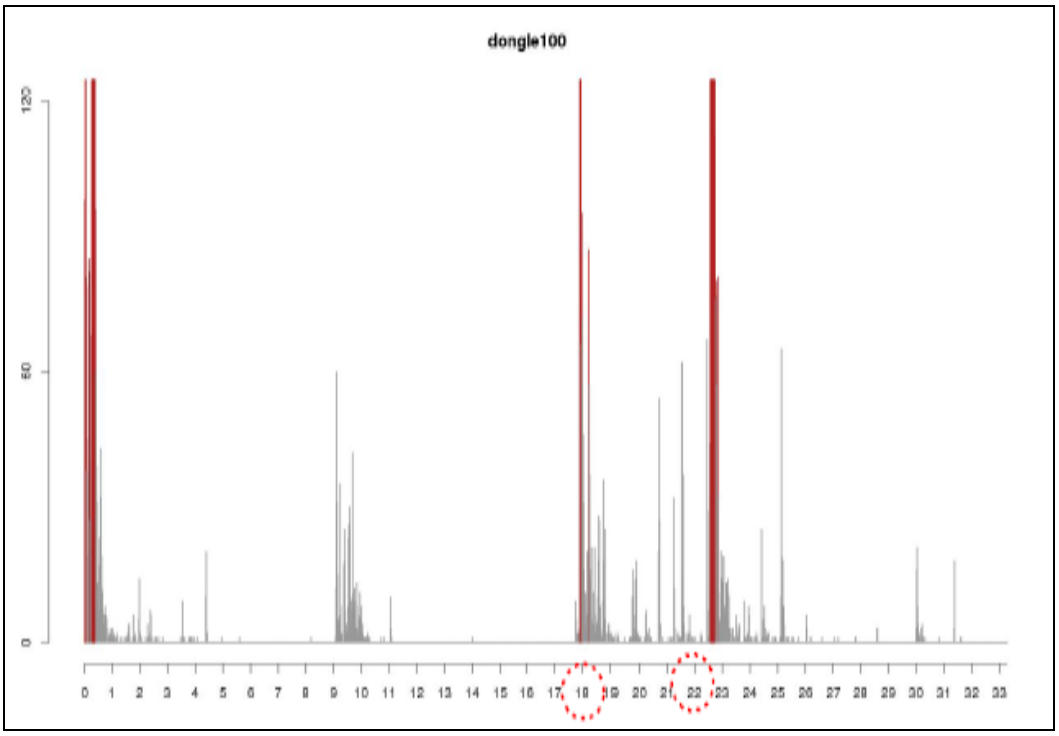

Figure 7. Detection of the Time of Issue Amplification

(3) Volume of tweets at time of issue amplification: Due to the increase of tweet amount by re-tweet, the elapse time of issue spread was very short, and in the case of the influencer's participation, the elapse time of issue amplification became very short. As for the tweets with many re-tweets, the 100 re-tweets were made on a large scale in a short time of 30 minutes average and the lapse time for 100 re-tweets was less than the double of the 50 re-tweets. If the re-tweet amount exceeded the critical value, then re-tweets were made more rapidly.

For example, the proposed method can predict the suicide rate by analyzing the frequency of keywords such as adolescent suicide number and suicide as in Figure 8.

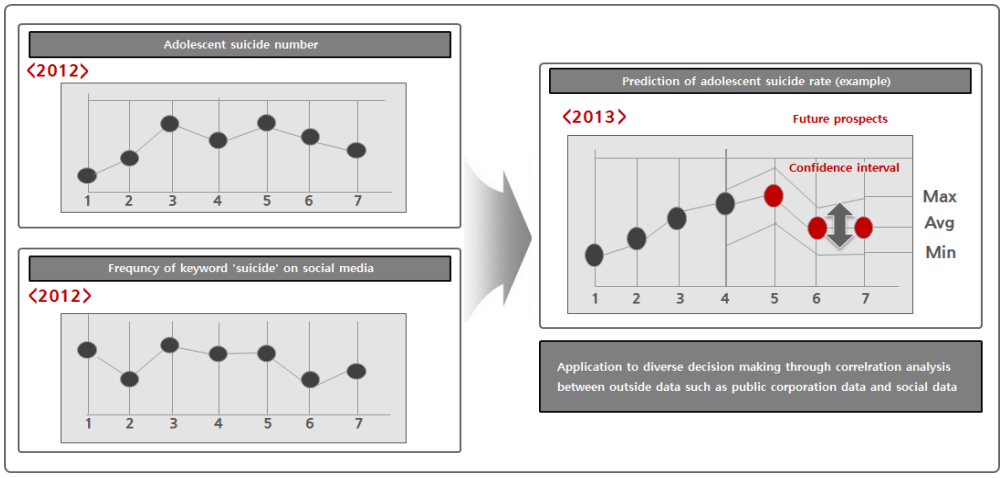

Figure 8. Example of Issue Detection and Prediction Analysis

4. Conclusions

A business organization has to exert every effort to respect clients' opinions, maintain a good relationship with them and keep providing quality products or services. To this end, it is extremely important to identify customers' responses and feedback. The role of big data on the web has recently been emphasized as an effective channel for real-time identification of customers' needs, since a great number of customers use such social media as blogs and SNS to express their opinions [10].

Recently IT companies have been competitively developing Social Big Data Analysis Tools. The Analysis Tools collect and accumulate the customers' opinions and analyze the contents using key words. The role of big data on the web has recently been emphasized as an effective channel for real time identification of customers' needs, since a great number of customers use such social media as blogs and SNS to express their opinions [11].

The Tools discover what kinds of public opinions are formed for the specific subjects, and how they are spread out.

This paper proposed a more effective and accurate method of profiling target collection sites that can collect big data on the social media and identify customers' feedback and responses. A set of major websites for each service industry are selected and listed to collect valid information. The experimental application of the system to the tourism industry found that the parameter of customers increased by 5,200 during the past five months.

It is expected that the present profiling model for target collection sites can be applied to other industries including medical industry than the tourism industry so that more accurate evaluation of customer reputation can be obtained [12].

Also, this study proposed a method which detects and predicts issues in the analysis of social Big Data. This method compares and analyzes the social media data pattern related to target institutions or corporations by combining diverse data pattern analysis tools based on Buzz Monitoring system which analyzes the Big Data.

The Issue-detection Methodology defines the independent variables as contents and writers which affect the spread of the negative public opinions and the dependent variables as average reaching time and speed of the issues. The influence of negative public opinions is detected concerning issues based on the numbers of tweets and re-tweets. The service providing corporations may prepare appropriate measures by the issue detection prior to the spread of the negative public opinions.

Acknowledgments

Funding for this paper was provided by Namseoul University.

References

[1] K. P. Nam, "System Implementation of the Customer Satisfaction Survey Using Internet", The Korean Journal of Applied Statistics, vol. 18, no. 3, **(2005),** pp. 713-727**.**
[2] J. S. Kim, "Big Data Analysis Methodology for Customer-oriented Services in Ubiquitous Environment" J. of KSCD (Korea Society of communication Design), vol. 19, **(2012),** pp. 131-145.
[3] J. Manyika, "Big Data: The next frontier for innovation, competition, and productivity", McKinsey Global Institute Report, **(2011).**
[4] E. J. Song, "A Study on the Case Analysis of Customer Reputation based on Big Data", J. of the Korean Institute of Industrial Engineer, vol. 17 no. 10, **(2013),** pp. 2439-2466.
[5] E. J. Song, S. J. Hong and M. S. Kang, "A Study on the System for Data Collection and Analysis on Social Network", Proceedings of International Workshop on Networking and Communication, ASTL vol. 44, **(2013).**
[6] Parasuraman, "SERVQUAL: A Multiple-Item Scale for Measuring Consumer Perceptions of Service Quality and Its Implication for Future Research", Journal of Retailing, vol. 64**, (1988).**

[7] M. S. Knag, "A Study on the Scoring Customer Feedback System for B2C Service", in Proceeding of the Korean Institute of Information and Communication Sciences Conference vol.17 no.1, 929-930**, (2013).**
[8] E. J. Song and M. S. Kang, "A Study on Collection Site Profiling for Analysis Custom Reputation", 2014 International Conference on Future Information & Communication Engineering vol. 6, no.1, **(2014)**, pp. 411-414.
[9] C. H. Lee, J. Hur and H. J. Oh, "Technology Trends of Issue Detection and Predictive Analysis on Social Big Data", ETRI Electronics and Telecommunications Trends, **(2013)**, pp. 62-71.
[10] K. Cheong, H. Y. Seo and S. D. Cho, "Classifications and Content Analyses of Social Networking Services Research", Journal of The Korean Knowledge Information Technology Society, vol. 6, no. 5, pp. 82-98 **(2011).**
[11] K. J. Park, "Big Data Eco System", J. of the Korean Institute of Industrial Engineers (KIIE), vol. 19, no. 3, **(2012),** pp. 41-47.
[12] J. T. Kim, "Analyses of Characteristics of U-Healthcare System Based on Wireless Communication", Journal of information and communication convergence engineering, vol. 10, no. 4, **(2012),** pp. 337-342.

Author

Eun-Jee Song received her B.S. degree from the Department of Mathematics, Sookmyung Women's University, in 1984. She earned her M.S. and Ph.D. degrees from the Department of Information Engineering, Nagoya University, Japan in 1988 and 1991, respectively. She was an exchange professor at the Department of Computer Science, the University of Auckland, New Zealand, in 2007. She is currently a full and tenured professor of the Department of Computer Science, Namseoul University, Cheonan, Korea.

