Social Event Radar: A Bilingual Context Mining and Sentiment Analysis Summarization System

Wen-Tai Hsieh
Department of IM,
National Taiwan University
wentai@iii.org.tw

Chen-Ming Wu
Institute for Information Industry
cmwu@iii.org.tw

Tsun Ku
Institute for Information Industry
cujing@iii.org.tw

Seng-cho T. Chou
Department of IM,
National Taiwan University
chou@im.ntu.edu.tw

Abstract

Social Event Radar is a new social networking-based service platform, that aim to alert as well as monitor any merchandise flaws, food-safety related issues, unexpected eruption of diseases or campaign issues towards to the Government, enterprises of any kind or election parties, through keyword expansion detection module, using bilingual sentiment opinion analysis tool kit to conclude the specific event social dashboard and deliver the outcome helping authorities to plan "risk control" strategy. With the rapid development of social network, people can now easily publish their opinions on the Internet. On the other hand, people can also obtain various opinions from others in a few seconds even though they do not know each other. A typical approach to obtain required information is to use a search engine with some relevant keywords. We thus take the social media and forum as our major data source and aim at collecting specific issues efficiently and effectively in this work.

1 Introduction

The primary function of S.E.R. technology is simple and clear: as a realtime risk control management technology to assist monitoring huge amount of new media related information and giving a warning for utility users' sake in efficiency way.

In general, S.E.R. technology constantly crawling all new media based information data relating to the client 24-hour a day so that the influential opinion/reports can be monitored, recorded, conveniently analyzed and more importantly is to send a warning signal before the issue outburst and ruining the authorities' reputation. These monitor and alert services are based on the socialnomics theory and provide two main sets of service functionalities to clients for access online: Monitor and alert of new media related information under the concept of cloud computing including two functionalities.

First functionality is the monitoring set. With the dramatic growth of Web's popularity, time becomes the most crucial factor. Monitoring functionalities of S.E.R. technology provides an access to the service platform realtime and online. All scalable mass social data coming from social network, forum, news portals, blogosphere of its login time, its social account and the content are monitored and recorded. In order to find key

Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics, pages 163–168,
Jeju, Republic of Korea, 8-14 July 2012.

opinion leaders and influential, the S.E.R. technology used social network influence analysis to identify a node and base on the recorded data to sort and analyze opinion trends statistics for every customer's needs.

Second functionality is alert module. Alert functionalities of the S.E.R. technology automatically give a warning text-messages or an e-mail within 6 hours whenever the golden intersection happened, meaning the 1-day moving average is higher than the 7-days moving average line, in order to plan its reaction scheme in early stage.

In empirical studies, we present our application of a Social Event Radar. We also use a practical case to illustrate our system which is applied in industries and society. The rest of this paper is organized as follows. Preliminaries and related works are reviewed in Section 2. The primary functionality and academic theory are mentioned in Section 3. Practical example and influence are explored in Section 4. S.E.R. detail operations are shown in Section 5. Finally, this paper concludes with Section 6.

2 Preliminaries

For the purpose of identifying the opinions in the blogosphere, First of all, mining in blog entries from the perspective of content and sentiment is explored in Section 2.1. Second, sentiment analysis in blog entries is discussed in Section 2.2. Third, information diffusion is mentioned in Section 2.3.

2.1 Topic Detection in Blog Entries

Even within the communities of similar interests, there are various topics discussed among people. In order to extract these subjects, cluster-liked methods Viermetz (2007) and Yoon (2009)are proposed to explore the interesting subjects.

Topic-based events may have high impacts on the articles in blogosphere. However, it is impossible to view all the topics because of the large amount. By using the technique of topic detection and tracking (Wang, 2008), the related stories can be identified with a stream of media. It is convenient for users who intend to see what is going on through the blogosphere. The subjects are not only classified in the first step, but also rank their importance to help user read these articles.

After decomposing a topic into a keyword set, a concept space is appropriate for representing relations among people, article and keywords. A concept space is graph of terms occurring within objects linked to each other by the frequency with which they occur together. Hsieh (2009) explored the possibility of discovering relations between tags and bookmarks in a folksonomy system. By applying concept space, the relationship of topic can be measured by two keyword sets.

Some researches calculate the similarity to identify the characteristic. One of the indicators is used to define the opinion in blog entries which is "Blogs tend to have certain levels of topic consistency among their blog entries." The indicator uses the KL distance to identify the similarity of blog entries (Song, 2007). However, the opinion blog is easy to read and do not change their blog topics iteratively, this is the key factor that similarity comparison can be applied on this feature.

2.2 Opinion Discovery in Blog Entries

The numbers of online comments on products or subjects grow rapidly. Although many comments are long, there are only a few sentences containing distinctive opinion. Sentiment analysis is often used to extract the opinions in blog pages.

Opinion can be recognized from various aspects such as a word. The semantic relationship between opinion expression and topic terms is emphasized (Bo, 2004). It means that using the polarity of positive and negative terms in order to present the sentiment tendency from a document.
Within a given topic, similarity approach is often used to classify the sentences as opinions. Similarity approach measures sentence similarity based on shared words and synonym words with each sentence in documents and makes an average score. According to the highest score, the sentences can assign to the sentiment or opinion category (Varlamis, 2008).

Subjectivity in natural language refers to aspects of language used to express opinions and evaluation. Subjectivity classification can prevent the polarity classifier from considering irrelevant misleading text. Subjectivity detection can compress comments into much shorter sentences which still retain its polarity information comparable to the entire comments (Rosario, 2004; Yu, 2003).

2.3 Information Diffusion in Internet

The phenomenon of information diffusion is studied through the observation of evolving social relationship among bloggers (Gill, 2004; Wang, 2007). It is noted that a social network forms with bloggers and corresponding subscription relationship.

Information diffusion always concerns with temporal evolution. The blog topics are generated in proportion to what happened in real world. Media focus stands for how frequently and recently is the topic reported by new websites. User attention represents how much do bloggers like to read news stories about the topic. By utilizing these two factors, the news topics are ranked within a certain news story (Wang, 2008).

The phenomenon of information diffusion is driven by outside stimulation from real world (Gruhl, 2004). It focuses on the propagation of topics from blog to blog. The phenomenon can discuss from two directions. One is topic-oriented model which provides a robust structure to the whole interesting terms that bloggers care about. The other is individual-oriented model which helps users figure out which blogger has information impact to others.

3 BUILDING BLOCKS OF S.E.R TECHNOLOGY

The core technology building block of S.E.R. technology is the central data processing system that currently sits in III's project processing center. This core software system is now complete with a set of processing software that keeps analyzing the recorded data to produce reports and analytical information, all those monitoring functionalities provided to subscribers.

Two important technology building blocks for the success of the S.E.R. are the bilingual sentiment opinion analysis (BSOA) technique, and social network influence analysis (SNIA) technique. These techniques are keys to the successful collection and monitoring of new media information, which in turn is essential for identifying the key opinion web-leaders and influential intelligently. The following sections apply the academic theory combining with practical functionality into the S.E.R.

3.1 Bilingual Sentiment Opinion Analysis

BSOA technique under the S.E.R. technology is implemented along with lexicon based and domain knowledge. The research team starts with concept expansion technique for building up a measurable keyword network. By applying particularly Polysemy Processing Double negation Processing Adverb of Degree Processing sophisticated algorithm as shown in Figure 1, so that to rule out the irrelevant factors in an accurate and efficiency way.

Aim at the Chinese applications; we develop the system algorithm based on the specialty of Chinese language. The key approach crawl the hidden sentiment linking words, and then to build the association set. We can, therefore, identify feature-oriented sentiment orientation of opinion more conveniently and accurately by using this association set analysis.

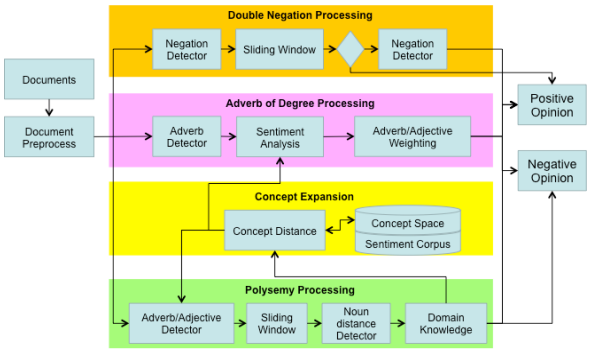

Figure 1. Bilingual Sentiment Opinion Analysis

3.2 Social Network Influence Analysis

Who are the key opinion leaders in the opinion world? How critical do the leaders diffusion power matters? Who do they influence? The more information we have, so as the social networking channels, the more obstacles of monitoring and finding the real influential we are facing right now.

Within a vast computer network, the individual computers are on what so-called the periphery of the network. Those nodes who have many links pointing to them is not always the most influential in the group. We use a more sophisticated algorithm that takes into account both the direct and indirect links in the network. This SNIA technique under the S.E.R. technology provides a more accurate evaluation and prediction of who really influences thought and affects the whole. Using the same algorithm, in reverse, we can

quickly show the direct and indirect influence clusters of each key opinion leader.

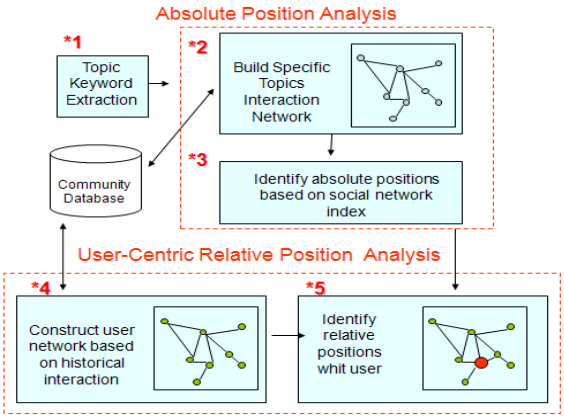

Figure 2. Social Network Influence Analysis

3.3 The monitoring methodology of agenda-tendency

In the web-society, the system architecture of monitoring and identifying on vast web reviews is one thing, being aware of when to start the risk control action plan is another story. We develop 3 different forms of analysis charts — long term average moving line，tendency line，1-day, 7-day and monthly average moving line. For example, the moment when the 1-day moving average line is higher than the 7-day moving average line, it means the undiscovered issue is going to be outburst shortly, and it is the time the authority to take action dealing with the consequences. One news report reconfirmed that a wrong manipulated marketing promotion program using an “iPhone5” smart-phone as its complementary gift and was shown on the analysis chart 9 days before it revealed on the television news causing the company’s reputation being damaged badly.

4 PRATICAL EXAMPLE

To make our proposed scheme into practice, corresponding systems are applying in the following example. S.E.R. plays an important role to support the enterprise, government and public society.

4.1 Food-safety Related Issues

S.E.R. research and development team built up the DEPH [di(2-ethylhexyl)phthalate] searching website within 2 days and made an officially announcement in June. 1st, 2011 under the pressure of the outbreak of Taiwan’s food contamination storm, which in general estimated causing NT$10,000 million approximately profit lost in Taiwan’s food industry. This DEPH website was to use the S.E.R. technology not only to collect 5 authorities’ data (Food and Drug Administration of Health Department in Executive Yuan, Taipei City government) 24 hours a day but also gathering 3 news portals — Google, Yahoo, and UDN, 303 web online the latest news information approx., allowed every personal could instantly check whether their everyday food/drink has or failed passing the toxin examination by simply key-in any related words (jelly, orange juice, bubble tea). This website was highly recommended by the Ministry of Economic Affairs because of it fundamentally eased people’s fear at the time.

4.2 Brand/Product Monitoring

A world leading smart phone company applying the S.E.R Technology service platform to set up its customer relationship management (CRM) platform for identifying the undiscovered product-defects issues, monitoring the web-opinion trends that targeting issues between its own and competitor’s products/services mostly. This data processing and analyzing cost was accordingly estimated saving 70 % cost approximately.

4.3 Online to Offline Marketing

In order to develop new business in the word-of-mouth market, Lion Travel which is the biggest travel agency in Taiwan sat up a branch “Xinmedia”. The first important thing for a new company to enter the word-of-mouth market is to own a sufficient number of experts who can affect most people’s opinion to advertisers, however, this is a hard work right now. S.E.R. helps Xinmedia to easily find many traveling opinion leader, and those leaders can be products spokesperson to more accurately meet the business needs. More and more advertisers agree the importance of the word-of-mouth market, because Xinmedia do created better accomplishments for advertisers’ sales by experts’ opinion.

5 S.E.R. DETAIL OPERATIONS

In the following scenario, S.E.R. monitors more than twenty smartphone forums. In Figure 3, the cellphone "One X" is getting popular than others. From the news, we know this cellphone is upcoming release to the market and it becomes a topical subject.

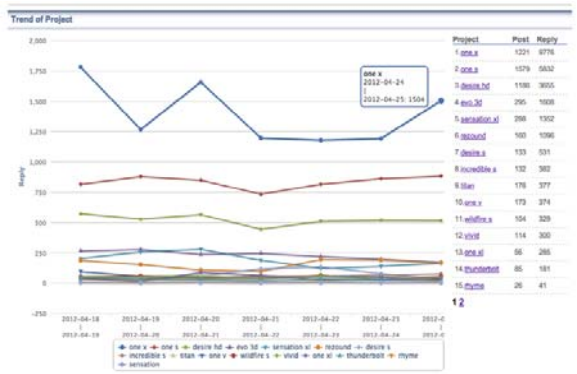

Figure 3: An example of Word-of-mouth of products

Beyond the products, some details are discussed with product in a topic. Thus, we use TF-IDF and fixed keyword to extract the important issue. These issues are coordinated with time slice and generated dynamically. It points out the most discussed issue with the product. In Figure 4, In this case, the "screen" issue is raising up after "ics" (ice cream sandwich, an android software version) may become the most concern issue that people care about.

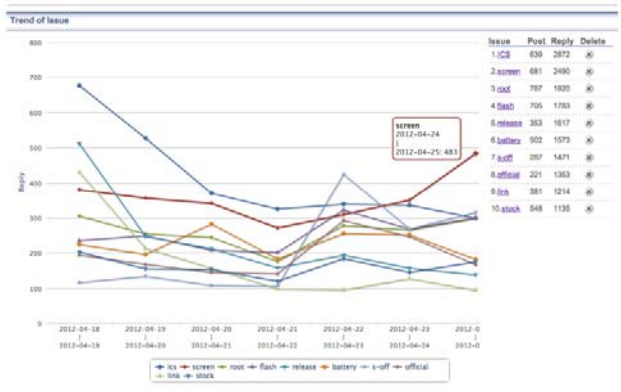

Figure 4. An example of hot topics

For different project, S.E.R. supports the training mode to assist user to train their specific domain knowledge. User can easily tag their important keyword to their customized category. With this benefit, we can accept different domain source and do not afraid data anomaly. We also apply training mechanism automatically if the tagging word arrive the training standard.

As shown in Figure 5, top side shows the analyzed information of whole topic thread. We just show the first post of this thread. As we can see, we provide three training mode, Category, Sentiment and Same Problem. The red word shows the positive sentiment and blue word shows the negative sentiment respectively. The special case is the "Same Problem". In forum, some author may just type "+1", "me2", "me too" to show they face the same problem. Therefore, we have to identify what they agreed or what they said. We solve this problem by using the relation between the same problem word and its name entity.

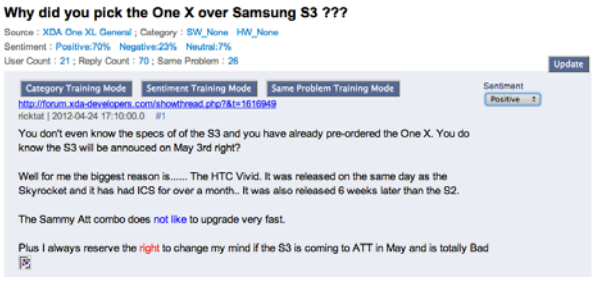

Figure 5: Training Mode – S.E.R. supports category training, sentiment training and same problem training

To senior manager, they may not spend times on detail issue. S.E.R. provides a quick summary of relevant issue into a cluster and shows a ratio to indicate which issue is important.

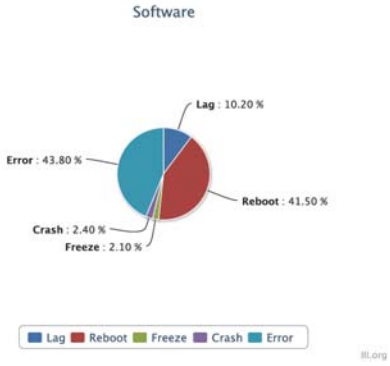

Figure 6. Quick Summary – Software relevant issues

6 Conclusions

In this networked era, known social issues get monitored and analyzed over the Net. Information gathering and analysis over Internet have become

so important for efficient and effective responses to social events. S.E.R technology is an Internet mining technology that detects monitors and analyzes more than Net-related social incidents.

An impending event that's not yet attracted any attention, regardless of whether of known nature or of undetermined characteristic, gets lit up on the S.E.R radar screen – provided a relevant set of detection conditions are set in the S.E.R engine. S.E.R technology, like its related conventional counterparts, is certainly capable for monitoring and analyzing commercial and social, public events. It is the idea "to detect something uncertain out there" that distinguishes S.E.R from others.

It is also the same idea that is potentially capable of saving big financially for our society. It may seem to be – in fact it is – hindsight to talk about the DEPH food contamination incident of Taiwan in 2011, discussing how it would have been detected using this technology. But, the "morning-after case analysis" provides a good lesson to suggest that additional tests are worthwhile – thus the look into another issue of food additives: the curdlan gum.

Certainly there is – at this stage – not yet any example of successful uncovering of impending events of significant social impact by this technology, but with proper setting of an S.E.R engine by a set of adequate parameters, the team is confident that S.E.R will eventually reveal something astonishing – and helpful to our society.

7 Acknowledgments

This study is conducted under the "Social Intelligence Analysis Service Platform" project of the Institute for Information Industry which is subsidized by the Ministry of Economy Affairs of the Republic of China.

8 References

Hsieh, W.-T., Jay Stu, Chen, Y.-L., Seng-cho Timothy Chou. 2009. A collaborative desktop tagging system for group knowledge management based on concept space. Expert Syst. Appl., 36(5), 9513-9523

Wang, J.-C., Chiang, M.-J., Ho, J.-C., Hsieh, W.-T., Huang, I.-K.. 2007. Knowing Who to Know in Knowledge Sharing Communities: A Social Network Analysis Approach, In Proceeding of The Seventh International Conference on Electronic Business, 345-351.

Bo, P. and Lillian, L. 2004. A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts. In Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics, 271-279.

Gill, K. E. 2004. How can We Measure the Influence of the Blogosphere, In Proceedings of the 2nd Workshop on the Weblogging Ecosystem, 17-22.

Gruhl, D., Guha, R., Liben-Nowell, D. and Tomkins, A. 2004. Information Diffusion through Blogspace, In Proceedings of the 13th International Conference on World Wide Web, 491-501.

Rosario, B. and Hearst, M. A. 2004. Classifying Semantic Relations in Bioscience Texts, In Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics, 430.

Song, X., Chi, Y., Hino, K., and Tseng, B. Identifying Opinion Leaders in the Blogosphere. 2007. In Proceedings of the 16th ACM conference on Conference on Information and Knowledge Management, 971-974.

Varlamis, I., Vassalos, V., and Palaios, A. 2008. Monitoring the Evolution of Interests in the Blogosphere. In Proceedings of the 24th International Conference on Data Engineering Workshops.

Viermetz, M. and Skubacz, M. 2007. Using Topic Discovery to Segment Large Communication Graphs for Social Network Analysis, In Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence, 95-99.

Wang, C., Zhang, M., Ru, L., and Ma, S. 2008. Automatic Online News Topic Ranking Using Media Focus and User Attention based on Aging Theory, In Proceeding of the 17th ACM Conference on Information and Knowledge Management, 1033-1042.

Yoon, S.-H., Shin, J.-H., Kim, S.-W., and Park, S. 2009. Extraction of a Latent Blog Community based on Subject, In Proceeding of the 18th ACM Conference on Information and Knowledge Management, 1529-1532.

Yu, H and Hatzivassiloglou, V. 2003. Towards Answering Opinion Questions: Separating Facts from Opinions and Identifying the Polarity of Opinion Sentences, In Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing - Volume 10, 129-136.