Semantics Extraction from Social Computing: A Framework of Reputation Analysis on Buzz Marketing Sites

Takako Hashimoto[1] and Yukari Shirota[2]

[1] Chiba University of Commerce, 1-3-1 Konodai, Ichikawa, Chiba, Japan
[2] Gakushuin University, 1-5-1 Mejiro, Toshima-ku, Tokyo, Japan
takako@cuc.ac.jp,
yukari.shirota@gakushuin.ac.jp

Abstract. Social computing services, which enable people to easily communicate and effectively share the information through the Web, have rapidly spread recently. In the marketing research domain, buzz marketing sites as social computing services have become important in recognizing the reputation of products hold with users. This paper proposes a reputation analysis framework for the buzz marketing sites. Our framework consists of four steps: the first is to extract the topics of the product using natural language processing. The input data comprises consumer messages on buzz marketing sites. Next, important topics on the products are extracted. The third step is to detect emerging consumer needs by identifying new burst topics. Finally, the results are visualized. Based on our framework, product characteristics and emerging consumer needs are extracted and reputations are visualized.

Keywords: Web intelligence, Social Computing, Buzz Marketing, Data Mining.

1 Introduction

Social computing services like blogs, SNSs (social networking services) and buzz marketing sites, which enable people to easily communicate and effectively share the information through the Web, have rapidly spread. We can say that communications in social computing services have generated new consensuses and new intelligence. In buzz marketing sites especially, varied consumers write review messages about a product. They also add their comments on others' messages. These communications affect consumer behavior. Social computing services have become highly-influential in the marketing research domain. In this environment, reputation analysis from messages in social computing services has become significant.

The purpose of this paper is to propose a reputation analysis framework to extract product characteristics and analyze consumer needs from the messages on buzz marketing sites. Our system targets both consumers and marketing planners. It's an application which provides the potential use in marketing and

S. Kikuchi, S. Sachdeva, and S. Bhalla (Eds.): DNIS 2010, LNCS 5999, pp. 244–255, 2010.

an improvement of products by manufactures. Through our framework, they can identify product characteristics and recognize emerging consumer needs related to the products.

Our framework consists of four steps: the first is to extract the topics of the product using natural language processing. The input data comprises consumer messages on buzz marketing sites. Next, important topics on the products are extracted. The third step is to detect emerging consumer needs by identifying new burst topics. Finally, the results are visualized.

The remainder of the paper is structured as follows: Chapter 2 describes the research background and related work. Chapter 3 shows the preliminary survey results on buzz marketing sites. Chapter 4 proposes the reputation analysis framework on buzz marketing sites. Chapter 5 concludes the paper and sets a path for future work.

2 Background and Related Work

A reputation analysis is of special concern in the study of web intelligence. In the advanced networked society which provides different communications through social computing services, a reputation analysis from various information sources like blog and buzz marketing sites, has become a key technology. In general, the main purpose of a reputation analysis is positive/negative comments detectin and their visualization. User utilize the results from a reputation analysis for their decision-making.

There are several reputation analysis systems [1], [2], [3]. Existing services focus on the detection of positive/negative comments, as pointed out above. Our framework also takes a similar approach. In addition, we plan to forecast the emerging needs by finding the trend in the real world. That is to say, our proposed reputation analysis framework are aiming to achieve the more integrative technique.

To estimate the positive/negative degree of a certain message, there is a method to clarify sentences with the sentiment in the document [4]. We plan to use this technique to assess the positive/negative degree. A method to clarify the pros and cons of the sentences is also proposed [5]. This work uses personality characteristics to improve accuracy. Our framework is due to use similar personality characteristics, not only to detect the pros and cons, but also to extract other semantics.

To extract the reputation, several techniques are available to detect positive/negative degree using the dictionary. The key point is to construct an efficient dictionary. Kamps et al. proposed a method to calculate the distance between good and bad for the word [6]. A co-occurrence dictionary is also used to detect positive/negative degree [7], [8]. We also plan to use a co-occurrence dictionary to extract the topics and detect positive/negative degree more precisely. We will evaluate existing techniques to meet our purpose in the next phase of our research.

Regarding document clustering, a method to form the clusters according to the positive/negative degree is proposed [9]. We cluster by clarifying documents based on topics. Reputation analysis is then processed according to the positive/negative degree.

3 Survey of Buzz Marketing Sites

We begin by looking at messages on the buzz marketing sites. From among the buzz marketing sites on the Internet, for this paper, we chose kakaku.com [10] as our example of buzz marketing sites. The kakaku.com site is the most popular 'customer purchasing support site' in Japan. It provides price information on electrical appliances, vehicles, toys, and various other products. Buzz marketing sites and shopping malls are also provided. Around 20 million user accesses per month were recorded as of September 2009.

As a preliminary experiment, we surveyed messages on the buzz marketing sites of kakaku.com. Each site for individual products are opened in the buzz marketing sites on kakaku.com. Consumers communicate with each other by adding their messages on the site. We read these messages and classified them to confirm whether or not product characteristics and emerging consumer needs can be extracted.

3.1 Messages on Front Loading Washing Machines with Automatic Drying System

We first checked the messages on three models of front loading washing machines with automatic drying. Front loading washing machines with automatic drying have become popular with Japanese families in recent years. However, their large size and vibration noise have been problems, and prices are still high. Therefore, many messages from consumers appear on the buzz marketing sites. As a preliminary experiment, we did the following:

1. Read all messages on corresponding machines
2. Extract the main topic of each message
3. Extract the feature of focus in each topic
4. Count the number of messages in each topic

Table 1 lists the topics, the feature terms and the number of messages on the three types of the washing machines we selected. The following eight topics are abstracted according to the messages: *Installation, Noise, Cleaning performance, Dryer performance, Price, Comparison with other models, Troubles* and *Other*. For each topic, we also derived feature terms (Table 1).

Figure 1 shows the number of messages on each topic using a cobweb chart. For the overall products (Model A, Model B and Model C), users seem to talk about *Comparison with other models*. In addition, depending on the machine, the number of messages differs with each topic. For example, for the Model C, there are many messages about *Troubles*. This indicates that users actively talk

Table 1. Topics, feature terms, and the number of messages for front loading washing machines with automatic drying system (Messages were collected on Nov. 12, 2009)

Topic	Feature Terms	Products (appearance time in the market)		
		Model A (Nov. 2008)	Model B (Oct. 2008)	Model C (Jun. 2009)
Installation	Size, Space, Width, Height, Measurement, Self-install	82	50	9
Noise	Noise, Vibration	27	14	20
Cleaning performance	Damage, Rip, Discoloration, Fluff, Rejuvenation, Stain	43	21	56
Dryer performance	Odors, Wrinkles, Shrinking, Fabric Care, Speed	66	35	6
Price	Expenditure, Cheap, Stock, "Shop A", "Shop B", "Shop C"	48	89	44
Comparison with other models	"Company A", gCompany B", "Company C", Predecessor	96	126	37
Troubles	Error Signals, Bugs, Nogood	30	45	105
Other		56	37	50
	Total	448	417	327

about Troubles on Model C. On the other hand, with the Model A, messages on *Installation* appear more often. Actually, the size of Model A is smaller than Model B and Model C. Users seems to be interested in Model A's easy installation. Model B has many messeages on *Price*. In fact, the price of Model B is slightly higher than Model A and Model C. Therefore, on Model B, users are concerned with the topic about price. It seems reasonable to suppose that these biases show the features of each machine and become a key for analyzing the product's reputation with consumers.

3.2 Messages on Electronic Air Cleaners

Table 2 lists the topics, feature terms and number of messages for the five types of electronic air cleaners we selected. Two types (Model G and Model H)came to

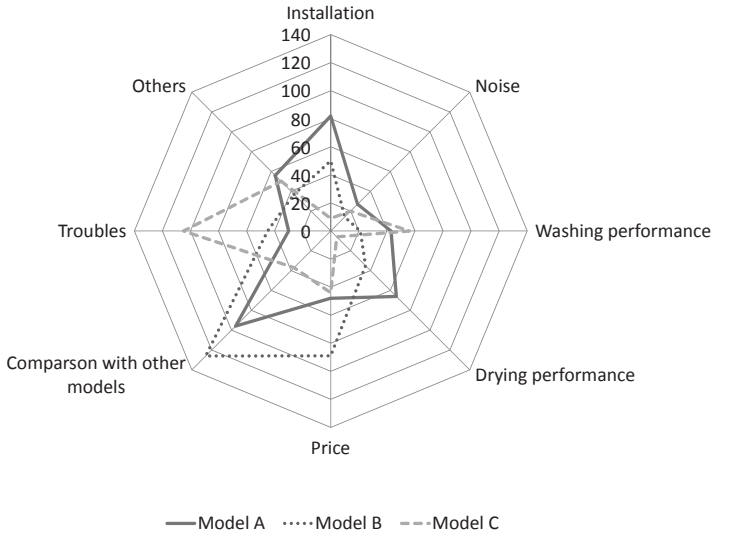

Fig. 1. Number of messages on each topic for corresponding products (Washing machines)

market at around the same time (in 2008) and the other three(Model D, Model E and Model F) entered about a year later (in 2009). Just as with the washing machines, we read all messages, classified the topics, extracted feature terms for each topic and manually counted the number of messages. Figure 2 shows the number of messages on each topic using a cobweb chart.

With the electronic air cleaners, topics such as *Installation, Noise, Cleaning performance, Price, Comparison with other models, Troubles* and *Other* appear just as with the washing machines. For the overall products (Model D, Model E, Model F, Model G and Model H), users seem to talk about *Prices*. In addition, for example, on Model G, there are many messages about *Troubles*. This indicates that users actively talk about troubles on Model G. Model F and Model H especially have many messages on *Price*. On Model F and Model H, users are concerned with the topic about price.

Beyond that, a topic related to *New influenza* appears for models that came to market in 2009. The reason for this burst on the new influenza seems to be that news of the new influenza rapidly grew in the spring of 2009. We believe that the detection of this kind of the bursty topic could be useful in identifying emerging consumer needs.

For burst detection, we propose a framework to check consumer messages periodically in the next chapter.

Table 2. Topics, feature terms, and the number of messages for electronic air cleaners (Messages were collected on Nov. 12, 2009)

Topic	Feature Terms	Products (appearance time in the market)				
		Model D (Sep. 2009)	Model E (Sep. 2009)	Model F (Sep. 2009)	Model G (Nov. 2008)	Model H (Sep. 2008)
Installation	Size, Space	4	0	1	0	0
Noise	Noise	8	4	3	2	0
Air cleaning performance	Humidification, , Setting, Ion	17	21	15	19	19
Price	Expenditure, Cheap, Stock, "Shop A", "Shop B"	25	22	31	8	33
Comparison with other models	"Company A", "Company B"	21	16	12	15	12
Troubles	Feed-water, Tank, Smell	8	0	11	10	26
New influenza	Virus, Prevention	19	23	14	0	0
Other		10	0	6	9	0
	Total	112	86	93	79	74

4 Framework for Reputation Analysis on Buzz Marketing Sites

This chapter discusses a reputation analysis framework to extract product characteristics and analyze emerging consumer needs from messages on buzz marketing sites.

Our framework consists of the following four steps:

1. Topic extraction
2. Important topic detection
3. Emerging needs detection
4. Visualization

The following sections describe each step.

4.1 Topic Extraction

In this step, input data is the set of messages about products, e.g., front loading washing machines on buzz marketing sites. We define one message as one

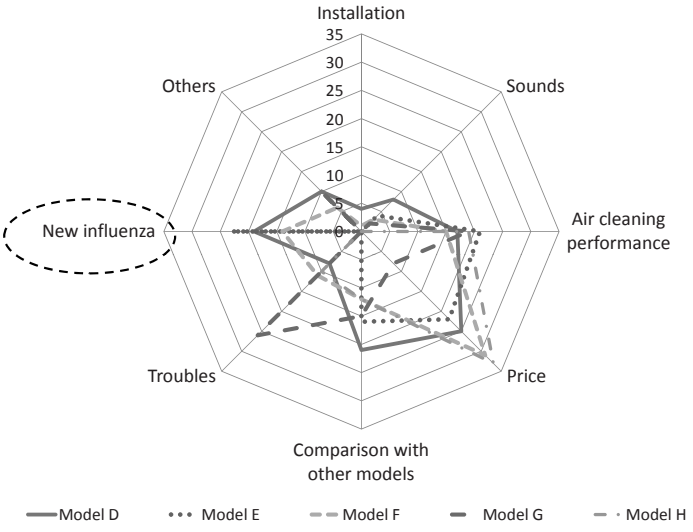

Fig. 2. Number of messages on each topic for corresponding products (Electronic Air Cleaners)

document and extract words using morphological analysis; we then calculate the tf-idf value for each word. We form the document clusters based on vectors of the tf-idf values using the cosine similarity [11]. Each cluster equals the topic of corresponding products. With topic extraction, we plan to use a concept-based co-occurrence dictionary [12] to improve the accuracy of document clustering.

4.2 Important Topic Detection

Important topics fall into two categories. One is important topics for the set of corresponding products, which express characteristics of the overall corresponding products. The other comprises important topics for individual products, which express the characteristics of each product. For example, consider washing machines, the topic related to *Comparison with other machines* is active for overall machines. For example, consider washing machines. The topic related to *Comparison with other machines* is active as a whole. On the other hand, *Troubles* is the bursty topic for Model C.

With reputation analysis, it is important to detect both categories of significant topics. We define the following three parameters to express the important topic for an individual product:

1. Contributing rate of each topic for a set of products P

$$CRPT_j = \frac{\sum_{i=1}^{l} m_{i,j}}{\sum_{j=1}^{n}(\sum_{i=1}^{l} m_{i,j})}$$

$CRPT_j$ expresses the important topic on the products as the overall trend. For example, the topic *Comparison with other models* would be the most important topic, because the number of the messages for it was the biggest (Table 1).
Where l is the number of products for a set of products P. P consists of several models (for example, a set of electronic air cleaners). n is the number of topics for P. $m_{i,j}$ is the number of messages about a topic T_j on a product P_i. $CRPT_j$ is the contribution rate of T_j on P . If T_j is important for P, $\sum_{i=1}^{l} m_{i,j}$ tends to be high. That is to say, $CRPT_j$ will be high.

2. Contributing rate of each topic on individual product

$$CRP_iT_j = \frac{m_{i,j}}{\sum_{j=1}^{n} m_{i,j}}$$

CRP_iT_j is the contribution rate of T_j on P_i .
CRP_iT_j expresses the main topic for individual products. We assume that significant topics differ from product to product. For example, for the Model C, there are many messages about *Troubles* (Table 1). The topic related to *Troubles* may express the characteristics of Model C. As a result, CRP_iT_j becomes high. This indicates that *Troubles* is significant on Model C.

Topics with high contribution rates ($CRPT_j$ and CRP_iT_j) are detected as important topics.

3. Positive/negative degree of topic on individual products

To analyze reputation, it is important to identify the positive/negative degree of each topic for the product. For example, the topic related to *Troubles* is significant on Model C. The question is whether the messages about *Troubles* are positive or negative.
We plan to use a dictionary (ontology) consisting of words with positive/negative values and a method to derive the positive/negative degree. Turney et al. proposed the method to derive positive/negative degree according to the semantic orientation of the phrases in the review that contain adjectives

or adverbs[13]. And Takamura et al. provide the dictionary on semantic orientations (positive/negative values) of words through the Web[14]. Based on this dictionary and Turney's method, we will calculate the positive/negative degree as follows :

$$PNP_iT_j = \frac{1}{q}\sum_{k=1}^{q} f(P_i, T_j, w_k)$$
$$q = Q(P_i, T_j)$$
$$w = W(P_i, T_j)$$

Where PNP_iT_j is the positive/negative degree of T_j of P_i. q is the number of words in T_j of P_i. $Q(P_i, T_j)$ is the function to calculate the number of words of ontology in T_j of P_i. The list of ontology words w for T_j of $P_i$is derived from the function $W(P_i, T_j)$. $f(P_i, T_j, w_k)$ is the function to extract the positive/negative value for each word w_k appearing in T_j of P_i based on the dictionary on semantic orientations (positive/negative values) of words[14]. PNP_iT_j is the mean value of the positive/negative values of words derived from T_j . The positive/negative degree of T_j of P_i depends on the sign of PNP_iT_j. For example, with the Model A, there are actually many negative words about *Dryer performance*(Table 1). Therefore PNP_iT_j of *Dryer performance* for the Model A may be a negative number.

Topics with high positive/negative degrees (PNP_iT_j) are also detected as important topics.

4.3 Emerging Needs Detection

To detect emerging needs, we plan to find changes in topics. In our framework, the messages are acquired periodically (e.g., once a week), and then the document clusters are formed. If a new cluster is generated, we recognize that new changes (or new needs) have been detected. For example, the topic related to *New influenza* can be found for models that came to market in 2009 (Table 2). We can say that the new needs about the new influenza have appeared. We assume that these new needs will be propagated to other products. That is to say, the needs about the new influenza will appear in other product as the new needs in the near future as well.

4.4 Visualization

Based on important topic extraction and emerging needs detection, the results are visualized to users. Figure 1 and Figure 2 are examples of visualization. However, the visualization method has not yet been perfected. We continue to discuss visualization in detail.

4.5 System Structure

Figure 3 shows the system structure of our proposed framework. The system consists of the following four modules and one database: the topic extraction module,

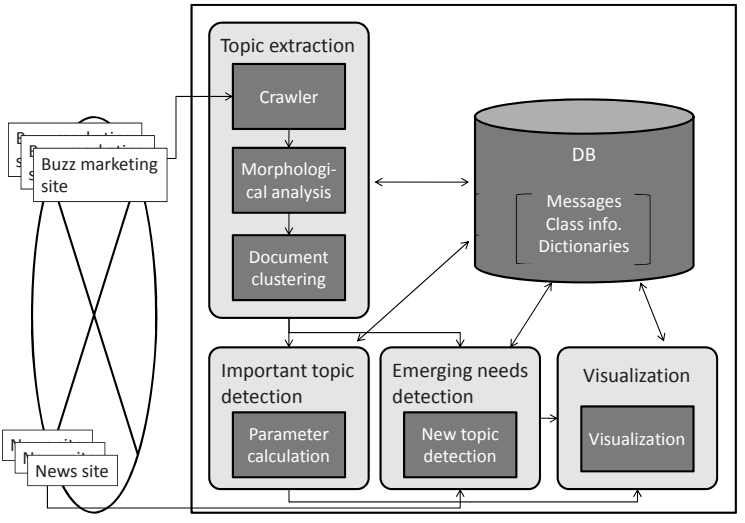

Fig. 3. System structure of our proposed framework

the important topic detection module, the emerging needs detection module, and the visualization module. The database has crawled messages, document class information, and the dictionaries. The topic extraction module acquires messages from the buzz marketing sites. It then extracts words using morphological analysis and forms the document clusters. To improve the accuracy of topic extraction, the module refers to the dictionaries in the database. The main process for the important topic detection module is to calculate the parameters we proposed in Chapter 3. Results of the calculation are stored on the database. The emerging needs detection module detects new burst topics. This module is executed concurrently with the important topic detection module. To precisely detect a new topic, the module refers to the news site and acquires candidates for the new topic. The visualization module visualizes the result of both the important topic detection module and the emerging needs detection module.

4.6 Problems with Reputation Analysis

Based on the above, we have found the following problems with reputation analysis.

1. Ontology provision:
   To improve the accuracy for topic extraction, it is necessary to prepare the high quality ontology. Furthermore, we have to consider the mechanism for ontology updating to detect emerging terms like the new influenza. Ontology generation and update are key issues with reputation analysis.

2. Credibility of the topic:
   With reputation analysis, credibility is the key issue. Even if there is a significant topic for a certain product, it is not always true. For example, the topic on *Troubles* is active in Model C (Table 1), but the main subject of the topic about troubles was misuse by the user, and many emotional messages appeared. In a case like this, it seems reasonable to suppose that credibility of the topic is low.

3. Burst products vs. non-burst products:
   The number of messages on the buzz marketing sites depends on the product. Of course, sales success seems to affect burstiness of the site. However, the degree of sentiment also seems to have an influence. We plan to analyze the reason for bursty products by taking into account emotional characteristics.

4. Roles in discussion:
   It is important to consider what kind of roles should appear in the discussion. Regarding the bursty discussion, both the agitator and the adversary are important. The follower seems to be significant as well. For reputation analysis, we have to look at the roles in the discussion.

5 Conclusion

In this paper, we proposed a framework for reputation analysis on buzz marketing sites. The purpose of our framework is to extract product characteristics and analyze consumer needs. Our framework extracts topics for the products using natural language processing of consumer messages. Important topics for the product are then detected. Emerging consumer needs are also detected. Finally, the results are visualized for the user. Our next step is to experiment with evaluating the effectiveness of our framework. Based on results, we will then improve the framework. At the same time, we will work on solving other issues, such as ontology provision, message credibility, burst discussion detection, and role detection on the buzz marketing site.

References

1. Morinaga, S., Yamanishi, K., Tateishi, K., Fukushima, T.: Mining product reputation on the web. In: Proceedings of the 8th ACM SIGKDD Intenational Conference on Knowledge Discovery and Data Mining (KDD 2002), pp. 341–349. ACM, New York (2002)
2. Yi, J., Niblack, W.: Sentiment mining in webfoutain. In: Proceedings of the 21st International Conference on Data Engineering (ICDE 2005), pp. 1073–1083 (2005)
3. Dave, K., Laurence, S., Pennock, D.M.: Mining the peanut gallery: Opinion extraction and semantic cllasification of product reviews. In: Proceedings of the 12th International World Wide Web Conference (WWW 2003), pp. 519–528. ACM, New York (2003)

4. Wiebe, J., Wilson, T., Bruce, R., Bell, M., Martin, M.: Learning subjective language. Computational Linguistics 30(3) (2004)
5. Gallery, M., Mckeown, K., Hirshberg, J., Shriberg, E.: Identifying agreement and disagreement in conversational speech; Use of bayesian networks to model pragmatic dependencies. In: Proceedings of 42nd Meeting of the Association for Computational Linguistics (ACL 2004), pp. 669–676 (2004)
6. Kamps, J., Marx, M., Mokken, R.J., Rijke de, M.: Using wordnet to measure semantic orientations of adjectives. In: Proceedings of the 4th International Conference on Language Resources and Evaluation, LREC 2004 (2004)
7. Turney, P.D.: Thumbs up? Thumbs down? Semantic orientation applied to unsupervised classification of reviews. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL 2002), pp. 417–424 (2002)
8. Nasugawa, T., Kanayama, H.: Acquisition of Sentiment Lexicon by Using Context Coherence. In: IPSJ SIG Notes, pp.109–116 (2004) (in Japanese)
9. Kaji, N., Kitsuregawa, M.: Dependency-based Probabilistic Model for Sentiment Classification. In: Proceedings of Data Engineering Workshop 2006 (2006)
10. Kakaku.com, `http://corporate.kakaku.com/en/`
11. Oguma, J., Utsumi, A.: Document clustering that uses co-occurrence information on word. In: Proceedings of the Annual Conference on JSAI (CD-ROM) (2007)
12. Arita, I., Kikuchi, H., Shirai, K.: Word Clustering Using Concurrent Search Queries. In: IPSJ SIG Notes, pp. 115–120 (2007)
13. Turney, P.D., Littman, M.L.: Unsupervised learning of semantic orientation from a hundred-billion-word corpus: NRC Technical report ERB-1094, Institute for Information Technology,11 pages (2002)
14. Takamura, H., Inui, T., Okumura, M.: Extracting Semantic Orientations of Words using Spin Model. In: Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL 2005), pp. 133–140 (2005)