Social Media Analysis
— Consumer Interest Detection over Time Series Concept Graphs —

HASHIMOTO, Takako

Introduction

Social media 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 media 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 media have become highly-influential in the marketing research domain.

Data mining techniques for product marketing to analyze word-of-mouth in social media have recently become an active area of research[3~9]. In analyzing product reviews or reputation by word-of-mouth in social media, almost all existing research focuses first on specific products, and extracts typical evaluation expressions such as "favorite," "dislike," "expensive," and "useful." They then calculate positive/negative degrees of extracted expressions. We have also researched data mining techniques on home electrical appliances such as air purifiers and front loading washing machines with automatic drying systems and proposed a reputation analysis framework for buzz marketing sites[10]. It may be easy to analyze a specific product's reputation, because the target product's characteristics can be illustrated by the specific ontology for the product, which is constructed with relatively little effort. On the other hand, it is very difficult to analyze unexpected consumer interest for "unspecified products." Because the target product is not explicit, it is not possible to prepare a specific ontology in advance. In this paper, we would like to discover unexpected consumer interest from word-of-mouth in social media as well as expected consumer interest.

In our data mining experiments concerning the super-flu spawn in 2009, we discovered an unexpected consumer interest. In threads about digital single-lens reflex cameras (digital SLR cameras), we discovered that many persons wrote about the flu. The flu pandemic made consumers hold off buying digital SLR cameras. In this case, we guessed that the flu made people cancel plans for children's PE festivals and trips during Golden Week in Japan because their children were confined at home. And those who had been planning to take photos at those events were reluctant to buy digital cameras. We could easily expect more air purifiers to be sold

due to the flu pandemic. However, the reluctance in buying digital single-lens reflex cameras because of the flu was not something we'd expect. The relation between the flu and digital SLR cameras can be recognized as an unforeseen and indirect relationship. We've already proposed Graph-based Consumer Behavior Analysis from Buzz Marketing Sites[1,2]. Our previos papers clarify the sort of unforeseen and indirect relationships between a current topic and unspecified products. In analysis, we adopt the concept graph due to Hirokawa[3], which makes relevance hypernym relations of keywords appearing in a set of documents. Beyond that, time series variation of graph structures is considered to detect consumer interest. This paper introduces the previous work on consumer interest detection over time series concept praphs.

The following section illustrates the unexpected consumer interest we discovered concerning the super-flu. Section 3 refers to existing research. Section 4 shows the consumer interest analysis framework on buzz marketing sites. In section 5, we use Hirokawa's concept graph to analyze the relationship between the flu pandemic and consumers holding off buying digital SLR cameras. Finally, section 6 gives concluding remarks and describes the direction of future work.

1. Expected and Unexpected Consumer Interest

This section illustrates expected and unexpected consumer interest. We focus on the super-flu pandemic in 2009. First, we conducted text mining on BBS (bulletin board system) of kakaku.com[4], which is the most popular buzz marketing site in Japan. We used it to research the effects of the flu on consumer interest. On the site, we can read word of mouth episodes about various products and various current affairs and events. One person begins a thread with one topic. Others then post their word of mouth views sequentially until the first person closes the thread. We call the posted individual document a post document.

1.1 Expected Consumer Interests

We define "expected consumer interest" as consumer interest that has explicit

(1) Hashimoto T., Kuboyama T., Shirota Y., "Graph-based Consumer Behavior Analysis from Buzz Marketing Sites," Proc. of 21st European Japanese Conference on Information Modelling and Knowledge Bases, pp.60-71 (2011).

(2) Kuboyama T., Hashimoto T., Shirota Y., "Consumer Behavior Analysis from Buzz Marketing Sites over Time Series Concept Graphs," Proc. of 15th International Conference on Knowledge-Based and Intelligent Information & Engineering Systems (2011) (to appear).

(3) Shimoji, Y., Wada, T., Hirokawa, S.: Dynamic Thesaurus Construction from English-Japanese Dictionary, The Second International Conference on Complex, Intelligent and Software Intensive Systems, pp.918-923 (2008)

(4) kakaku.com, http://kakaku.com/

relationships with current topics. Figure 1 shows temporal development of the number of post documents that include the word "flu." We conducted counting for all products on the BBS site. In Japan, the first infected patient of the super-flu was detected in May 2009 and was important news. In September to November 2009, the great epidemic was noticed and precautions were strengthened. The number of post documents with the word "flu" in Figure 1 correlates the symptoms in autumn with news reports. We then show the number of post documents with the word "flu" for four products (Digital camera, Digital SLR camera, Lens, and Air purifier). We started by conducting text mining over all products and found four numeral effects. We discovered that the number of post documents on the "digital SLR camera" and the "air purifier" were greater than others in Figure 2. Virus elimination functions－the province of air purifiers－were quite popular, because air purifier manufacturers

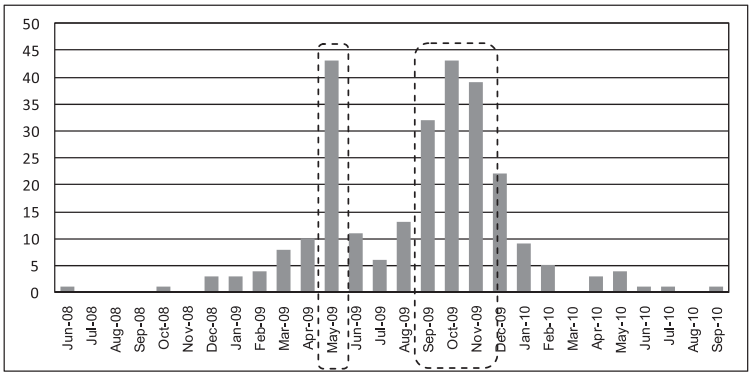

Figure 1. The number of post documents that included the word "flu" from BBS of kakaku.com (from June 2008 to September 2010).

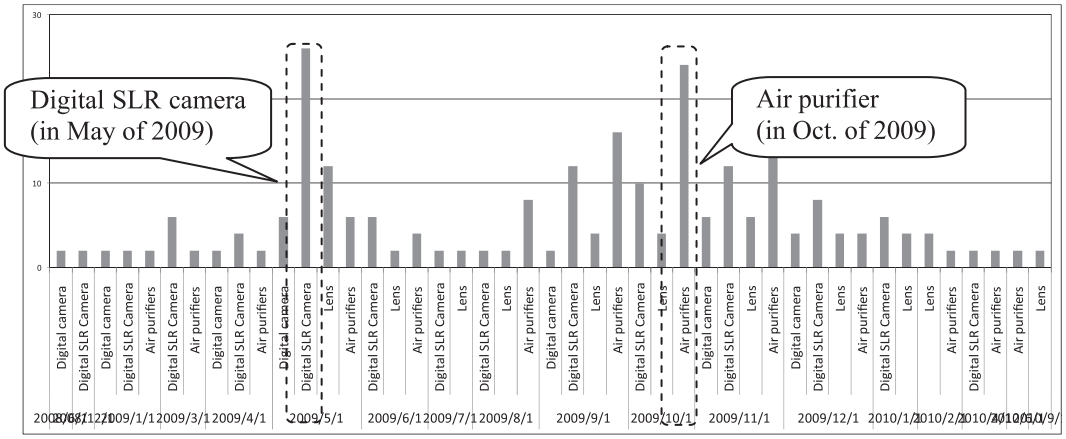

Figure 2. The number of post documents of main products (Digital camera, Digital SLR camera, Lens, and Air purifier) that included the word "flu" from BBS of kakaku.com (from June 2008 to September 2010).

propagated them through the web. Therefore, the relationship between the flu and the air purifier is intuitively expected. Figure 3 illustrates the relationship between real sales of the air purifiers and the number of post documents, where we detect a strong correlation. We recognize this kind of explicit relationships as expected consumer interest.

1.2 Unexpected Consumer Interests

We never intuitively forecasted the negative relationship between digital SLR camera sales and the flu pandemic. However, the number of post documents that include the word "flu" increased most in May 2009. Reading word-of-mouth contents, we found many messages like "owing to the flu, we cannot take a vacation (or a business trip)" and "children's PE festivals may be called off owing to the flu" repeatedly. From these comments, we can guess that various events had been cancelled due to the flu pandemic, that consumers could not go out to take photos, and that, consequently, they had held off buying cameras.

We provide supportive evidence of our guess in Figure 4, where 2009 sales of digital cameras are shown. In 2008, an ordinary year, the slack sales of 2009 does not exist. In April to May and September to November 2009, however, we found a drastic drop in digital camera sales in our domestic market as shown in Figure 4. In September to November 2009, there is a significant reverse increase-decrease pattern compared to 2008. The drop in sales is related to the number of post documents with the word "flu," we consequently guess there is a strong correlation between the flu and digital camera sales. The relationship between the flu and digital camera sales cannot be detected without mining technology, because it was unforeseen. We define this kind of an unforeseen and indirect relationship triggered by a current topic as unexpected consumer interest. This would be a new approach to marketing analysis.

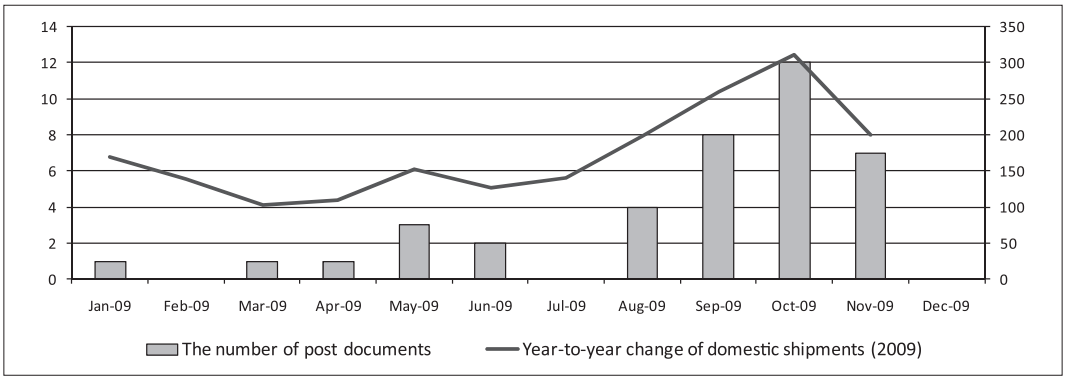

Figure 3. The number of post documents in the BBS of kakaku.com and the volume of shipments for air purifiers. (Cited: GfK Marketing Services Japan Ltd., http://www.gfkjpn.co.jp/).

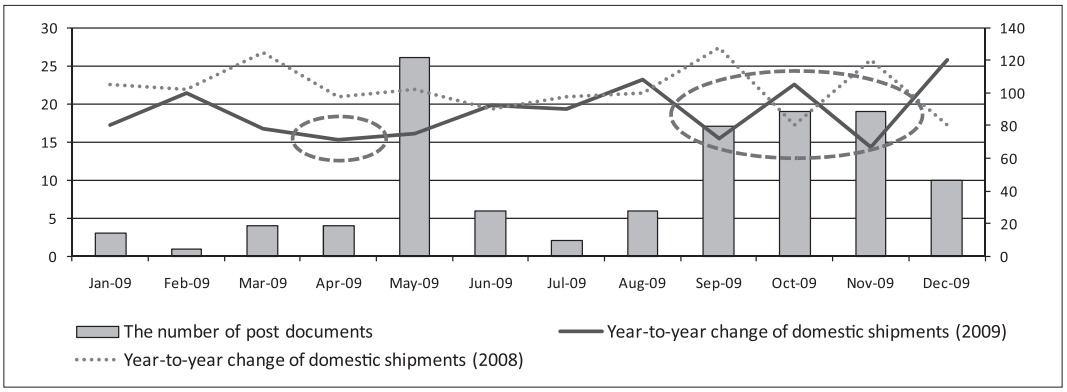

Figure 4. The number of post documents in the BBS of kakaku.com and the volume of shipments for digital single-lens reflex camera in 2009. (Cited: The Camera Information Center: Camera information Center Report, http://www.camera-info.net/index.htm).

3. Related Work

3.1 Research on Reputation Analysis

Various researchers have analyzed product reviews and reputation from social media[5,6,7,8]. Nagano et. al[5] propose the word-of-mouth engine to present product reputation on the Web. In their system, users first specify the products by taking pictures using cell-phone cameras. The system then retrieves word-of-mouth information and extracts typical evaluation words like "favorite," "dislike," "expensive," and "useful" about the specific product. It also calculates positive/negative degrees. Kobayashi et. al[6] define the main portions of an opinion as (object, attribute, opinion). Asano et. al[7] also define the basic element of reputation as (object, evaluation point, expression). To extract reputation from word-of-mouth information, both propose a technique for efficiently building an object name dictionary, an attribute expression dictionary (ontology), and an opinion word dictionary for the specific object domain. Spangler et. al[8] propose an automated way to monitor social media to analyze the specific corporate brand, reputation, consumer preferences and buying habits. They also offer a mechanism for developing the ontology, near-real-time gathering of word-

(5) Nagano, S., Inaba, M., Mizoguchi, Y., Iida, T., Kawamura, T.: Ontology-Based Topic Extraction Service from Weblogs. IEEE International Conference on Semantic Computing, pp.468-475 (2008)

(6) Kobayashi, N., Inui, K., Matusmoto, Y., Tateishi, K., Fukushima,, S.; Collecting evaluative expressions by a text mining technique, IPSJ SIG NOTE, Vol.154, No.12, pp. 77-84 (2003) .

(7) Asano, H., Hirano, T., Kobayasi, N., Matsuno, Y.: Subjective Information Indexing Technology Analyzing Word-of-mouth Content on the Web, NTT Technical Review, Vol.6 No.9 Sep. 2008, pp.1-7 (2008)

(8) Spangler, W.S., Chen, Y., Proctor, L., Lelescu, A., Behal, A., He, B., Griffin, T.D., Liu, A., Wade, B., Davis T.: COBRA - mining web for COrporate Brand and Reputation Analysis. Web Intelligence and Agent Systems (WIAS) 7(3), pp.243-254 (2009)

of-mouth information and the calculation of positive/negative measures. This related work targets specific products, extracts evaluation expressions from word-of-mouth in social media and calculates sentiment orientations of extracted expressions to analyze product reviews and reputation. They require specific ontology. Our proposed method, however, does not target specific products, and a specific ontology is not needed. We focus on a current topic and visualize the unforeseen relations between a current topic and unspecified products from buzz marketing sites. Through the visualization, we can detect unexpected consumer interest.

3.2 Research on Visualization of Reputation Analysis Result

Regarding research on visualization of relations in the results of reputation analysis, Sekiguchi et. al[9] treat recent blogger posts and analyze word co-occurrence and the rate words are repeated. They visualize the relation between words and show topics in social media through the visualization results. Wang et. al[10] propose a graphic reputation analysis system for Japanese. It presents information on the relation between the products and users' evaluations using simple graphs (pie charts and line graphs). Both also require that specific products be targeted. Our proposed method has a similar approach, which focuses on word co-occurrence and the visualization of relations. However, we visualize the relation between a current topic and unspecified products. Further, our method can show unexpected consumer interest by considering the time series variation of graph structures. We can say that our method analyzes implicit reputations. That is the novelty of our method.

4. Framework for Consumer Interest Analysis on Buzz Marketing Sites

This section discusses a consumer interest analysis framework to detect expected and unexpected consumer interest from messages on buzz marketing sites (Figure 5).

Our framework consists of the following three steps:

1. Concept graph generation
2. Consumer interest detection
3. Visualization

The followings describe each step.

---

(9) Sekiguchi, Y., Kawashima H., Uchiyama, T.: Discovery of Related Topics Using Serieses of Blogsites' Entries, The 22nd Annual Conference of the Japanese Society for Artificial Intelligence, 211-1 (2008)

(10) Wang, G., Araki, K.: A Graphic Reputation Analysis System for Mining Japanese Weblog Based on both Unstructured and Structured Information, AINA Workshops 2008, pp.1240-1245 (2008)

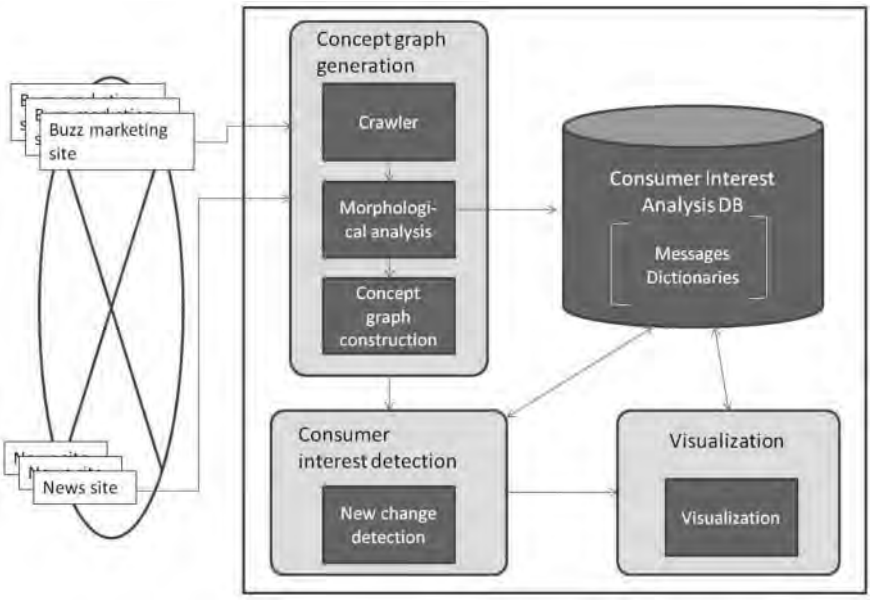

Figure 5. Consumer interest analysis framework.

- Concept Graph Generation
  In this step, input data is the set of messages on buzz marketing sites. We define one message as one document and crawl messages. We then extract important words using morphological analysis and construct concept graph. In the next chapter we discuss this step in detail.

- Emerging needs detection
  To detect emerging needs, we plan to find changes in graph structures. In our framework, the messages are acquired periodically (e.g., once a week), and then the concept graphs are formed. If a new structure is generated, we recognize that new changes (or new needs) have been detected.

- Visualization
  Based on above steps, the results are visualized to users.

5. Consumer Interest Analysis by Concept Graphs

This section explains our analysis method in finding the unexpected consumer interest concerning the flu pandemic in 2008.

5.1 Concept Graph

We use a the concept graph proposed by Hirokawa et al.[3] to show the unforeseen

relations between the current topic and unspecified products. Hirokawa et al, proposed a simple method to construct a hierarchy of words from a set of documents automatically and dynamically. The method first retrieves the set of documents according to given keywords, and extracts related words. Then hypernym relations of these related words are obtained using co-occurrence frequencies. A concept graph is a directed acyclic graph whose nodes are characteristic words of the set of documents and whose edges represent the upper and lower relation of words. It can present meaningful structures. For example, suppose a set of documents is retrieved by the query "wine" from an English-Japanese dictionary. The concept graph of the trivial hyponym "white wine," "chardonnay," "cuve," and so on which indicate the names of areas and brands of wine, is then constructed.

Hirokawa et. al formalize upper-lower relationships among words in documents as a concept graph. The set of whole target documents is represented as "U." Given a subset of "U" as "X," and keywords "u" and "v," "$df(u, X)$" represents the number of documents in X that contain the keyword "u," and "$df(u^*v, X)$" represents the number of documents that contain both "u" and "v" in X. The relevance between "v" and "u" is defined as follows:

$$r(v, u) = \frac{df(u^*v, X)}{df(v, X)} \tag{1}$$

If $r(v, u)>0.5$ and $df(u, X)>df(v, X)$ then "u" is defined to be greater than "v" from the standpoint of document frequency. The hypernym/hyponym relation then determines an order structure among characteristic words and can be drawn as a directed acyclic graph. Visualization of a concept graph can be obtained by placing words of high frequency on the left and ones with lower frequency on the right. Thus a directed edge looks like an arrow from left to right.

Iino et. al[8] showed that the concept graph can effectively depict the time series variation of organizational structure by researchers, which is a form of implicit information in patent documents. The method is efficient and objective for discovering the implicit relation among a set of documents.

5.2 Concept Graph Generation from Buzz Marketing Sites

We decided to use the BBS (bulletin board system) of "kakaku.dom[2]" because the site is the most popular buzz marketing site in Japan. We would like to find post documents among many products related to a specific topic of current affairs, such as the flu. We therefore think the kakaku.com BBS is suitable for us to conduct a crossover retrieve on many products.

The flow for constructing a concept graph of a specific current topic follows:

1. Select the most impressive topic word of current topic "w."

2. Search the kakaku.com Web site to find a set of post documents that includes the topic word "w." The result is defined as a set of post documents D=$\{d_i\}$ where d_i represents one post document that includes the "w." Each d_i has four attributes ($posted\_time$, $target\_id$, $user\_id$, $content_i$). The $content_i$ represents content of the post document and points to the i-th content among a whole set of post document contents. The target_id shows a target product or topic of the post document such as a digital SLR camera, a camera lens, and an air purifier. The user_id specifies the person who posted it.

3. We extract keywords that are nouns, verbs, adjectives, and adverbs from d_i using morphological analysis and then calculate the value of tf-idf (Term Frequency-Inverse Document Frequency) for an individual keyword.

4. For each d_i, we select a set of keywords of which tf-idf values are greater than a certain threshold level "T." The posted date is then delimited monthly at the beginning of the month, and D is clustered by month.

5. For each cluster, we construct Hirokawa's concept graph.

In this paper, we set w="flu" and T=2.0, and retrieve post documents from January 2009 to December 2009 in kakaku.com. Table 1-1 lists the excerpts in the result of step 4. In May 2009, there were keywords with high tf-idf value (>2.0) such as "Lens," "Bike," "Papa" and "Mama" on digital SLR cameras. These keywords are not typical opinions for the flu, and digital SLR cameras seem to be irrelevant to the flu, because it is difficult to extract these keywords using existing research approaches that target specific products in advance. In October of 2009, however, there are keywords like "Ion," "Purify," "Plasma Cluster" and "Antibody" regarding the function of air purifiers (Table 1-2). These keywords are typical opinions against flu and we can say there is a strong relation between air purifiers and the flu. It is easy to imagine that they are opinions about the flu. That means we can say that it is easy to extract these keywords using existing research approaches.

Table 1-1. the excerpt in the result of keyword extraction (with high tf-idf calues) of step 4 (in May of 2009).

ID	Product Name	Keywords
9599867	Digital SLR Camera	flu, Lens
9560041	Digital SLR Camera	flu, XIAN, Bike, PiPiPi
9586600	Digital SLR Camera	flu, Papa, Mama, Lens, Shoot
9587481	Digital SLR Camera	flu, Virus
9563058	Digital Camera	flu, Oversees Travel
9566776	Air Purifier	flu, Ion, Virus, Product
9628120	Air Purifier	flu, Virus, Daikin, Plasma Cluster

Table 1-2. the excerpt in the result of keyword extraction (with high tf-idf values) of step 4 (in October of 2009).

ID	Product Name	Keywords
10262671	Digital SLR Camera	flu, Shoot
10335307	Digital SLR Camera	flu, Shoot, Break
10288631	Air Purifier	flu, Virus, Air, Experiment, Purify
10317588	Air Purifier	flu, Nanoe, Ion, Streamer, Daikin, Discharge
10311912	Air Purifier	flu, Ion, Virus, Streamer, Plasma Cluster, Air, Antibody
10342041	Air Purifier	Influ, Air, Product, Purify
10342158	Air Purifier	Influ, Streamer, Purify, Experiment

According to our preliminary survey, however, with regard to digital SLR cameras, because of the flu pandemic, people who had plans for children's PE festivals and trips during Golden Week in Japan were confined to their homes and those who had been planning to take photos of the events were reluctant to buy digital cameras. From this viewpoint, keywords such like "PE festival," "trip," "Golden Week," and "cancel" should have been extracted from post documents. In general, we used tf-idf values to extract keywords. However, when we use tf-idf values, values of keywords extracted from regular post documents tend to become high. On the other hand, values of keywords we would like to extract tend to become low. We have to improve measures to extract useful keywords that more precisely express unexpected user interest.

Figures 5 and 6 show the concept graph as of May 2009 and October 2009. Concept graphs are directed acyclic graphs whose nodes are characteristic words (keywords) of the set of post documents in kakaku.com BBS and whose edges represent not only upper and lower relation of words but also the name of the product having a correlation between both nodes. There is a large island structure

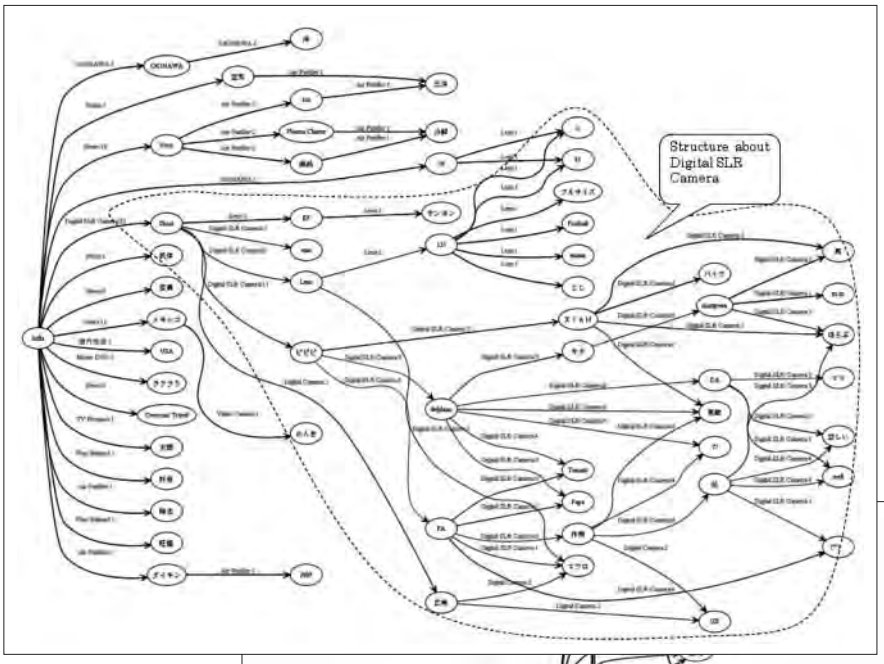

Figure 6.
Concept graph as of May 2009.

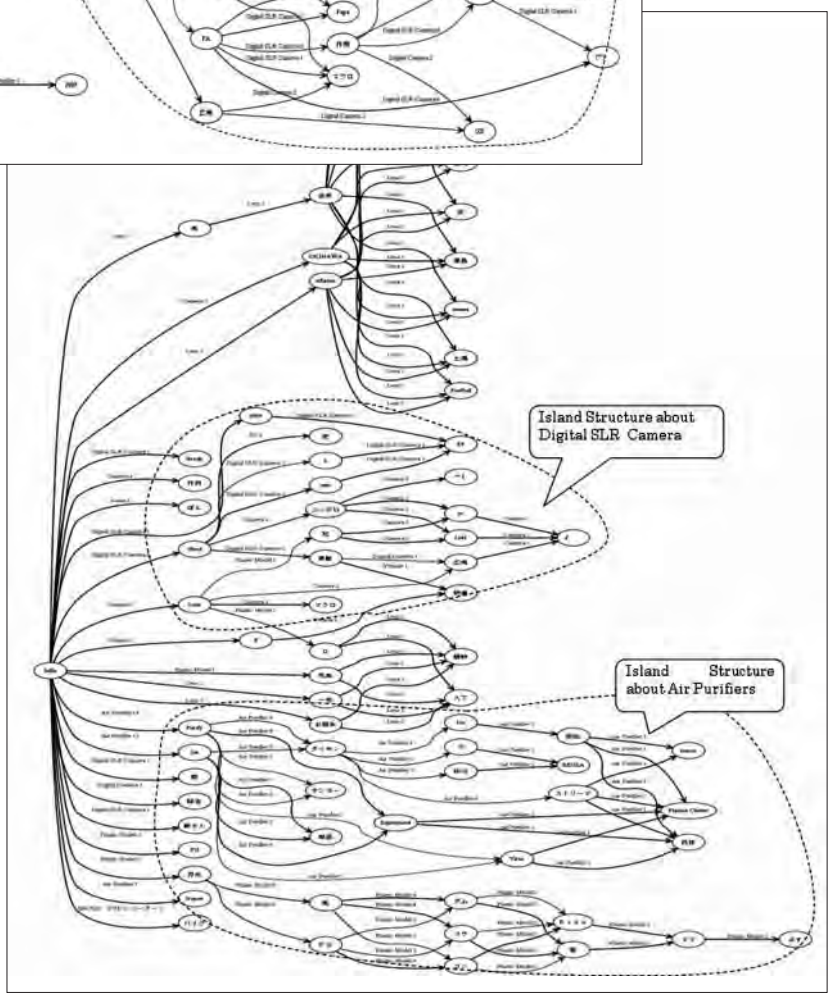

Figure 7. Concept graph as of October 2009.

about digital SLR cameras in Figure 6, because it shows that users who are interested in digital SLR cameras take a strong interest in the flu as well. A relationship between the flu and digital SLR cameras is indicated. On the other hand, in Figure 7, there is a large structure about air purifiers as well as the structure about digital SLR cameras. It also indicates that users who are interested in air purifiers take a strong interest in the flu. The relationship between the flu and air purifiers is indicated.

6. User Interest Detection Based on Time Series Variation

A concept graph is helpful to discover unexpected consumer interest. By analyzing time series variation of concept graphs, user interest can be detected more precisely. Figure 8 shows monthly concept graphs about the flu extracted from messages of kakaku.com (from January 2009 to December 2009). Regarding this time series variation, we first recognized that major structure changes happened in January, May, and September. Our hypothesis was that these major changes caused specific consumer interest. In Japan, the first infected patient of the super-flu was detected in May 2009 and was reported in the mass media. Further, from September to November, the great epidemic was noticed and precautions were strengthened. These major structure changes happened according to the topical problem "flu."

6.1 Expected Consumer Interest Detection

As mentioned, we define expected consumer interest as consumer interest having explicit relationships with current topics. This subsection illustrates how expected consumer interest is shown in our concept graphs.

The air purifier island structures are indicated with gray rectangles in the Figure 8 concept graphs. We see there are newly-created structures about air purifiers in the concept graphs of May and August. With this structure analysis, we see that the drastic structure changes happened in May and August Compared to real sales of air purifiers, the amount changes started in May and August. We can therefore guess that the structure change in May and August illustrates consumer interest.

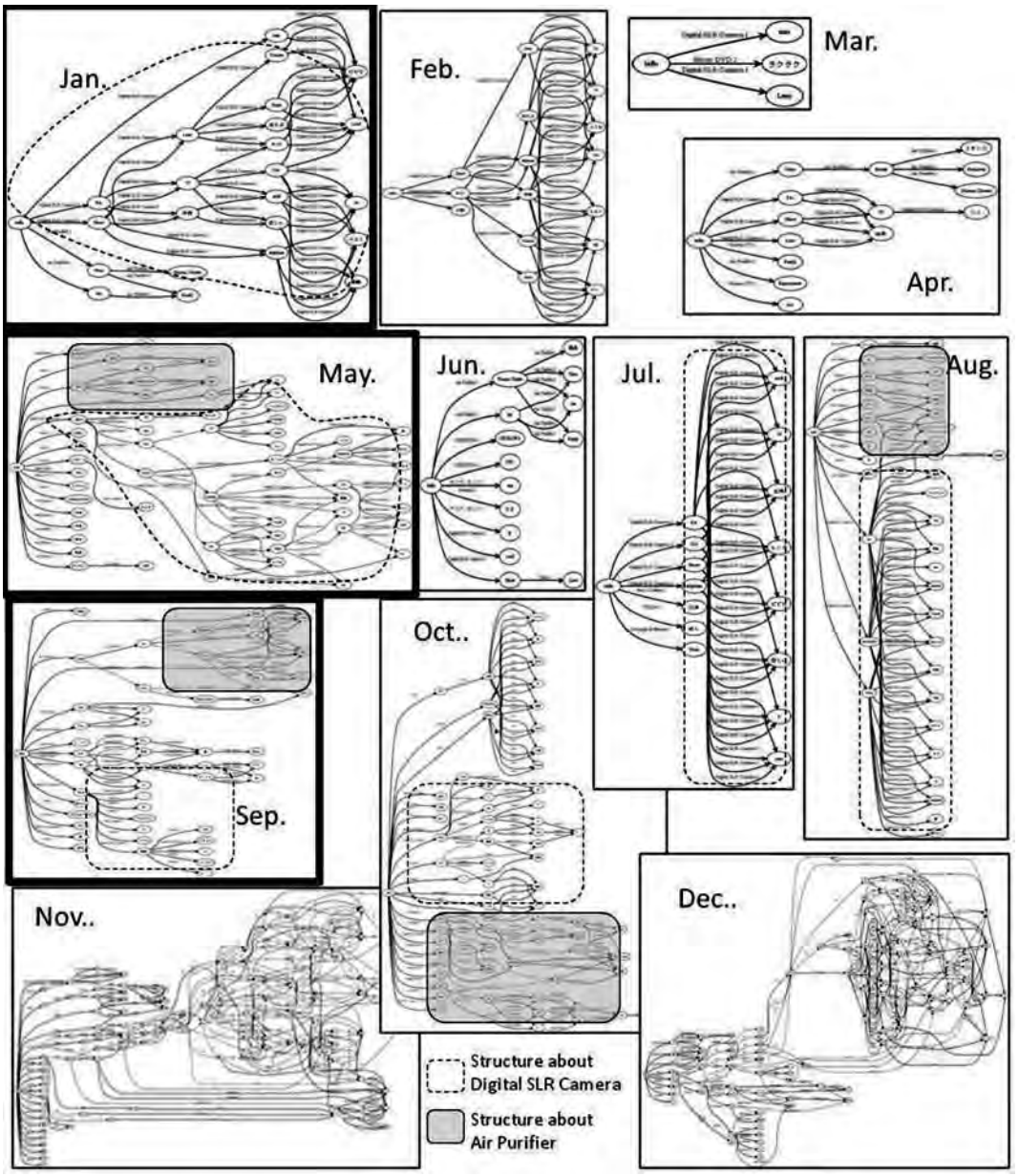

Figure 8. Concept graphs about the flu from kakaku.com (2009)

6.2 Unexpected Consumer Interest Detection

We define unexpected consumer interest as consumer interest that has an unforeseen and indirect relationship triggered by a current topic. For example, the relationship between the flu and digital SLR cameras is an unforeseen relationship.

The digital SLR camera's structures are indicated with clear rectangles in the Figure 8 concept graphs. Graph structures about digital SLR cameras are also recognized in the concept graphs of January, May, July, August, September and October of 2009. With this structure analysis, we recognize that the major structure changes happened in May, July, and September. We guess that these structure changes cause unexpected consumer interest. Compared to real sales of digital SLR cameras (Figure 3), sales increased in June (after May) and October (after September). We can guess, therefore, that the structure change in May and September illustrates consumer interest. We can say that these structure changes can express changes in consumer interest.

This sort of unexpected consumer interest detection based on graph structure changes cannot be conducted using existing reputation analysis research. It is not possible because existing research focuses on specific products first, and then extracts typical evaluation expressions such as "favorite," "dislike," "expensive," and "useful." Time series analysis based on concept graph structures is useful to visualize unexpected user interest.

7. Conclusion and Future Work

This paper visualizes expected and unexpected consumer interest from messages on buzz marketing sites using Hirokawa's concept graph.

To detect unexpected consumer interest more precisely, we must improve the following:

1) Consider a new scoring measure to replace tf-idf
   As mentioned in section 5.2, when we use tf-idf values, values of keywords we would like to extract tend to become low. Because values of keywords extracted from regular post documents tend to become high, we must improve the measure to extract useful keywords to more precisely express unexpected user interest.

2) Introduction of "graph edit distance measure"
   In this paper, we have visually evaluated the graph structure changes. For more precise evaluation, we will introduce the "graph edit distance measure[13]."

3) Integrate with marketing data
To develop a system that can extract unexpected consumer interest semi-automatically, the results of concept graph analysis should be integrated with appropriate marketing data.

Beyond the improvements listed above, we will acquire other data examples that can express unexpected consumer interest from buzz marketing sites, and evaluate the effectiveness of our proposed method using the concept graph.

謝　辞
本研究は，平成22年度千葉商科大学学術研究助成金の交付を得て実施した研究の成果である。

(11) Iino, Y., Hirokawa, S.: Time Series Analysis of R&D Team Using Patent Information, Lecture Notes in Computer Science, 2009, Volume 5712/2009, pp.464-471 (2009)
(12) Hashimoto, T., Shirota Y.: Semantics Extraction from Social Computing: A Framework of Reputation Analysis on Buzz Marketing Sites, Lecture Notes in Computer Science, 2010, Volume 5999/2010, pp.244-255 (2010)
(13) Bunke, H.: On a relation between graph edit distance and maximum common subgraph, Pattern Recognition Letters, Volume 18, Issue 8, August 1997, pp.689-694 (1997)

— Abstract —

Social media, 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 media have become important in recognizing the reputation of products hold with users. This paper proposes a method to discover consumer interests from buzz marketing sites. For example, in 2009, the super-flu virus spawned significant effects on various product marketing domains around the globe. Using text mining technology, we found an unforeseen relationship between the flu pandemic and the reluctance of consumers to buy digital single-lens reflex camera. In this paper, the unforeseen relationship between a current topic and products is modeled and visualized using a directed graph that shows implicit knowledge. Consumer interest is further analyzed based on the time series variation of directed graph structures.

Keywords. Data mining, marketing research, Web Intelligence, visual analysis