Mobile Service for Reputation Extraction from Weblogs – Public Experiment and Evaluation

Takahiro Kawamura, Shinichi Nagano, Masumi Inaba, and **Yumiko Mizoguchi**
Corporate Research & Development Center,
Toshiba Corp.

Abstract

In this paper, we introduce a mobile service that extracts reputations of a product from weblogs by cellular phones during shopping. If the user takes a photo of a product barcode on the package with a cellular phone camera, Ubiquitous Metadata Scouter first gets the product metadata (name, manufacturer, etc.) from the internet and collects blogs that review the product. Also, it analyzes the blog contents with NLP techniques and ontologies. Then, it indicates the overall reputation (positive or negative), and other related products that are the subject of much discussion in the blogs. This paper illustrates each function of this service and a public experiment and evaluation at a real consumer electronics store and book-store in Tokyo in March 2006.

Introduction

In a ubiquitous computing environment, it would be important for users to bind their situation and the related useful information on the internet. A typical example is the showing of movie information if the user is heading to a movie theater, or suggesting sightseeing points if the user is on a trip. In conventional desktop computing, an information system such as a web browser can show lots of results at once, because most users have big displays, familiar interfaces, and a broadband connection, and so it is easy to check and view the results repeatedly. For example, most personal computer users are already accustomed to checking 100 results output by Google, and they click and click again until they finally get the desired information. However, in contrast to PCs, typical devices for ubiquitous (currently, just mobile) computing such as cellular phones have small displays, limited operations, and a narrowband connection. Therefore, to easily get the useful information via the internet by those devices, the operations must be more automatic, and only the really necessary information should appear on the displays. In fact, a survey of cellular phone carriers shows that the number of people using the internet via cellular phones has hit a ceiling in recent years in Japan.

Thus, we propose using semantic web technology, that is, metadata and ontology, to help extract only the necessary information for the users in ubiquitous computing environments. In this paper, we introduce Ubiquitous Metadata Scouter(UMS), a practical example of the application of metadata and ontologies to a natural language processing technique .

What is Ubiquitous Metadata Scouter?

Recently, blogs have multiplied exponentially, and lots of bloggers quickly check press releases concerning new products, and then publish reviews on the products from their own view points on their blogs. On the other hand, users who are considering buying a product tend to refer to these blogs when making purchase decisions. Further, those users may also become bloggers and publish their own reviews. Consequently, WOM (Word-Of-Mouth) information is growing rapidly via the internet.

Ubiquitous Metadata Scouter is a mobile service to get WOM information for a product of interest in a ubiquitous computing environment. If the user scans a product barcode by using a cellular phone's camera (more than 90% of new phones in Japan are equipped with cameras), UMS gets the product metadata corresponding to the barcode number from the internet and collects the blogs. In addition to showing the snippets extracted as positive or negative opinions from the original blog bodies according to the order of the amount of opinions, it shows a summary such as $POSITIVE\ xx\%\ vs.\ NEGATIVE\ yy\%$ and the associated products that have co-occurred with the scanned product. For example, if the user snaps a barcode of a book, UMS finds metadata including the book title, publisher, author from UPC/EAN/JAN or ISBN, and collects blogs mentioning the book. Then, it analyzes contents of blogs by referring to blogs' metadata and ontologies, and indicates $POSITIVE\ 73\%\ vs.\ NEGATIVE\ 27\%$ (P/N determination). Also, it shows other books such as those by the same author or in the same genre (Associate Topic extraction). Then, it lists some blogs with snippets that seem to be opinion sentences (Sorting and Filtering). Further, Fig. 1 shows an example output for a product. In the following sections, we briefly introduce the above three functions.

P/N determination

Several P/N determination techniques have already been proposed in the literature on natural language summarization. One of the proposed techniques involves retrieving triples $< subject, property, value >$ such as $<$

Figure 1: Example result

$car, speed, fast >$ for a target subject by checking modification relation through morphological analysis and syntactic parsing(Kobayashi *et al.* 2005)(Dave, Lawrence, & Pennock 2003)(Turney 2002). Further, there are several applied techniques such as extension to a variety of documents and styles of output. For example, target documents have been changed to rating comments on an auction site(Kusumura, Hijikata, & Nishida 2004), and the results have been represented by a radar chart in which the properties are axes. In our research, we added the two improvements described below to the above-mentioned technique in order to apply UMS specifically to blogs.

Use of blog correlation from metadata The first improvement is an exploitation of blog metadata. We are considering the correlations among blog entries, which are determined by blog metadata, RSS (RDF Site Summary) 1.0(Brickley 2000), in order to put accord weight depending on the importance of each blog entry (article in a blog) and bias their opinions. In most previous research, the target documents are a set of (unlinked) documents such as newspapers prepared as an experimental corpus. Then, the fact that the users will encounter several opinions by tracing the links of the web pages is outside the scope. However, if there are two blogs, one of which has lots of favorable comments and trackbacks (a function to notify a blog author that "I have put a link to your blog entry"), and one with no comments or trackbacks, then the user who reads those two blogs will have different levels of trust in them. Further, the user will have different levels of trust in an opinion of an author submitting many reviews on a product category and that of a chance author.

Some web mining research focuses on trackbacks(Kimura *et al.* 2005), but no research has been reported that actively makes use of implicit but useful blog correlations retrieved by RSS, such as frequent combinations of blog authors and their interests, flows of pros and cons along with trackback links, and so forth. Blogs are not diaries or advertising space, but constitute communities loosely coupled by trackbacks. Therefore, instead of limiting the document type to blogs, our P/N determination extracts and utilizes the blog correlations determined by their RSSs. Specifically, we are using the following weighting heuristics. Fig. 2 illustrates the correlations among blogs. We believe that those heuristics make the p/n result closer to the reader's actual impression.

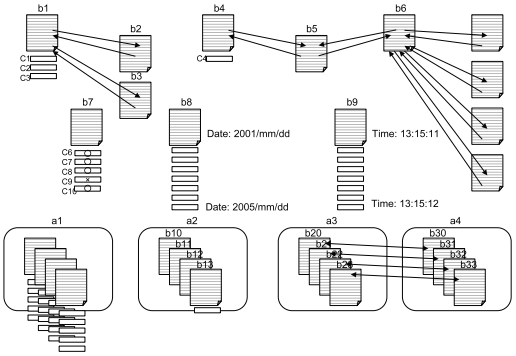

Figure 2: Blog correlation measured by RSS metadata

1. accord weight to opinions expressed by trackbacks rather than in comments. (Non-anonymity, b2 and b3)
2. accord weight depending on the amount of favorable trackbacks and comments from different authors. (Wide acceptance, b6)
3. accord weight to opinions of an author reviewing related products. (Expert, able to make comparisons from various perspectives)
4. accord weight to an agreement among the flow of disagreements in responses, that is, trackbacks and comments, and vice versa. (Courageous attitude, c9 in b7)
5. accord weight to opinions that are getting responses for a long period although their time stamps are old. (Pioneer, b8)
6. accord weight opinions that have high value as indicated by dividing the time from the first response to the last one by the number of responses. (High acceleration, b9)
7. accord weight to opinions of an author whose average number of responses is high. (Opinion leader, a1)
8. decrease the weight accorded to blogs that have no response. (No ads, b10-12)
9. decrease the weight accorded to blogs of authors who have lots of blogs with no response (No agency, a2)
10. accord weight to frequently exchanged opinions between a few authors. (Debate, a3 and a4)
11. Finally, according to a survey report(Web Advertising Bureau 2006), 70% of authors tend to make more positive comments than negative ones. So, in the case of negative comments such as complaints, we determine that the intention of the author is higher than in the case of good comments, and then accord greater weight to the blogger's comments.

Use of ontologies The second improvement is the use of product ontology and evaluation expression ontology. Since the ontologies are written in Japanese, overviews of the ontologies are shown in Fig. 3.

The product onology is a graph of products in a domain such as Book or DVD. It is not just a product category,

but it has properties belonging to a product class and instances that have values for the properties. For example, the book class has properties such as author and publisher, and a book instance has the actual author's name and the publisher's name. This dependency of the product (the instance) and the properties (its values) greatly contributes to improvement of the accuracy and speed of the above syntactic parsing as the following reason. The syntactic parsing is a process to trace the modification relation from a target word in a sentence. Then, we basically search evaluation expressions such as "good" or "expensive" in the relation. One of the target words is of course the product name (subject). However, unlike sentences in newspapers, many of the casual sentences in blogs lack subjects. Therefore, if we know the properties or the values associated with the product name that might appear in the sentences, we can take them as the target words to trace the modification relation, and carefully get the evaluation expressions about the product. This combination of NLP and ontology raises the accuracy of our P/N determination and reduces the time required for round-robin matching to remaining expressions.

The evaluation expression ontology defines the relationship between attributes such as Functionality, Design, Speed, and, their values like "good", "beautiful", "high". Therefore, we can adopt criteria other than just positive vs. negative in future, although we believe this simple summarization of p/n is the best for mobile users. Also, it defines strength and weakness of the evaluation expressions such as Best, Good to Worst by the "degree-of" relation. Thus, our P/N determination is not just a sum of positive (+) or negative (-) points, but takes the degree of each expression into consideration. Of course, it can reverse positive or negative meaning according to the attribute even if it has the same value, such as $< car, speed, high >$ and $< car, cost, high >$. Further, if the expression does not directly mean positive or negative, it might be used by referring the "part-of" relation to other evaluation expressions. For example, some adverbs are commonly used with certain positive or negative expressions. Needless to say, several other expressions with the same meaning are merged to a concept by referring to the "instance-of" relation. This is a thesaurus use of the ontology.

Associate Topic extraction

Associate Topic (AT) extraction is a function to find other products that are similar to a certain product. However, since nowadays a many blogs are like spams or ads, a simple method based on word frequency would fail to identify hot products. So, we make use of the blog correlation as well as the P/N determination. Bloggers often communicate their own opinions to others' blogs by means of trackbacks. This correlation among blog entries is called a blog thread, where a certain topic is intensively discussed among bloggers. Other readers would have more trust in the opinions in a blog thread than in an opinion in a single blog entry without any trackback or comment. Therefore, the approach based on the blog thread seems effective for finding associated topics.

The AT extraction is based on TF/IDF, but we are not counting product name frequency, but the frequency of authors who mentioned the product. Also, we defined a utility function in terms of the degree of association and the number of compared products within a blog entry, where the degree becomes the maximum when the number of products is 2 to 4. Then, UMS analyzes the blog threads, and if a product is mentioned in a blog entry that has many trackbacks and comments, the product is given a high degree of association. Additionally, if the blog entry that has a trackback to the original blog mentioning the product also has many trackbacks, the degree is increased proportionally. This is a page rank algorithm for the blog search.

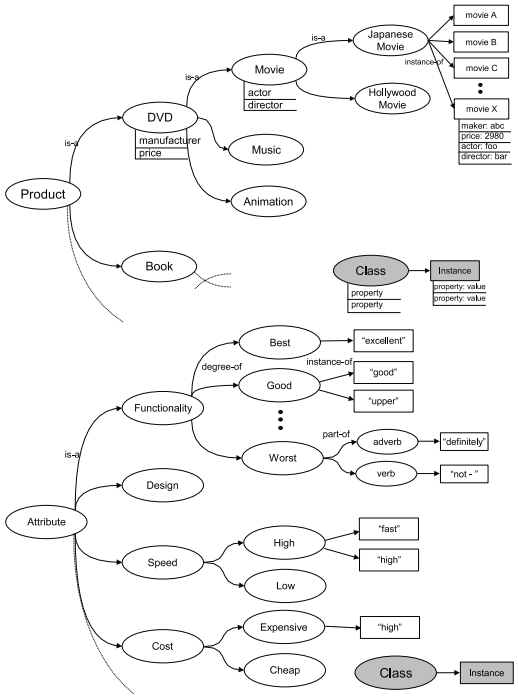

Figure 3: Product Ontology(above) and Evaluation Expression Ontology(below)

Sorting and Filtering

Finally, the user will also want to read the actual blog entries. So we show a maximum of 20k of blog entries that are selected as worth reading after spam elimination. Here, we also use ontology and metadata. When searching for blog entries related to a product, we add attributes of the product and/or concepts near the product in the product ontology as additional search keys in order to restrict the product domain and eliminate products with the same name in a different domain. Also, the blog correlations are used on top of spam filter techniques for e-mails. For example, blogs without any trackbacks or comments are removed. Further, the result of the P/N determination is used, and then blogs that have obvious positive or strong negative opinions are prioritized, etc.

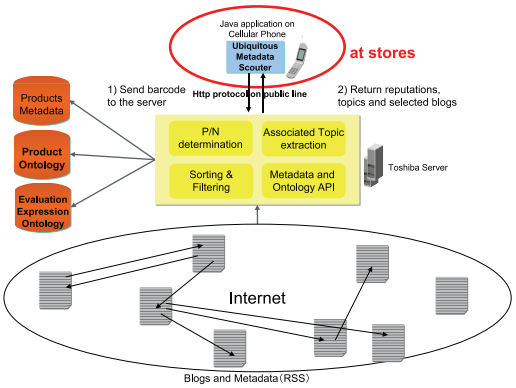

Figure 4: System architecture

Table 1: P/N impression, analysis and result

Target Product	Users Impression	
	Positive (%)	Negative (%)
DVD	86.8	13.2
Book	95.0	5.0
HDD Recorder	81.8	18.2
	Experts Analysis	
	Positive (%)	Negative (%)
DVD	62.5	37.5
Book	71.4	28.6
HDD Recorder	28.6	71.4
	Output of P/N determination	
	Positive (%)	Negative (%)
DVD	86.4	13.6
Book	91.7	8.3
HDD Recorder	100.0	0.0

Public Experiment and Evaluation

Recently, we finished the development of the evaluation version of UMS. Then, we conducted a public experiment of this service at a consumer electronics store and a book-store in Tokyo in March 2006. This section shows the evaluation result for the accuracy of reputation extraction and its speed, such as turnaround time, clarified by the experiment.

In the experiment, test users receive cellular phones with our client application installed. They scan barcodes on products of interest to them at the stores and check the displayed results. Then, they complete a questionnaire. It should be noted that we collected profiles of the test user candidates in advance and included some groups, each consisting of a few people organized by age, sex and internet expertise. There were 19 users (men 10, women 9). The system architecture is shown in Fig. 4.

P/N determination

In the following, we show the evaluation of the accuracy of the P/N determination. First, Table 1 shows the users' positive and negative impression ratios (avg.) for target products, which were gathered after the users had read several (correlated) blog threads concerning the products. In this evaluation, the number of target products is three. It is a small number, but it takes almost 30 minutes to read blog threads for a product. We selected a product for each of three genres: DVD, book, and digital product [1]. These genres were selected because cheaper products are readily available in stores, although those products' reputations may be less impressive, and more expensive ones would be considered in advance.

On the other hand, Experts Analysis in Table 1 shows the ratio of supporters and opponents (not supporters) counted by NLP researchers, after carefully reading all blog entries in the same blog threads without checking the correlation. By comparing these tables, we can confirm that there is a gap between users' sensitivity to blog reading (Users Impression) and the real count (Experts Analysis). This supports our assumption mentioned in section 2.1 that the P/N determination for the users is not a simple sum, but would be affected by the correlations among blog entries and the degree of expressions.

Before showing the result of the P/N determination, we illustrate its procedure in this experiment. Firstly, a blog crawler collects a maximum of 100 blog entries containing (part of) the product name. Then, we take some sentences around a sentence that includes the product name (for considerations of processing time), and extract the properties and the value pairs found in the modification relations from the target words obtained from the product ontology. After according some weights to each blog entry based on the blog correlation, we apply the evaluation expression ontology to the property-value pairs and calculate the degree of positive and negative for the product. The number of nodes in the product ontology is approx. 1,500,000, and the number of nodes in the evaluation expressions ontology is approx. 10,000.

According to this procedure, first, we get the result of our P/N determination without two improvements: metadata and ontology. We compared this result with Experts Analysis in Table 1, and then obtained the accuracy of our base mechanism of the P/N determination. In Table 2, we can see the result matches more than 80% to the experts' analysis, indicating a certain degree of accuracy, except for the HDD recorder. The reason for the low accuracy for HDD recorder is that the collected blogs contained many ads placed by retailers. The experts in Table 1 did not count them as commercials, but the P/N determination often counted them as positives due to the lack of any clue to determine they were ads. This indicates a need to improve the accuracy of Sorting and Filtering function soon.

Then, we got the result of the P/N determination with the two improvements. It is shown in Output of P/N determination in Table 1. We compared this result with Users Impression in Table 1, and then found that our P/N determination can get the degree of positive and negative close to the users' impression, at least in this experiment. Thus, we believe our

[1] the DVD is "Charle and The Chocolate Factory", the book is "Tokyo Tower" in Japanese, and the digital product is an HDD recorder "TOSHIBA RD-X6".

Table 2: Accuracy of P/N determination to experts analysis (without improvements)

Target Product	Accuracy
DVD	87.5
Book	85.7
HDD Recorder	57.1

Table 3: User assessment

Category	Points
DVDs	3.6
Books	3.5
Electronics	2.8
Average	3.3

proposed improvements have an effect to a certain extent, except again for the case of the HDD Recorder due to the above-mentioned reason.

Furthermore, Table 3 shows the answers to a question asking the test users if the outputs of the P/N determination fit their impression for all the products scanned in the experiment (avg. 20). The scale is one to five points. In the DVD and Book categories, we can see the outputs fit approx. 70% to the users' impression. But again, electronics including HDD Recorders are inconsistent with the impression probably due to ad blogs.

Associated Topic extraction

In this section, we try to verify an assumption that associated topics extracted from the blog thread are more suited to the users' impression than those retrieved according to the word frequency. In the experiment, we first asked the test users to carefully read the blog threads for the same three target products, and to recommend some of the products which were frequently compared with the target products. Then, we divided the number of recommendations for each product by the number of the test users to obtain the degree of association for each product. Users Selection in Table 4 shows the results for a target product in the DVD category [2].

Next, we calculated TF/IDF for the above recommended products from the same blog threads. The results are shown in the right-hand column in Table 4. Here we can see that the results for some products match the users' impression, but the results for Product A to G are not aligned with the impression. A reason for this result is that the frequency of those products was not high; in fact, it was just once. However, these products appeared in a hub blog in the corresponding thread, which has many trackbacks. So the simple frequency method ignored these associations. On the other hand, the middle column in the same table shows the result

[2] the target is "Charle and The Chocolate Factory", and the product A to I are: "Pirates of The Caribbean", "Finding Neverland", "Edward Scissorhands", "SLeepy Hollow", "Tim Burton's Corpse Bride", "Charle and The Chocolate Factory (Book)", "The Load of The Rings", "Willy Wonka & The Chocolate Factory", "Nightmare Before Christmas".

Table 4: AT impression, result, and TF/IDF

Products	Users Selection (%)	AT (%)	TF/IDF (%)
Product A	50.0	77.8	29.5
Product B	39.0	77.8	29.5
Product C	39.0	36.8	15.4
Product D	22.0	8.3	4.2
Product E	22.0	8.3	4.2
Product F	17.0	8.3	4.2
Product G	6.0	8.3	4.2
Product H	33.0	49.4	29.6
Product I	11.0	19.4	14.1

of the AT extraction. We find that those products are taken as the associated topics due to the reflection of the blog correlations. In fact, the average difference between the users' impression and TF/IDF was 12.2%, but the difference in the case of the AT extraction was 7.2%.

Further, we are using the product ontology in the AT extraction to eliminate the frequent words other than product names in the same domain. In the above comparison, TF/IDF was calculated only for the products recommended by the users, in order to check the improvement of the AT extraction. But, if we pick up the frequent words gathered by TF/IDF, they include many general terms that are not product names. However, the AT extraction filters out these irrelevant words by referring to the product ontology.

Search Time

Currently, it takes approx. 10 - 30 (sec) to collect a maximum of 100 blogs, do the P/N determination and the AT extraction, and then return the results with a maximum of 20k blog bodies to the cellular phone. The time varies due to the amount of collected sentences. It should be noted that the blog crawling is started after receiving a search request in the current system. We plan to make a robot and an indexing mechanism in the near future. It should also be noted that Product Metadata, which bind barcode numbers to product information such as title, author, manufacturer, publisher, and date, are prepared for 1,400,000 products.

Fig. 5 illustrates the relationship between the search time and the users' impression. The users evaluated the turnaround time for each search with five points. It is found that acceptable waiting time for mobile users is much longer than for users in the desktop environment, but a response within 20 (sec) is necessary for a good impression.

Then, Fig. 6 shows how much time each function consumed in a user's multiple searches. After all, blog collection is time-consuming and another reason is our current server spec., Pentium 4, 3.2 GHz and 1GB memory.

Discussion on Practical Use

The cooperation of retailers is a prerequisite for the practical application of this system in stores. We believe the system offers the following advantages to retailers.

- It could provide supportive information that prompts consumers to make a purchase decision who might otherwise

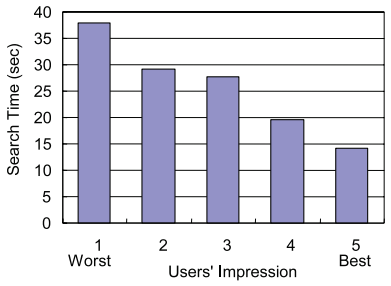

Figure 5: Search time and users' impression

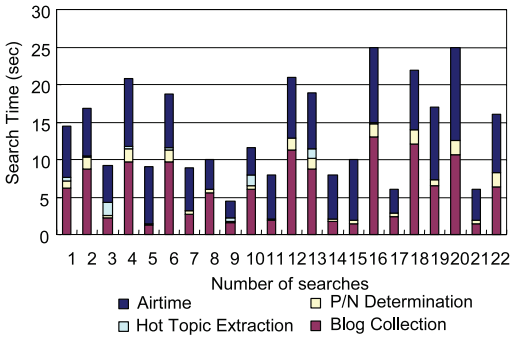

Figure 6: Search time for each component

hesitate. In fact, 70% of product review blog entries are positive according to a survey(Web Advertising Bureau 2006).

- A good reputation on the internet can be directly linked to sales in bricks-and-mortar stores. Moreover, if retailers use this system, they can take beneficial actions in advance, such as reducing inventory of products that have a bad reputation on the internet.
- In terms of the AT extraction, if the retailers do not have the suggested products at that time, they can be guided to order them.

Further, it should be noted that this service does not compare prices at stores nor the reputations of stores. Thus, it would not result in information that could undermine the profitability of retailers.

Related Work

In terms of using barcodes, there is Amazon Scan Search(Amazon 2005). When the user scans a barcode on a book with a cellular phone, the corresponding page in Amazon.com is shown if such a page exists. Also, Microsoft is experimenting with AURA (Advanced User Resource Annotation System)(Brush *et al.* 2005) in the US. In this system, PDAs with attached CF-type barcode readers scan the barcodes on products and search the related information on Google or eBay. Further, the system includes a web site where users can freely annotate products.

In terms of using WOM, threre are some blog search engines featuring word-of-mouth information(Nanno *et al.* 2004)(Furuse *et al.* 2007). It employs positive and negative classifications, and shows the degree of burst (trend) with a time line.

Conclusion and Future Work

This paper proposed a mobile service to retrieve WOM information from weblogs using semantic information by means of barcodes and a cellular phone. We think Ubiquitous Metadata Scouter is poised to provide the instantaneous benefits of using semantics in a ubiquitous computing environment. In light of the results of the public experiment, subjects for future work include improvements to the accuracy of the precision ratio and to the performance, such as making possible multiple access.

References

Amazon. 2005. Amazon scan search. `http://www.amazon.co.jp/exec/obidos/tg/feature/-/546374/249-9948758-541%6367`.

Brickley, D. 2000. Rdf site summary (rss) 1.0. `http://purl.org/rss/1.0/spec`.

Brush, A.; Turner, T.; Smith, M.; and Gupta, N. 2005. Scanning objects in the wild: Assessing an object triggered information system. In *Proc. of 7th Int. Conf. on Ubiquitous Computing (UbiComp2005)*.

Dave, K.; Lawrence, S.; and Pennock, D. M. 2003. Mining the peanut gallery: Opinion extraction and semantic classification of product reviews. In *Proc. of the 12th Int. World Wide Web Conference (WWW2003)*.

Furuse, O.; Hiroshima, N.; Yamada, S.; and Kataoka, R. 2007. Opinion sentence search engine on open-domain blog. In *Proc. of 20th Int. Joint Conf. of Artificial Intelligence (IJCAI2007)*.

Kimura, M.; Saito, K.; Kazama, K.; and Sato, S. 2005. Detecting search engine spam from a trackback network in blogspace. In *Proc. of 9th Int. Conf. on Knowledge-Based and Intelligent Information and Engineering Systems (KES2005)*.

Kobayashi, N.; K. Inui, Y. M.; Tateishi, K.; and Fukushima, T. 2005. Collecting evaluative expressions for opinion extraction. In *Proc. of First Int. Joint Conf. (IJCNLP2004)*, LNAI 3248.

Kusumura, Y.; Hijikata, Y.; and Nishida, S. 2004. Ntm-agent:text mining agent for net auction. *IEICE Transactions of Information and Systems* E87-D(6):1386–1396.

Nanno, T.; Fujiki, T.; Suzuki, Y.; and Okumura, M. 2004. Automatically collecting, monitoring, and mining japanese weblogs. In *Proc of 13th Int. World Wide Web Conference (WWW2004)*.

Turney, P. D. 2002. Thumbs up or thumbs down? semantic orientation applied to unsupervised classification of reviews. In *Proc. of 40th Annual Meeting of the Association for Computational Linguistics (ACL02)*.

Web Advertising Bureau. 2006. `http://www.wab.ne.jp/english/index.html`.