Subjective Information Indexing Technology Analyzing Word-of-mouth Content on the Web

Hisako Asano†*, Toru Hirano, Nozomi Kobayashi, and Yoshihiro Matsuo*

Abstract

This article introduces subjective information indexing technology for extracting, in advance, a large amount of diverse word-of-mouth content scattered throughout the World Wide Web and turning that content into knowledge. This technology will make it easy for users to access relevant word-of-mouth information in weblogs (blogs) and will facilitate the analysis and tabular display of that information from various points of view.

1. Word-of-mouth information

Consumer generated media (CGM) such as weblogs (blogs) and online forums are becoming a familiar element of our daily lives. Whenever something is happening in society, there is a good chance that many people will be checking CGM for word-of-mouth content about that subject. If the subject is limited to a specific genre, such as traveling, users can easily obtain a great deal of information by visiting word-of-mouth sites concerned with travel. However, the subjects that users want to check change from day to day if not from moment to moment. For example, at one time a user may be interested in how everyone is reacting to the news that a certain well-known celebrity couple has announced wedding plans, while at another time, he or she may be curious about the opinions of other people about a certain cell phone appearing in a commercial. However, users may have to perform various kinds of searches on their own depending on the subject in question: a user might use a blog search to learn about people's reactions to that wedding announcement and might access word-of-mouth sites to get opinions about that cell phone. In short, a good deal of knowledge as well as time and effort are necessary to obtain various types of word-of-mouth information.

† NTT Cyber Space Laboratories
Yokosuka-shi, 239-0847 Japan
Email: asano.hisako@lab.ntt.co.jp

2. Subjective information indexing technology

Subjective information indexing technology was developed to extract and make use of a wide variety of opinions written about all kinds of subjects including people, products, and retail establishments and to simplify access to word-of-mouth/subjective information, like that described above. This technology consists of subjective information extraction and summarization (**Fig. 1**). These two processes are explained below.

Subjective information extraction automatically extracts subjective information (such as "the X905i has a fine screen" or "the night view from Building X is beautiful") from documents such as blogs. We consider subjective information to consist of a triplet (subject, feature, and evaluation), and it is these three elements of subjective information that we seek to extract. Here, subject indicates the evaluation topic (e.g., X905i (a cell phone model)), feature indicates what aspect of the subject is being evaluated (e.g., the screen), and evaluation expresses an opinion about it

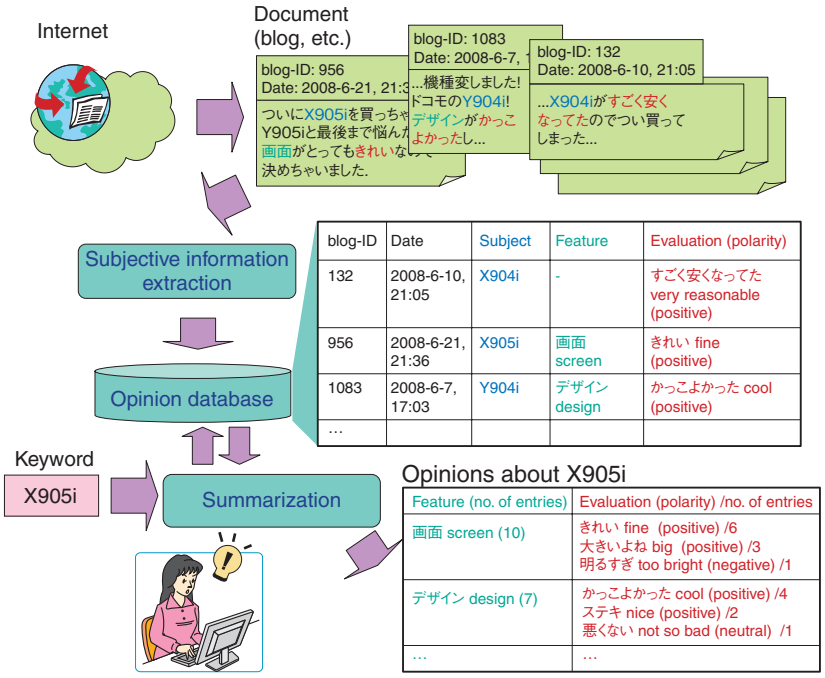

Fig. 1. Outline of subjective information indexing technology.

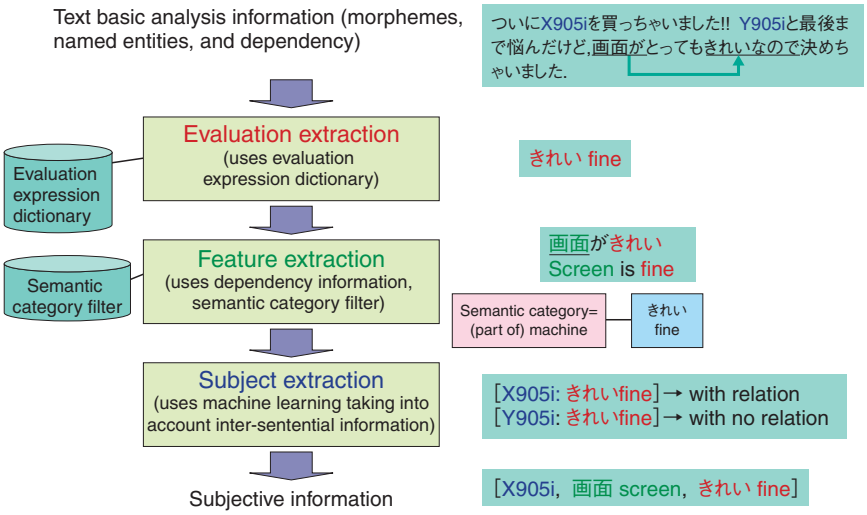

Fig. 2. Subjective information extraction process.

(e.g., fine).

Summarization tabulates opinions of interest to the user along various axes such as feature (screen, design, etc.) and time (e.g., 1-month frequency distribution) and processes and outputs that information for display and analysis purposes. The main characteristics of these processes are described in Sections 3 and 4.

3. Subjective information extraction process

The flow of the subjective information extraction process is shown in **Fig. 2**. This process makes use of basic analysis information described in the fourth article in this special feature entitled “Basic Japanese Text Analysis Technology as a Platform for Knowledge Extraction”.

3.1 Evaluation extraction

The process begins by extracting evaluations that appear as expressions like きれい (fine) and 元気いっぱい (vigorous). Evaluation extraction makes use of an evaluation expression dictionary having several tens of thousands of expressions. This dictionary consists of evaluation expression patterns and their associated polarity (positive, negative, neutral), each represented as a word sequence like きれい fine (positive) or 元気いっぱい vigorous (positive). In addition, the word sequences preceding and following an extracted evaluation expression pattern are used to adjust the range of that expression and its polarity. (Example: きれい fine (positive) → 全然きれいじゃなかった not fine at all (negative)).

3.2 Feature extraction

Next, the process extracts features corresponding to extracted evaluations. It does this by using dependency information and a semantic category filter that indicates whether a semantic relation exists between feature candidate words and evaluation expressions. For example, since the subject of きれい (fine), the extracted evaluation expression in Fig. 2, is 画面 (screen), and because きれい (fine) can be an evaluation of the semantic category[*1] of "(part of) a machine" to which 画面 (screen) belongs, 画面 (screen) passes through the semantic category filter and is extracted as a feature.

Using semantic categories in this way suppresses the erroneous extraction of features. For example, given the statement X905iは友だちが使いやすいって (my friend says that the X905i is easy to use), the semantic category filter prevents 友だち (friend) from being extracted as the feature modified by 使いやすい (easy to use).

3.3 Subject extraction

Finally, the process extracts the subject corresponding to extracted evaluation expressions using the machine learning technique. First, it prepares all possible pairs of previously extracted evaluations and subject candidates. A subject candidate may be a named entity such as a person name or location name or a general term indicating a topic of discussion (e.g., flu). In the example in Fig. 2, we get two such pairs: [subject candidate: evaluation] = [X905i: きれい fine], [Y905i: きれい fine]. For each of these pairs, the process determines whether the subject candidate in question could be the subject of that evaluation. The procedure for examining the pair [Y905i: きれい fine] is depicted in **Fig. 3**. We note here that this subject extraction process does not limit itself to dependency information and other types of information in just one sentence. It also uses information that spans multiple sentences, namely, omitted information and text boundary information [2]. In the figure, something appears to be omitted before きれい (fine) in the second sentence (in this case, X905i), and this is treated as omitted information. Text boundary information indicates whether the pair of words in question lies in the range of the same topic. For the pair [Y905i: きれい fine] in the figure, it is determined that there is no text boundary (i.e., there is only one topic) within the same sentence. In short, omitted information and text boundary information are used as a basis for determining whether a subject can be obtained with respect to all of the pairs under scrutiny. The subject candidate of the pair for which it is determined that a subject can be obtained is regarded as the subject of that evaluation expression.

3.4 Storage of subjective information

In the example in Fig. 2, the above evaluation, feature, and subject extraction processes extract a triplet of subjective information: [X905i, screen, fine]. The precision of extracting it from blogs is about 80%.

To enable the extracted subjective information to be used in the summarization process and for output display, notations appearing in the original text are stored in a database along with their standard forms and end forms. Some examples of such notational variations are given in **Fig. 4**.

Standard form is normalized notation that excludes, for example, degree words and adjunct words. This makes it possible in summarization processing to group together and process information that, while having different notation, signifies the same thing. The normalization process uses standard forms of words, and in the case of declinable words, it also converts the original expression to its end form (example: とっても強くって very powerfully →強い powerful). For subjects, however, the normalization process will use, if available, ground information as described in the third article in this special feature entitled "Grounding Named Entities for Knowledge Extraction".

Normalization (identification) of a subject is particularly useful in reducing misses in opinion searching. For example, consider an opinion search based on the keywords 電電太郎 (*Denden Tarō*), given that

*1 Semantic category: The process uses about 2700 categories from Nihongo GoiTaikei [Japanese Lexicon] [1].

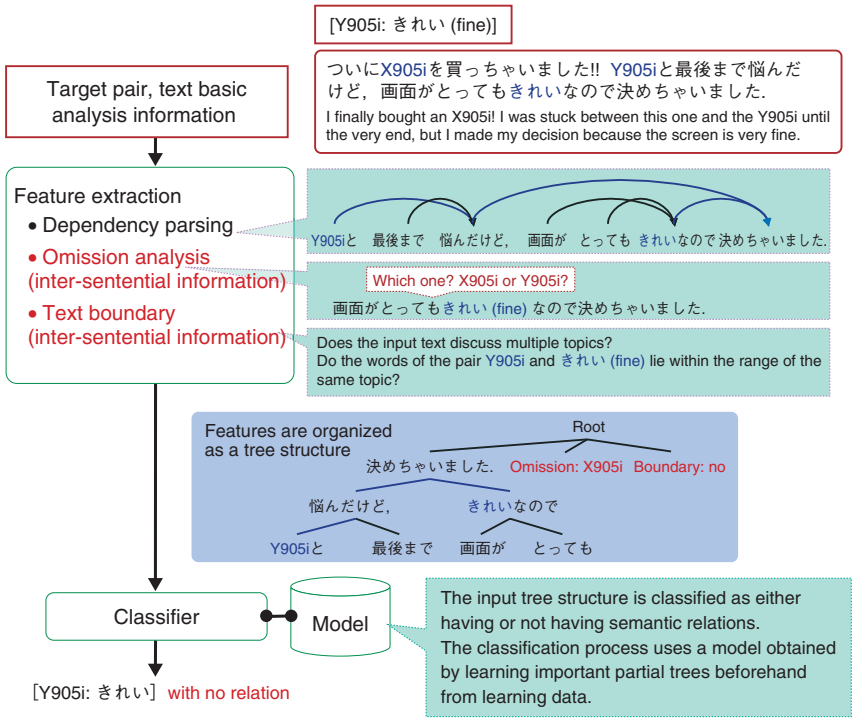

Fig. 3. Subject extraction.

…昨日サッカーを見に行きました．タローくんの最後のセーブがとってもかっこよくって，…
... I went to see the soccer match yesterday. Taro's final save was really cool, ...

Notation type	Subject	Feature	Evaluation
Original notation	タロー *Tarō*	最後のセーブ final save	とってもかっこよくって really cool
Standard form	電電太郎 *Denden Tarō* (identified by grounding technique)	セーブ save	かっこいい cool
End form	—	—	とってもかっこよい really cool

Fig. 4. Notational variation of subjective information for summarization and display.

a soccer player called 電電太郎 (*Denden Tarō*) is also well known as simply タロー (*Tarō*). Although the expression 電電太郎 (*Denden Tarō*) is not written in the text shown in Fig. 4, the fact that the subject's standard form タロー (*Tarō*) is 電電太郎 (*Denden Tarō*) makes it possible to extract the opinion "セーブがかっこいい" (the save was really cool) from that text and apply it to search results.

In the case of notation in which the trailing portion of an evaluation expression is used in a declinable form, only that portion will be converted to the end form, which can be used for displaying opinions in a tag cloud[*2] style.

4. Summarization process

The summarization process outputs the results of

*2 Tag cloud: A display of tags corresponding to certain items, commonly used on Web sites that use tags for social bookmarking and other purposes. Tags with higher usage frequency are displayed in larger font sizes.

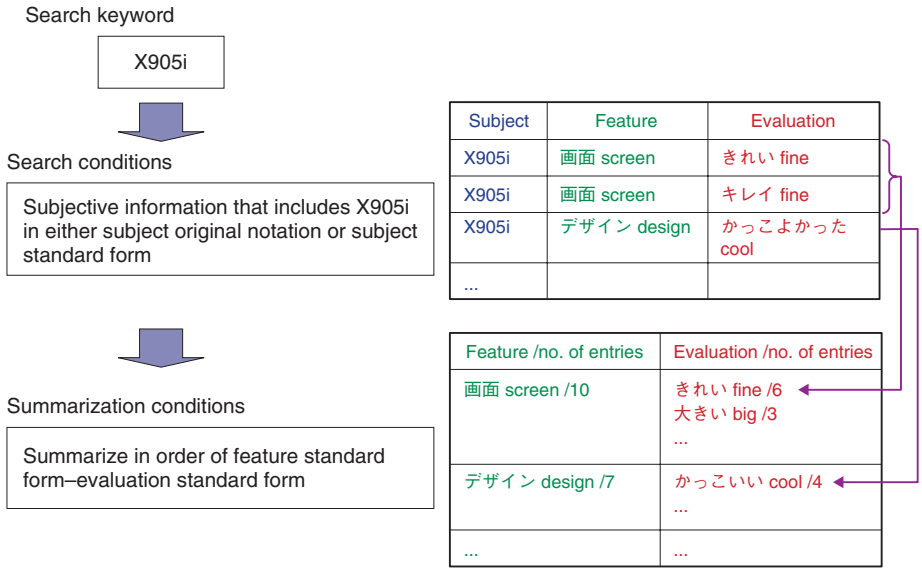

Fig. 5. Summarization process in opinion search.

opinion tabulation with respect to subjective information stored in the opinion database and additional information (like dates and times) associated with the source documents from which that subjective information was extracted. Results are output based on search conditions that determine what subjective information is to be gathered and summarization conditions that determine how that information is to be grouped.

The flow of summarization processing in opinion searching is shown in **Fig. 5**. The process begins by specifying search conditions, which here consist of subjective information that includes the search keyword X905i as the subject of evaluation. The process now returns the results of that search. Next, the process specifies summarization conditions, which here state that features shall be listed in standard form (screen, design, …) with each feature combined with an evaluation expression also in standard form (fine, big, …). This process results in the display of subjective information in units of features.

5. goo Opinion Analysis service

Subjective information indexing technology has been used on the goo portal site operated by NTT Resonant since the launch of the goo Opinion Analysis service [3] in May 2007. This service has three functions: analyze, compare, and search for related terms.

A screen shot of the analyze function is shown in **Fig. 6**. It displays the results of an opinion analysis for any keyword. The feature area on the left side displays retrieved opinions about each type of feature (equivalent to the output of the summarization process shown in Fig. 5). This display lets a user focus on a specific feature of a certain product and examine the opinion summarized for that feature. The time-series graph area on the right side shows the frequency of positive, negative, and neutral subjective information in the form of a time-series graph. Also shown are opinions for four different periods (weeks) in tag-cloud format.

The compare function quantifies and compares the opinions for two to three keywords. The “search for related terms” function searches for words expressing topics and opinions related to the input keyword.

6. Future plans

This article described subjective information indexing technology and introduced the goo Opinion Analysis service as an example of applying that technology to an opinion-searching service. In future research, we plan to investigate the information needed for providing even more detailed opinion analysis services and for applying this technology to targeted advertising with the aim of developing more portal services. In parallel, we plan to investigate the expansion of this technology to the corporate-oriented marketing business. For example, this technology could be used as a mining tool to analyze the differ-

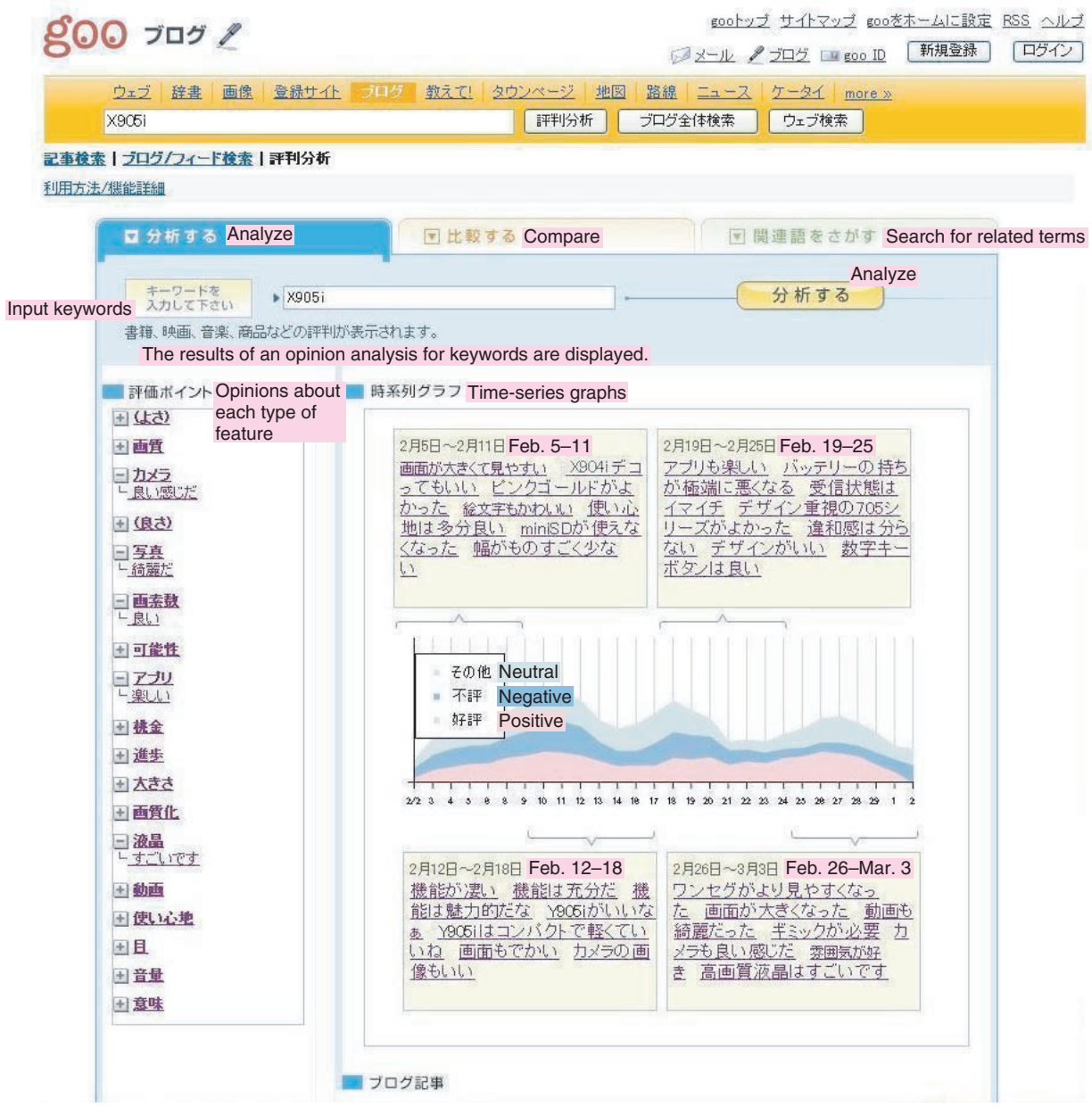

Fig. 6. goo Opinion Analysis service.

ences in features between one's own products and those of another company with the aim of developing products that stand out.

References

[1] S. Ikehara, M. Miyazaki, S. Shirai, A. Yokoo, H. Nakaiwa, K. Ogura, Y. Ooyama, and Y. Hayashi, "Nihongo GoiTaikei [Japanese Lexicon]," Iwanami Shoten, 1997 (in Japanese).

[2] T. Hirano, Y. Matsuo, and G. Kikui, "Detecting Semantic Relations between Named Entities in Text Using Contextual Features," Proc. of the 45th Annual Meeting of the Association for Computational Linguistics Companion Vol. Proc. of the Demo and Poster Sessions, pp. 157–160, 2007.

[3] http://blog.search.goo.ne.jp/wpa/ (in Japanese).

Hisako Asano
Senior Research Scientist, Speech, Acoustics and Language Laboratory, NTT Cyber Space Laboratories.
She received the B.E. degree in information engineering from Yokohama National University, Kanagawa, in 1991. She joined NTT Information Processing Laboratories in 1991. Her research interests include natural language processing, especially morphological analysis, information extraction, and text analysis for text-to-speech synthesis. She is a member of the Information Processing Society of Japan (IPSJ) and the Association for Natural Language Processing (NLP).

Toru Hirano
Research Engineer, Speech, Acoustics and Language Laboratory, NTT Cyber Space Laboratories.
He received the B.E. degree in systems engineering from Wakayama University, Wakayama, and the M.E. degree in information science from Nara Institute of Science and Technology, Nara, in 2003 and 2005, respectively. He joined NTT Cyber Space Laboratories in 2005. He is a member of NLP.

Nozomi Kobayashi
Research Engineer, Speech, Acoustics and Language Laboratory, NTT Cyber Space Laboratories.
She received the M.E. and Dr.Eng. degrees in information science from Nara Institute of Science and Technology, Nara, in 2004 and 2007, respectively. She joined NTT Cyber Space Laboratories in 2007. She is a member of IPSJ.

Yoshihiro Matsuo
Senior Research Engineer, Supervisor, Speech, Acoustics and Language Laboratory, NTT Cyber Space Laboratories.
He received the B.S. and M.S. degrees in physics from Osaka University, Osaka, in 1988 and 1990, respectively. He joined NTT Communications and Information Processing Laboratories in 1990. His research interests include multimedia indexing, information extraction, and opinion analysis. He is a member of IPSJ and NLP.