18

An Information Extraction Support System from BBS Using Topic Mapping Based on Structure of Conversation with a 3-dimensional Space

Ayako Hiramatsu*, Tatsuya Nakae**, Hiroshi Shibata**, Norihisa Komoda**
*Osaka Sangyo University, **Osaka University

Abstract: This paper proposes a support system for the information extraction from BBS (Bulletin Board System). On BBS in EC (Electronic Commerce) sites, participants who are consumers write frank opinions easily. This produces active conversation as enormous text data and includes unexpected opinions and consumers' requirements. In this system, opinions on BBS are classified by topics and arranged on a 3-dimensional space so that users (for example product designers) can easily understand how topics are interwoven into conversations or how topics are related to each other. Furthermore, two methods using the feature of the conversation structure on BBS are proposed for correcting the recognition of opinions.

Key words: information extraction, bulletin board system, electronic commerce

1. INTRODUCTION

Recently, the Internet is being used for marketing (Finch, 1999 and Finch and Luebbe, 1997) and on many EC (Electronic Commerce) sites for consumers, a BBS (Bulletin Board System) is provided as a way to gather consumers' opinions about request for product on the Internet. Opinions written on BBS include consumers' requirements for products and are useful for new product design. Furthermore, on BBS, many people talk about their problems, requests, and experiences from various views. So, a topic generated by someone can be developed to new related topics by other people. The conversation acts as a chain reaction and can provide novel and divergent ideas for product designers. However, the conversation among consumers on BBS is stored as just text data and product designers have a hard time of extracting effective and useful information from the massive

The original version of this chapter was revised: The copyright line was incorrect. This has been corrected. The Erratum to this chapter is available at DOI: 10.1007/978-0-387-35617-4_48
J. L. Monteiro et al. (eds.), *Towards the Knowledge Society*

amounts of text data. Therefore, there is a real need for a system that extracts information about consumers' requirement.

In the text data written on BBS (called the BBS data), the opinions about various topics are mixed along the passage of time. Besides, there are complex relationships between opinions such as an approval or an objection to an opinion and a reply to a question. By the way, existing support methods for the information extraction (Stefik, Foster, Bobrow, Kahn, Lanning, 1987 and Sugiyama, Misue, Watanabe, Nitta, Takeda, 1997) decide similarities among opinions and classify by contents using the cluster analysis or the multidimensional scaling. The information provided by the existing methods is only the results of classifications by contents and not based on the concept of conversations. The existing methods cannot provide the relations between classified groups and conversational flows. Hence, with the existing methods, product designers cannot understand how opinions are extended from a certain opinion in a focused topic, what topics are developed by conversations, and which groups of opinions are related to a certain topic. As mentioned above, the existing methods can support classifications of opinions by contents, but cannot provide conversational flows and flows of the developed topics so that the product designers cannot gain information for unexpected ideas.

We are developing a support system that aims at the following two things: 1.exactly extracting information about consumers' requirements from the conversations on BBS, 2.supporting unexpected ideas for new product design. The features of this proposed system are visualization BBS data using a 3-dimensional space and providing the relations between contents and conversational flows. The 3-dimensional space consists of the topic plane and the time axis. BBS data is classified to some clusters based on contents and the clusters are arranged on the topic plane. By adding the time axis to the topic plane, the arranged clusters are expressed as oval spheres whose sizes are changed by how long and how many opinions are included in the clusters. The relations between contents and conversational flows are expressed as arcs that might connect opinions in different clusters (It shows that the conversation strays to other topics). By tracing these arcs, users who are product designers can easily watch similar topics that participants' association created.

On BBS, participants often write their opinions in colloquial style, which often results in omission of important words. In the process of classifying opinions by contents, if we use only words for calculating similarities, the results of classifications have improper classification. To cope with this problem, our proposed system utilizes the feature of BBS conversation for the correction of word omission.

At the end of this paper, we show an experiment for appropriateness of the classification. Furthermore, with comparative experiments of the requirement extraction, we verify whether our proposed system can support

the exact extraction of information about consumers' requirements and the creation of unexpected ideas.

2. GATHERING OPINIONS WITH BBS

BBS is an application of computer online services, in which participants of a network are welcome to write and read their opinions on a bulletin board like Figure 1. In this chapter, we describe the features and problems of gathering opinions with BBS.

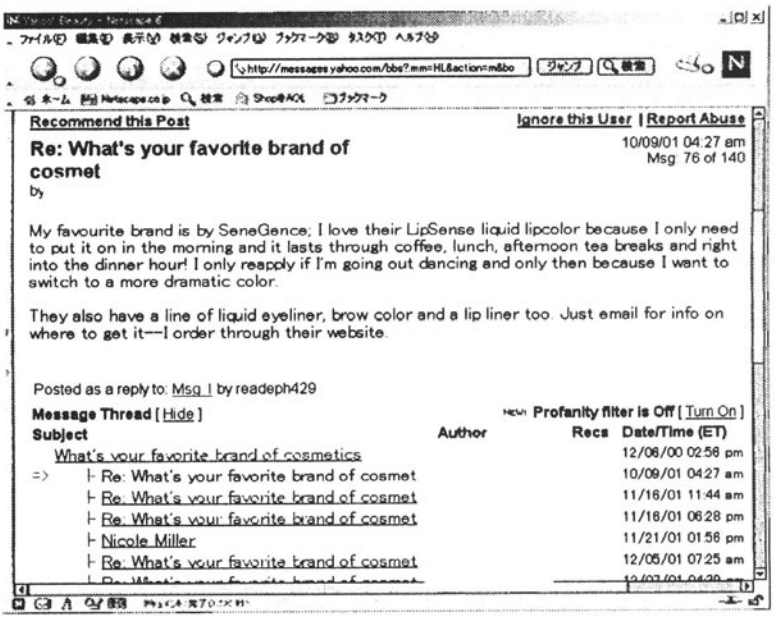


Figure 1: An example of BBS

2.1 Free Format Answer

The formats of gathering opinions from consumers are roughly two types: structured and free format. BBS is one way of gathering opinions using a free format. Compared with the structured formats with which respondents choose from pre-selected alternatives, with the free formats inquirers can get unexpected ideas, detailed opinions, and reasons for opinions. However, because of qualitative answers, gathered opinions cannot be analyzed quantitatively by conventional statistical methods. As an analyzing method for qualitative opinions, to gain effective information, inquirers classify opinions with the free format, based on the affinity diagram method (KJ method) (Swanson, 1995) etc. Because the opinions

gathered with BBS amount to the thousands in a week, the support of classifying opinions with computers is necessary.

2.2 Interaction among Consumers on BBS

The opinions on BBS are written in the form of conversations among consumers, because participants can write their opinions freely and easily. Here, we regard a series of written opinions (for instance a question and replies) as a conversation. BBS can perform the same function as the more unrestricted group interview. In the case of the questionnaire method both in the structured format and the free answer form, there is no relation among participants, each opinion is independent and the concept of the conversation does not exist. Therefore, we consider that the following features of conversation on BBS are important.

- Development of topics

  By active conversation among participants who are consumers, the ideas of consumers are generated and new topics are developed that both participants and product designers do not expect. Getting similar and unexpected topics, it is expected that the product designers can discover novel ideas for product planning.

- Word omission

  The conversation among consumers on BBS is realized by writing opinions in colloquial style. The content recognition of written opinions with colloquial style often lacks important keywords because participants drop the tacitly understood words and substitute reference terms for the repeatedly written words. Accordingly, the exactly recognition of contents cannot be realized without taking into account the context of the conversation flows. The product designers need to read opinions along the conversation flows.

2.3 Classifications by contents for BBS data

One of the famous methods for analyzing qualitative opinions is the KJ method (Swanson, 1995). The KJ method utilizes human graphical and visceral thought for structuring information and finds out the total semantic contents and the relationships of information from a large amount of data. The actual structuring method is the following: 1. Fragments of opinions are written on cards. 2. The opinion cards are arranged on the plane of a worktable by putting similar opinion cards close to each other. In the process of the structuring, if new views or relations are discovered, the opinions are re-structured again. With this structuring, the information written on the cards is classified by contents and analysts can find out some contributory hypothesis from the structured information. Such classifications let analysts

who are product designers know effective opinions. Moreover, the classifications help product designers to discover unexpected ideas by deriving the relations between various topics. However, on BBS of hot WWW sites, participants write thousands of opinions in a week. With a large number of opinions, manually classifying and associating becomes hard works.

Some support systems to classify qualitative opinions written in the free format based on the KJ method etc. have been proposed. For examples, Colab (Stefik, Foster, Bobrow, Kahn, Lanning, 1987), D-ABDUCTOR (Sugiyama, Misue, Watanabe, Nitta, Takeda, 1997), KJ editor (Ohiwa., Kawai, Koyama, 1990), and GUNGEN (Munemori and Nagasawa, 1996) were developed for structuralizing opinions based on the KJ method. In addition to these systems, some support systems (Romano, Bauer, Chen., and Nunamaker, 2000 and Chen, Titkova, Orwig, Nunamaker, 1998 and Frasconi, Gori, Soda, 1999) for recognizing qualitative opinions with other methods have been developed. However, these systems apply manually pre-processed data that computers can process easily. They do not deal with actual opinions or the conversation flows that are features of BBS.

Therefore, in this paper, for applying to BBS data, we propose a requirement extraction support method with the exact recognition of the contents and the developing topics. By applying this proposed method to an application on WWW browsers, we try to help product designers extract consumers' requirements and discover unexpected ideas.

3. REQUIREMENT EXTRACTION SUPPORT SYSTEM WITH 3-DIMENSIONAL USER INTERFACE

3.1 3-dimensional Expressions of BBS Data

The necessary functions of our proposed support system are the following two things. The first is to support the classifications by contents because of complexity and enormous quantity of BBS data. The second is that the relations between the topic clusters and the conversation structures are provided to show how topics are developed. To satisfy these two things in our proposed system, the topic clusters and the conversation structures are arranged in 3-dimensional space that consists of the topic plane and the time axis as shown in Figure 2.

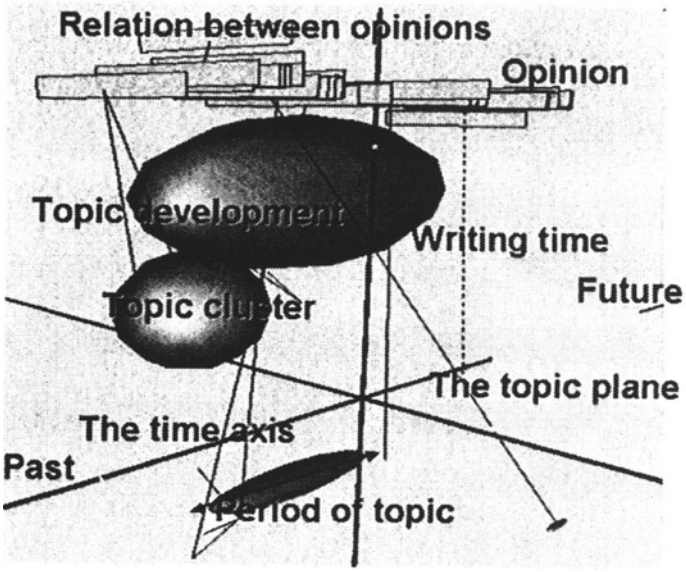


Figure 2: The 3-dimensional expression of BBS data

– The support for classifications by contents
The topic plane that consists of two axes shown in the Figure 2 presents the result of classified opinions by contents. To grasp contents of opinions on BBS, the degrees of similarities among opinions are defined basically with how many keywords are the same in opinions. The opinions are classified into topic clusters with the degree of similarities based on the cluster analysis. Using multidimensional scaling (Quantification Theory 4) hierarchically, the similarities between classified topics determine the distances between clusters and the similar topics are arranged closely on the topic plane. It helps users to become aware of the relations among topics.
– Providing the information about sequential developments of topics
In the case when two opinions are in a conversation flow, an arc connects the two opinions. The arc has a direction, which starts on a certain opinion and ends on its reply. Because the conversations flow with the progress of time, the inclusion of the time axis on the output space makes it possible to present the conversational relationship shown by the arcs. The arcs between opinions in the different topic clusters show the developments of topics. Tracing these arcs enable users to gain the information about how topics stray sequentially.

3.2 Requirement Extraction Support System

We are developing the requirement extraction support system that can transform the BBS data into the 3-dimensional information mentioned in the previous section. Figure 3 shows the outline of this proposed system. The

input data is only BBS data which is log file of BBS. The following shows the output and the problems for the transformation into 3-dimensional information。

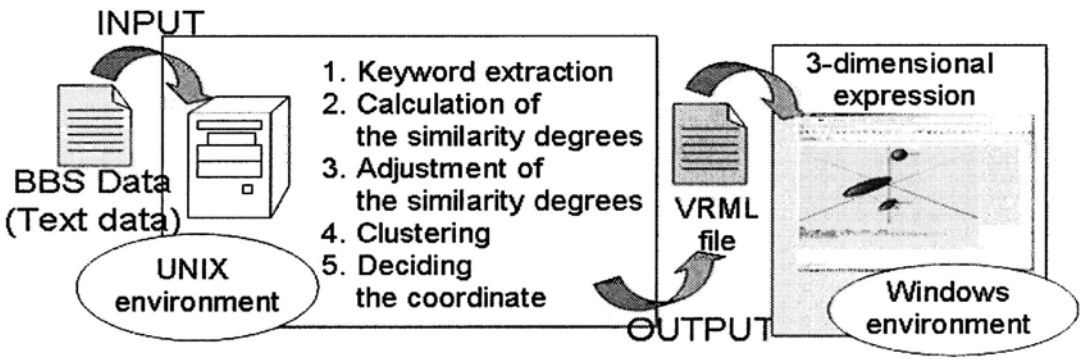


Figure 3: The outline of the proposed system

– Output

Using the standard language VRML (Lemay, Couch, and Murdock, 1996) that is excellent in the drawing of 3-dimensional information, the results of the analysis of the BBS data are presented. To express the 3-dimensional information within the limits of the screen, the output information is diagrammatically condensed with the Fish-Eye function (Shibata, Nozaki, Hiramatsu, Komoda, 2001). The topic or opinion labels expressed on the condensed diagrams are chosen with the order of priority based on what opinions the users click by mice. By using this priority, this system supports the movements of the attention opinions that are carried out frequently with the analysis process.

- The high-level output shows the result of the classification by contents and the conversation structures on BBS (shown in Figure 2).
- In the detailed output, the cluster chosen by a mouse click on the high-level output is opened and the detailed conversation and developed topics are presented (shown in Figure 4).

–Problems of the transformation

On BBS, with the conversation in the colloquial style, the omission of words for the content recognition often occurs. Such omission causes errors of the content recognition which is based on how many words are the same in different opinions. In the next chapter, we describe the solution to this problem.

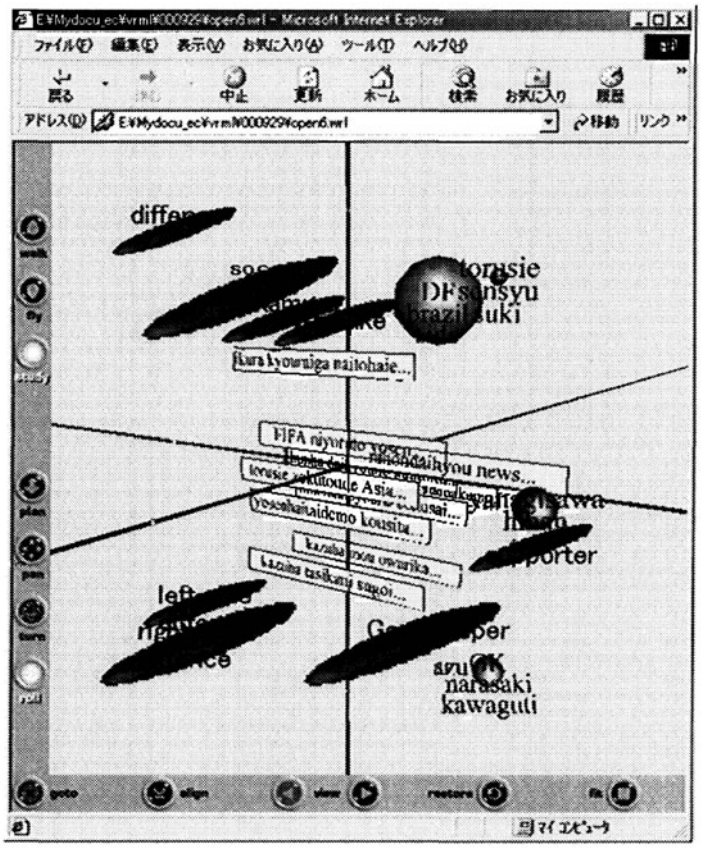


Figure 4: The detailed output

4. CORRECTION OF KEYWORD OMISSIONS FOR CONTENT RECOGNITION

In this chapter, we propose two methods to deal with omission of keywords because of the colloquial opinions.

4.1 Content Recognition with Keyword Matching

For the content recognition of opinions, the similarities have been calculated only by keyword information such as the included same words, the included similar words and the importance of the same included words. In the proposed system, the following formulas based on the Jaccard calculation formula calculate the similarity between opinions as the similarity coefficient J.

$$J_{mn} = \frac{\sum W_i a_i}{\sum W_i (a_i + b_i + c_i)} \quad (1)$$

J_{mn} is the similarity coefficient between the opinion m and the opinion n. a is the number of the same

words that both m and n include. b and c respectively are the numbers of words that one of m or n includes. W shows the weight of importance of each word.

4.2 False Recognition

For the conversation on BBS in colloquial style, keywords are often omitted in opinions and the word information cannot practice the exact similarity calculation and the content recognition. Figure 5 shows an example of the conversation flow, the similarity matrix based on the pre-explained similarity calculation between opinions and the clusters classified by the similarity.

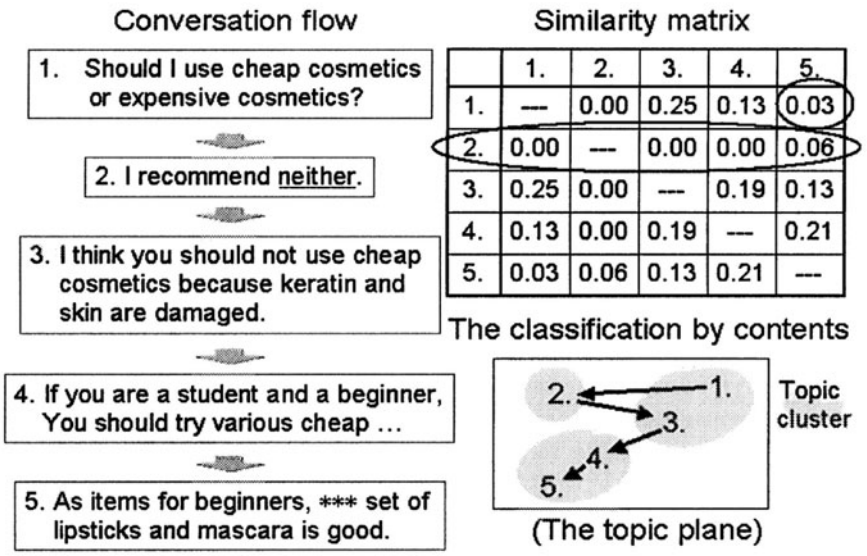


Figure 5: An Example of conversation on BBS, the similarity matrix and the classification

As shown in Figure 5, the result of the recognition of 5 opinions is the following.

- In Opinion 2, the demonstrative pronoun is used in the reply and the necessary keywords for the content recognition are omitted. For this reason, the similarities to other opinions in the same conversation are small and Opinion 2 is recognized as one opinion in a different topic.
- There are the common words each between Opinion 1 and 3, Opinion 3 and 4, and Opinion 4 and 5. The similarities between each pair of opinions are large. But, because there is no common word the similarity between Opinion 1 and 5 is small and these opinions are classified as two different clusters as the sequential topic development.

Usually one specific topic is talked about in one conversation. However, as mentioned above, the opinions in the one conversation are classified to different topics by the omission of keywords and the topic development. The classification by the omission of keywords is a false recognition and needs adjustment of similarities. On the other hand, the classification by the topic development is the exact recognition and doesn't need similarity adjustment. If the adjustment applies to all opinions in the one conversation, the latter exact classification is missed. In our proposed system, the similarities in the classification by the topic development are not adjusted and only the similarities with omissions of keywords are found out and adjusted. For this adjustment, we propose two methods explained in the next sections. The first method is based on what the similarities among opinions are concerned with conversation flows and finds the opinions which need to adjust its similarities. The second method is based on what the number of keywords in each opinion has a relation to the content recognitions. How many keywords are in an opinion is considered for the similarity calculation.

4.3 Identification of Omissions and Similarities Adjustments

Using the features of the conversation among consumers on BBS, we identify opinions including keyword omissions and adjust the similarities of the opinions. We show the process of the identification and the adjustment below.

1. Identification of omissions

   Like the similarity matrix among opinions shown in Figure 5, the opinion with omissions has very small similarities against the previous and next opinions. On the other hand, in the case of the topic development, the similarity between Opinion 1 and 5 is very small, but the similarities to the previous and next opinions of each opinion in the conversation keep certain values. It shows the topics are developed with some relevance. According to these differences, as shown in Figure 6(a), the opinion that has a pair of the very small similarity to the previous and next opinions is identified as an opinion including keyword omissions.

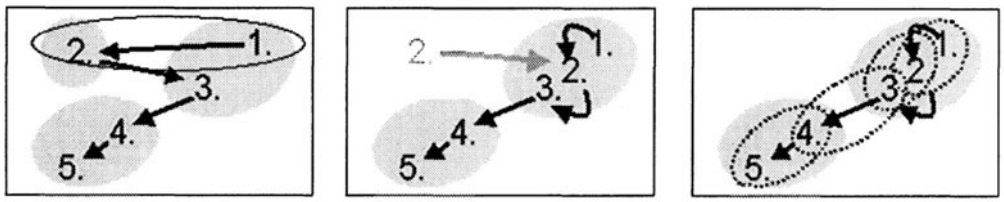
2. Adjustment similarities

   When a participant of BBS writes an opinion about a certain topic, he/she will write it as a reply for the most related opinion. Therefore, the similarities that the opinion with verbal omissions has are changed to a reply position as shown in Figure 6 (b).

3. Repeat of identification and adjustment

   The opinion on the first of the conversation usually includes no verbal omission. Therefore, we consider that the first opinion can be recognized exactly with the word information. For a start, the pair of the first opinion

and its reply is judged. The judgment is repeated while shifting the pair to the end of the conversation (Figure 6 (c)).



(a) The smaller similarity than a certain value

(b) Correction as a reply

(c) Repeatedly processing from the beginning to the last of the conversation

Figure 6: The identification and the adjustment of opinions with keywords omissions

With the above method, the information about the topic development is not lost and the keywords omissions are found out and adjusted.

4.4 Adjustment by the number of keywords

We considered that information quantity for the contents recognition that are included in an opinion is related to the number of retrieval keywords. Therefore, an opinion with few keywords cannot be recognized exactly and needs to adjust the similarity using conversation flows as contexts. For this adjustment, we use the following formula.

$$J_{mn} = J^{0}_{mn} + MW \times \frac{1}{K_n} \quad (2)$$

J_{mn} is the similarity coefficient between the opinion m and the opinion n. J^{0}_{mn} is the initial similarity before the adjustment. M is the max similarity in all target opinions. W is the weight parameter. K_n shows the number of keywords in the opinion n. (When K_n is 0, K_n is changed to 1.)

4.5 Experiment for Comparison of the Two Proposed Methods

For this experiment, we applied 53 opinions from BBS about PC and 50 opinions from BBS about TV games. On the BBS about PC, participants talk about the very narrow area topics. The BBS about TV games include wide area topics. These opinions are classified based on the similarity calculated by proposed two methods and a method with no adjustment. Namely, we

compared three classification results for each BBS. We evaluated these two methods with how these three classification results are different from the results classified by 10 human subjects. The difference is shown as values how many times the classification results need to move opinions to other cluster to become the same as results classified by human.

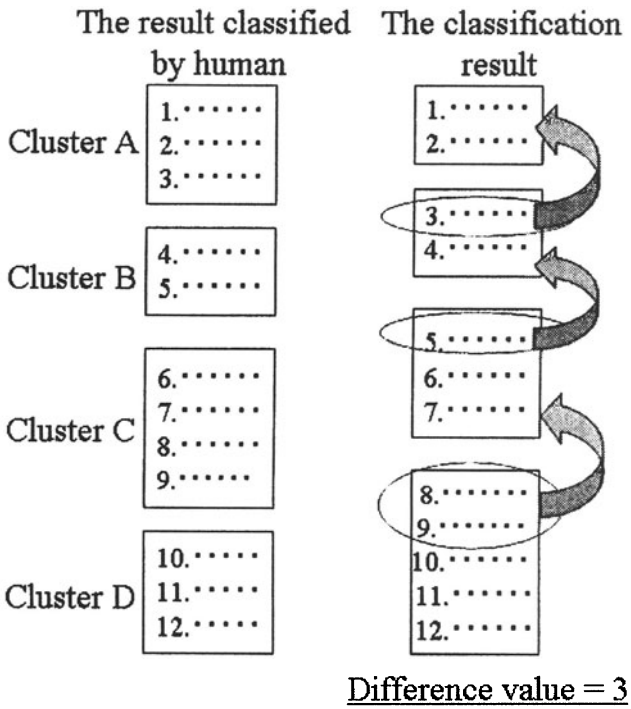


Figure 7: An example of calculating difference value

The example of difference value is shown in Figure 7. In this example, the classification result needs 3 times movements of opinions to become the same as the result classified by human and the difference value is 3. If difference values are small, the results are regarded as similar classifications. The difference values are calculated per one subject and one result. Therefore, one result has 10 difference values. The result of this experiment is shown as Table 1. The value in Table 1 is the average of 10 difference values. For the criterion, Table 2 shows the difference values among human results. In Table 2, the average difference values against each subject are shown. From these two tables, it is clear that the result of BBS about PC with the method by the number of keywords fell below the average value among human. This means that the method by the number of keywords could classify opinions of the narrow area topics like human. In the result of the BBS about TV games, the method by identification of omissions got the best value 3.2 of the three methods. Although this value is larger than the average among human, the value is within the standard deviation. Therefore, this experiment cleared to what kind of BBS each method is suited.

Adjusting Method / Target	No adjustment	By identification of Omissions	By the number of keywords
BBS about PC	3.6	3.7	2.5
BBS about TV games	3.5	3.2	3.5

Table 1: The average of difference values

Subject ID	BBS about PC		BBS about TV games	
	Average	Standard deviation	Average	Standard deviation
a	2.44	1.24	2.33	1.41
b	4.00	0.50	2.67	0.87
c	2.89	0.93	3.44	0.73
d	4.67	0.71	3.22	1.20
e	2.22	1.56	2.89	0.60
f	2.44	0.88	2.22	1.30
g	2.22	1.56	2.78	0.83
h	2.00	1.50	1.89	0.60
i	2.11	1.27	2.22	1.30
j	2.56	1.24	3.00	0.87
Total average	2.76	1.41	2.78	1.04

Table 2: The average of difference values among human

5. EVALUATION OF UNEXPECTED IDEAS SUPPORT

We have been developing the prototype of this support system that can provide the sequential developed topics with 3-dimentional expressions. Using this prototype system, we conducted a comparative experiment of consumers' requirement extraction. For this experiment, we use one hundred opinions concerned with cosmetics. For the comparison, the separate output method that provides the result of the classification and the conversation flows on BBS separately was also used. Each 3 subjects extracted requirement from BBS data with each support. Five other female students evaluated the extracted requirements according to the five-grade. Five female students evaluated the suitability and the novelty of the extracted

requirements. The suitability shows whether the subjects can extract requirements exactly. The novelty shows how unexpected and original the extracted requirements are. Therefore, if both two items are high level, the extracted requirement is useful for product designing. Moreover, we compared the number of read topics for extracting one requirement. With the separate output method, for the information of the contents classification and the conversation structure is not linked, users cannot refer related other topics to support the creation of an unexpected idea. Table 3 shows the result of the evaluation by questionnaires.

Table 3: The evaluation of extracted requirements (average)

Experiment method	Suitability	Novelty	The number of referred topics
The proposed system	4.2	3.5	2.8
The separate output	4.1	2.9	--------

From the result of the comparative experiment, the novelty of the extracted requirements by the proposed method exceeded 0.6 points in comparison with the separate output method. Besides, the reference of the other topics that cannot be performed with the separate output has been realized in the proposed system, and the users referred 2.8 topics for extracting a requirement on the average. As the result, users actually use the sequential developed topics provided by the 3-dimenstional outputs on the proposed system so that they create the unexpected ideas. Since the suitability by the proposed method keeps the same degree as the separate output method, the combinations of the other topics are not irresponsible and the related plural topics can support the exact grasp of consumers' requirements.

6. CONCLUSIONS

This paper has proposed the requirement extraction support system, which aimed at the exact grasp of consumers' requirements from opinions on BBS and the creation of unexpected ideas of the sequential developed topics. In this system, we proposed a similarity calculation method utilizing the conversational structure and realized the exact recognition and classification of opinions. The topic plane (which supports the understanding the classification visually) and the time axis compose the 3-dimensional space on which the opinions and the relationship between the conversations on BBS are arranged. It can display the link between the classifications and the conversational structures and can help product designers understand the

sequential developed topics easily. The prototype system based on the proposed method is actualized with programming language C and VRML that use 3-dimensional expressions. The requirement extraction experiments of the prototype system resulted in the exact classification with the contents corrections and support to the creation of suitable and divergent ideas with the grasp of the sequential developed topics.

7. REFERENCES

Byron J. Finch: *"Internet discussions as a source for consumer product customer involvement and quality information: an exploratory study,"* Journal of Operations Management No.17 pp.535-556 (1999).

Byron J. Finch and Richard L. Luebbe*: "Using Internet conversations to improve product quality: an exploratory study,"* International Journal of Quality and Reliability Management, Vol.14, No.8-9, pp849-865 (1997).

Stefik M., Foster G., Bobrow D. G., Kahn K., Lanning S.: *"Beyond the Chalkboard: Computer Suppport for Collaboration and Problem Solving in Meetings,"* Communication of ACM, Vol.30, No.1 (1987).

Sugiyama K., Misue K., Watanabe I., Nitta K., Takeda Y.*: "Emergent Media Environment for Idea Creation Support,"* Knowledge-Based Systems, Vol.10, No.1, pp.51-58 (1997).

Swanson, Roger C.: *"Quality Improvement Handbook: Team Guide to Tools & Techniques"*, Saint Lucie Press(1995).

H.Ohiwa., K.Kawai, and M.Koyama : *"Idea Processor and the KJ Method,"* Journal of Information Processing, Vol.13, No.1, pp44-48(1990).

Jun Munemori and Yoji Nagasawa: *"GUNGEN: groupware for a new idea generation support system"*, Information and Software Technology,Volume 38, Number 3; pp. 213-220(1996).

Romano,N.C.,Jr., Bauer,C., Chen,H., and Nunamaker,J.F.,Jr *: "The MindMine Comment Analysis Tool for Collaborative Attitude Solicitation, Analysis, Sense-Making and Visualization,"* Proc. of the 33rd Hawaii International Conference on System Sciences (2000).

Chen,H., Titkova,O., Orwig,R., and Nunamaker, J. F., Jr.: *"Information Visualization for Collaborative Computing,"* IEEE Computer, Vol. 31, No. 8, pp75-82 (1998).

Paolo Frasconi, Marco Gori, and Giovanni Soda*: "Data Categorization Using Decision Trellises,"* IEEE Transaction on Knowledge and Data Engineering, Vol.11, No.5 (1999).

Laura Lemay, Justin Couch, and Kelly Murdock: *"3D Graphics and VRML 2.0"*, Sams.net Publishing (1996).

Hiroshi Shibata, Takafumi Nozaki, Ayako Hiramatsu, and Norihisa Komoda: *"A Support System for Requirement Extraction from BBS using Hierarchical Fish-Eye User Interface,"* Proc. of 2001 IEEE International Conference on Systems, Man and Cybernetics (in CD-ROM)(2001).