Mining from Open Answers in Questionnaire Data

Hang Li*
NEC Corporation
4-1-1 Miyazaki, Miyamae-ku
Kawasaki, Kanagawa, JAPAN
+81-44-856-2143
lihang@ccm.cl.nec.co.jp

Kenji Yamanishi
NEC Corporation
4-1-1 Miyazaki, Miyamae-ku
Kawasaki, Kanagawa, JAPAN
+81-44-856-2143
k-yamanishi@cw.jp.nec.com

ABSTRACT

Surveys are an important part of marketing and customer relationship management, and open answers (i.e., answers to open questions) in particular may contain valuable information and provide an important basis for making business decisions. We have developed a text mining system that provides a new way for analyzing open answers in questionnaire data. The product is able to perform the following two functions: (A) accurate extraction of characteristics for individual analysis targets, (B) accurate extraction of the relationships among characteristics of analysis targets. In this paper, we describe the working of our text mining system. It employs two statistical learning techniques: rule analysis and Correspondence Analysis for performing the two functions. Our text mining system has already been put into use by a number of large corporations in Japan in the performance of text mining on various types of survey data, including open answers about brand images, open answers about company images, complaints about products, comments written on home pages, business reports, and help desk records. In this it has been found to be useful in forming a basis for effective business decisions.

Keywords

Text Mining, Survey, Questionnaire Data, Open Question, Classification Rules, Association Rules, Correspondence Analysis

1. ANALYSIS OF OPEN ANSWERS

Surveys are an important part of marketing and customer relationship management, and **open answers** (i.e., answers to open questions) in particular may contain valuable information and provide an important basis for making business decisions. Summaries of such answers made by human analysts, however, tend to be too intuition-based for satisfactory reliability, and the question of how to *automatically* summarize open answers and *automatically* mine useful information from them has become an important issue. Further, as a result of the ease of performing surveys on the web, more and more questionnaire data has become available, and manual handling of it all has become not only cumbersome but also costly. The development of a computer system that can automatically analyze open answers is sorely needed.

Table 1 Example Questionnaire Data

Car	Brand Image
Car A	For ordinary people
Car A	Easy to drive
...	...
Car B	High performance
Car B	Mobility
...	...

Let us consider the questionnaire data (translated from Japanese) in Table 1 above, referring to automobile brand images. If we are to automatically analyze the brand images of these cars, or more generally, the content of the open answers, such analysis should at least, in our view, perform the following two functions:

(A) accurate extraction of image characteristics for individual car types, and

(B) accurate extraction of the relationships among car types in terms of image.

In general, it is difficult to process answers in natural language because of the enormous varieties of linguistic expression. A more realistic approach would be to segment open answers into words and conduct an analysis at the word and phrase level. We believe that with this approach we can still obtain rather useful results. As will be described, a number of systems for survey results analysis have been developed on the basis of this assumption. To the best of our knowledge, however, there has been no system developed to date that can perform *both* of the two functions described above.

We have developed a survey analysis system that provides a new way to analyze open answers. The product[1] - our survey analysis system, which we we abbreviate as **SA** throughout this paper , works at the word and phrase level, and is able to perform the two functions described above. In this paper, we describe the working of **SA**. **SA** performs mining by using two statistical learning

KDD 01 San Francisco CA USA
Copyright ACM 2001 1-58113-391-x /01/08...$5.00

*Hang Li's current address is: Microsoft Research China, 5F Beijing Sigma Center No.49 Chichun Road Haidian District, Beijing China Tel: 86 10 6261 7711-3177 Email: hangli@microsoft.com

[1] This product is available by the name of SurveyAnalyzer in Japan. SurveyAnalyzer™ is a trademark of NEC Corporation in Japan.

techniques: **rule learning** (hereafter referred to as '**rule analysis**') and **Correspondence Analysis**.

Given, for example, the questionnaire data in Table 1, **SA** views each type of car as a category and views its brand image answers as texts (i.e., word sequences). It learns, for each type of car, **classification rules** for assigning brand image answers to the type, and it also learns **association rules** for associating the type of car with its answers. The rules obtained for each type of the car will clearly indicate the characteristics of the type (thus performing Function A).

SA views as random variables car types and keywords extracted from the full set of brand image answers, and it conducts **Correspondence Analysis** between the variables. It is thus able to visually display the correspondence relationship between cars and keywords on a two-dimensional map (Function B).

2. PREVIOUS WORK

Many methods have been proposed for analyzing **closed answers** (i.e., answers to closed questions, for which possible responses have been limited, as in check lists). These include those using such multivariate analysis techniques as cluster analysis (e.g., [1]) and correspondence analysis (e.g., [2]). Questionnaire data with only closed answers are relatively easy to handle because they are structured data. Questionnaire data having **open answers**, however, are not, even after they have been segmented into sequences of words, because the number of words is generally very large.

Methods for analyzing open answers by means of text-clustering techniques have been proposed (e.g., [8]). The idea here is to view each answer as a vector of words, and to cluster vectors on the basis of similarity measures. Such methods are effective for summarizing (grouping) answers, but they are not effective for extracting analysis target characteristics. Furthermore, since this approach is based on unsupervised learning, its accuracy can easily deteriorate with smaller data sizes.

Methods for analyzing open answers on the basis of associations between words have also been proposed (e.g., [5]). More specifically, associations between word pairs are calculated on the basis of their co-occurrences in open answers, and then are visually presented on a two-dimensional map. Although it is possible to find a rigorous way for calculating the associations between words, it would be difficult with this approach to find a rigorous way for positioning words on a map.

Other text mining systems have been developed, including, for example, those designed for discovering trends in test databases (e.g., [9]), and for extracting topics from a text (e.g., [13, 14, 11]). (See also [4, 3, 7, 6]). None of them, however, can be straightforwardly applied to questionnaire data analysis.

3. FEATURES OF THE MINING SYSTEM

SA inputs questionnaire data in the CSV format.[2] After a user designates an open question or several open questions (in a column or several columns in the CSV file), and also designates targets for analysis (in a column), **SA** automatically conducts its analyses over the designated data. In the above example, the types of cars are targets and the answers about brand images are open answers. (**SA** is currently usable only with Japanese language data, but it has been designed to be easily extendable to other languages).

SA has the following features and characteristics:

- It views analysis targets (e.g., products, companies) as categories and open answers (e.g., brand images, claims, comments) as classified texts (word sequences), and it learns classification rules (also referred to as 'stochastic decision lists') and association rules from the data. **SA** also outputs for each category a `word histogram' of that category.
- It views as random variables analysis targets and keywords extracted from open answers, and performs Correspondence Analysis on the random variables.
- It employs Stochastic Complexity [12] or Extended Stochastic Complexity [16] in rule analysis and Correspondence Analysis.
- It provides a visualization function for mining results. More specifically, it uses bar graphs to represent rules obtained and positioning maps to represent correspondence relationships.
- It provides a search function for displaying the relationships between mining results and original data.
- It has a powerful morphological analysis component and can segment a text into a sequence of both words and phrases. For example, it can automatically distinguish positive expressions and negative expressions such as `elegant' and `inelegant.' Users are also able to define their own technical terms, stop words, and synonymous words.
- It performs analysis very efficiently. For example, it takes only 10 seconds to process 1000 examples, even on a PC with Pentium 2 processor.

In the sections below, we describe the two main features of **SA**: rule analysis and Correspondence Analysis. We also discuss the relationship between them.

4. RULE ANALYSIS

4.1. Classification Rules

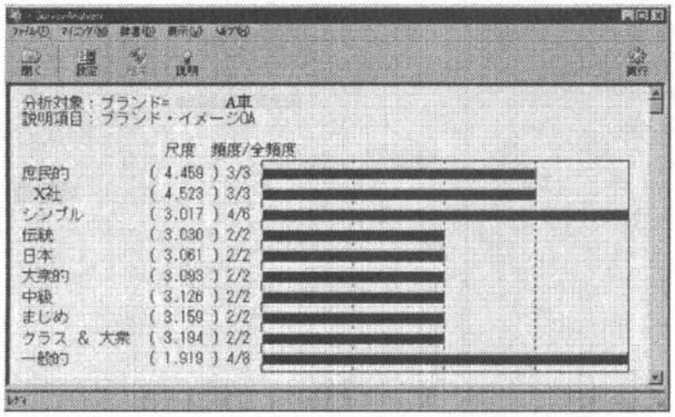
The task of rule-based text classification can be described as follows: we have a number of categories, each already containing a number of texts, and we are to automatically acquire some sort of rules from the categorized texts and classify new texts on the basis of the acquired rules. **SA** views each analysis target as a category and views open answers associated with the target as texts. For each target, it learns rules for assigning open answers. It views the obtained classification rules as mining results that represent the characteristics of the analysis target.

Classification rules (a stochastic decision list) for an analysis target consist of an ordered sequence of **IF-THEN-ELSE rules** for assignment of open answers to the target. Each rule has a condition for its assignment. This condition requires the simultaneous presence of several words or simply the presence of

[2] CSV (comma separated values) is a file format that can be read by, for example, Microsoft Excel.

a single word. Each rule also attaches a probability (relative frequency) value to its assignment.

Figure 1 Classification Rules for Car A



```
分析対象：ブランド=        A車
説明項目：ブランド・イメージOA

                 尺度   頻度/全頻度
庶民的        ( 4.459 ) 3/3
  X社         ( 4.523 ) 3/3
シンプル      ( 3.017 ) 4/6
伝統          ( 3.030 ) 2/2
日本          ( 3.061 ) 2/2
大衆的        ( 3.093 ) 2/2
中級          ( 3.126 ) 2/2
まじめ        ( 3.159 ) 2/2
クラス & 大衆 ( 3.194 ) 2/2
一般的        ( 1.919 ) 4/8
```

Table 2 Translation of Results in Figure 1

Condition	Score	Freq./Total Freq.
'for ordinary people'	4.459	3/3
'X, LTD'	4.523	3/3
'simplicity'	3.017	4/6
'tradition'	3.030	2/2
'Japan'	3.061	2/2
'common-people'	3.093	2/2
'middle-class'	3.126	2/2
'earnest'	3.159	2/2
'class&common-people'	3.194	2/2
'general'	1.919	4/8

For example, Figure 1 shows the classification rules in Japanese for the brand images of Car A, as output by **SA**. Table 2 shows a translation of the results in Figure 1. (Input questionnaire data, shown in abbreviated form in Table 1 contained brand image answers about six types of cars: Car A, Car B, etc, and there were 120 answers for each type.) The first rule indicates that if an open answer contains the expression 'for ordinary people' then it should be classified as an answer about Car A with a probability of 3/3. If it does not but does contain the word 'X,LTD,' then it should also be classified as an answer about Car A with a probability of 3/3. The open answer is examined in this way with respect to the rules, from beginning to end.

The classification rules are constructed on the basis of the given questionnaire data. As may be seen, the first rule indicates that there are three occurrences of 'for ordinary people' among all brand image answers, and all of them appear in answers about Car A. The second rule is constructed *after removing all answers* that contain the expression in the first rule, i.e., 'for ordinary people,' and it indicates that there are three occurrences of 'X,LTD' in the remaining answers, and that all of them appear in answers about Car A.

In Figure 1 and Table 2, **frequency** denotes the number of occurrences of a word or several words in answers about the target, **total frequency** denotes the total number of occurrences among answers at large, and **score** denotes the score of the rule, as based on Stochastic Complexity. Frequencies are displayed in the bar graph in Figure 1.

4.2. Association Rules

Table 3 Association Rules for Car A

Condition	Score	Freq./Total Freq.
'for ordinary people'	4.459	3/3
'X, LTD'	4.459	3/3
'tradition'	2.888	4/6
'popularity'	2.888	2/2
'Japan'	2.888	2/2
'common people'	2.888	2/2
'middle-class'	2.888	2/2
'earnest'	2.888	2/2
'class&common-people'	2.888	2/2
'simplicity'	2.859	4/6

Association rules for an analysis target consist of an ordered sequence of **IF-THEN-OR rules** that represent the strength of associations between open answers and the target. Each rule has a condition for that association which requires the simultaneous presence of several words or simply the presence of a single word. It also attaches a probability (relative frequency) value to its association.

Table 3 shows a translation of association rules for the brand images of Car A, as output by **SA**. The first rule indicates that there are three occurrences of 'for ordinary people' in the brand image answers at large, and that all of them appear in answers about Car A. The second rule indicates that *again in the brand image answers at large*, there are three occurrences of 'X, LTD,' and all of them appear in answers about Car A.

4.3. Extracting the Characteristics of an Analysis Target

Table 4 Word Histogram for Car A

Word	Freq. / Total Freq.
'car'	17/86
'image'	10/41
'luxury'	10/56
'luxury-car'	10/60
'Japan'	5/12
'good'	4/10
'ride'	4/27
'simplicity'	4/6
'common-people'	4/8
'family'	4/10

Let us look at the mining results of **word histogram**, which consist of words listed in descending order of their occurrences in answers about a target. Table 4 shows a translation of a word histogram for the brand images of Car A, as output by **SA**. Here 'car' and 'image' are listed at the top because they frequently appear in answers about Car A. These words cannot be viewed indicative of Car A characteristics in particular, however, because they also frequently appear in answers about other cars (cf., total frequency).

We can, on the other hand, *view words that appear significantly more frequently in answers for a target as those that are indicative of the target.* As will be made clearer later, words found in both classification rules and association rules for a target turn out to be precisely those that appear significantly more frequently in answers about the target.

Table 5 Classification Rules for Car C

Condition	**Score**	**Freq./TotalFreq.**
'outdoor'	10.325	5/5
'Z, LTD'	8.132	4/4
'mobility'	6.527	8/14
'fast'	5.694	3/3
'run'	5.057	2/2
'work'	3.341	2/2
'road'	3.380	2/2
'enjoyableness'	3.420	2/2
'boring'	3.438	3/4
'sporty'	1.891	2/3

Table 5 shows a translation of classification rules for the brand images of Car C, as output by **SA** . The first rule consists of the word 'outdoor' with a relative frequency of 5/5. In contrast to this, the first rule for Car A in Table 2 consists of the word 'for ordinary people,' with a relative frequency 3/3. In other words, 'for ordinary people' is a word more frequently appearing in Car A and more indicative of Car A, while 'outdoor' a word more frequently appearing in Car C and more indicative of Car C.

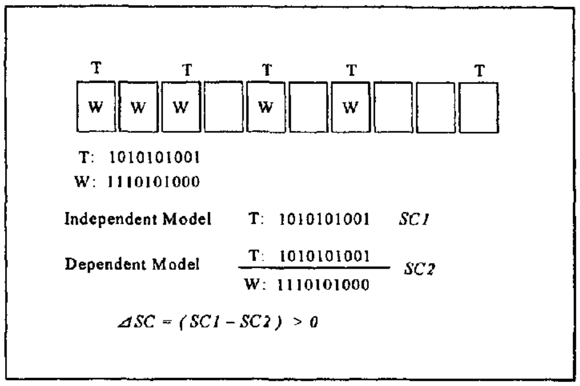
The use of classification rules is more suitable to summarization of open answers, while that of association rules more suitable to the discovering from open answers.

4.4. Algorithm

SA learns classification rules or association rules by using Stochastic Complexity (SC) or Extended Stochastic Complexity (ESC). Details of the algorithms for learning classification rules using SC and ESC can be found in [15, 10]; here we simply describe the concept behind the use of the SC.

Stochastic Complexity (SC) [12] is defined as the least code-length (also referred to as description length) required to encode the data with the help of a probability model. The **MDL** (Minimum Description Length) principle is a model selection criterion which asserts that, for a given data sequence, the lower a model's SC value is, the greater its likelihood of being a model which would actually generate the data. MDL offers many good properties as a criterion for statistical estimation.

Figure 2 Rule Selection Using SC



Suppose that there are 10 open answers (represented here by the rectangles in Figure 2). Some of them are associated with an analysis target, denoted by T, and some contain a specific word, denoted by W. The associations or non-associations with T, and similarly the presence or absence of W can be represented by a string of 1 and 0's.

We can then define two probability models: an independent model and a dependent model. The data sequence for T can be assumed to be generated by the independent model in which 1 or 0 for T is independently and identically generated by a Bernoulli model (i.e., coin flipping). The data sequence for T can also be assumed to be generated by the dependent model in which 1 or 0 for T is generated on the basis of the state of W (1 or 0). We can next calculate the SC value of the data sequence for T with respect to the independent model, and that with respect to the dependent model. If the difference between the former and the latter (denoted as ΔSC) is positive, i.e., the former is larger than the latter, then according to the MDL principle, we should view the dependent model as one that is most likely to have given rise to the data. Actually, when W appears significantly more frequently in the open answers of T, ΔSC will be significantly larger than 0, and thus the dependent model will be selected.

In the learning of classification rules, we calculate, on the basis of the entire data, the ΔSC value for each of the possible rules. Here, a rule may contain not only the presence of one word, but also the simultaneous presences of several words as its condition. We select as the first rule that for which the ΔSC value is the largest. We then remove from the data those that satisfy the condition of the first rule. For the remaining data, we again calculate the ΔSC value for each of the remaining possible rules, and select as the second rule that for which the ΔSC value is the largest. We repeat this process until we cannot find any rule which is significant in terms of ΔSC .

In the learning of association rules, we calculate, on the basis of the entire data, the ΔSC value for each of the possible rules, and sort the rules in descending order of their ΔSC values.

It turns out that those words which appear in both classification rules and association rules for an analysis target are exactly those that appear significantly more frequently in answers about the target (cf., Tables 2 and 3).

5. CORRESPONDNCE ANALYSIS

5.1. Positioning Map

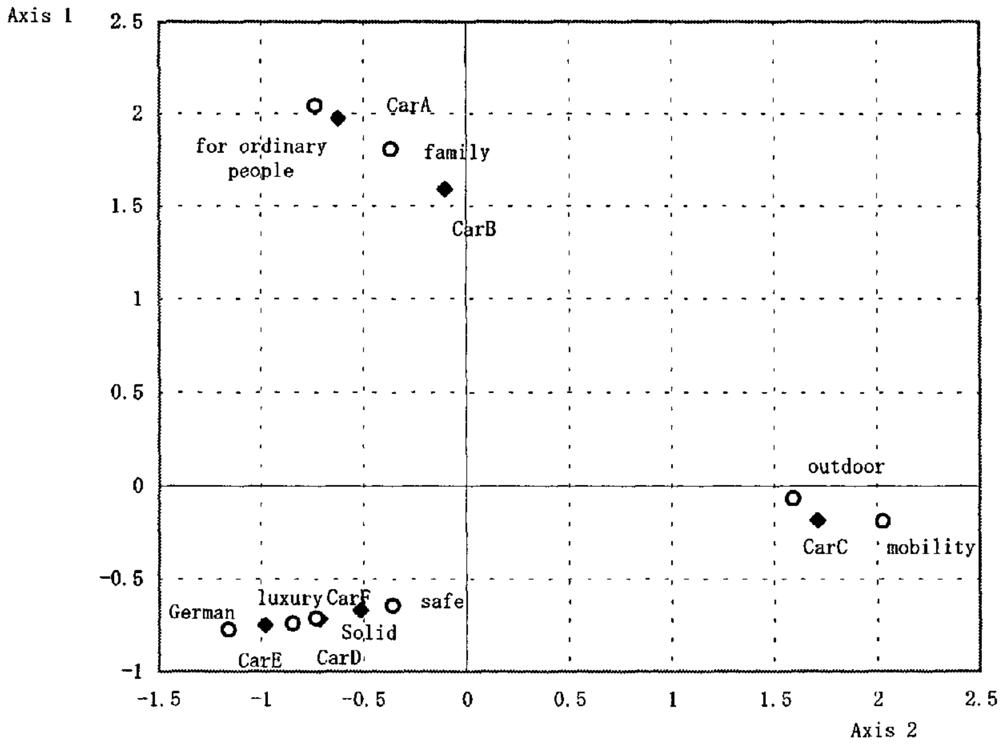
After a user designates analysis targets and open answers, **SA** can perform Correspondence Analysis and output a two-dimensional positioning map of the analysis targets and keywords extracted from the open answers. The map visually shows the relationships between the targets and the keywords, with distance on the map being a representation of correspondence (closeness).

Figure 3 shows the positioning map for the car data in Table 1. The car types form three groups: (I) Car A and Car B; (II) Car C; (III) Car D, Car E, and Car F. Group I is characterized by the words `for ordinary people' and `family,' Group II by the words `mobility' and `outdoor,' and Group III by the words `luxury,' `safe,' `solid,' and `German.' In this way, the positioning map clearly indicates the relationships between the image characteristics of the cars.

Table 6 Co-occurrence Data for Car Types and Keywords

	ordinary people	family	Outdoor	mobility	luxury	solid	safe	German
Car A	3	4	0	0	0	0	0	0
Car B	0	6	1	0	0	0	0	0
Car C	0	0	9	4	0	0	2	0
Car D	0	0	1	0	0	9	8	0
Car E	0	0	0	0	3	2	4	0
Car F	0	0	0	0	0	2	1	2

Figure 3. Positioning Map of Car Types and Keyword



5.2. Algorithm

Before performing Correspondence Analysis, **SA** extracts keywords from open answers designated by the user. Specifically it uses Stochastic Complexity to extract keywords that are indicative of individual analysis targets. The extracted keywords are in fact equivalent to those existing on the top of the association rules for the targets. Table 6 shows some example keywords extracted for the images of car types. **SA** then constructs a table like Table 6, which contains co-occurrence data between the targets and extracted keywords.

SA next views the targets and extracted keywords as random variables and performs Correspondence Analysis on those random variables. Such analysis is a rigorous way of positioning words on a two-dimensional or three-dimensional map. (The positioning map in Figure 3 was obtained on the basis of the co-occurrence data in Table 6). Detail regarding Correspondence Analysis can be found in [2]. Here, we simply describe the concept behind it.

Correspondence Analysis can be viewed as an extension of Principal Component Analysis (PCA) (which is similar to Singular Value Decomposition). Suppose that the co-occurrence data is represented as a matrix X, which has m rows and n columns. We can view the data as m n-dimensional vectors. The relationship between rows (in the above example, car types) thus can be represented by the distance between the corresponding vectors. Using PCA, we can compress the data. More precisely, we can approximately represent the m vectors with only two dimensions so that the amount of information (rigorously the variances of the random variables corresponding to the two dimesions) can be preserved as much as possible. Similarly, we can view the original data as n m-dimensional vectors and employ PCA to compress data. There is, however, no guarantee that the amounts of information preserved in the two cases are the same. Correspondence Analysis resolves this problem by conducting certain transformations on the original data. In this way, it can preserve the same amount of information for both columns and rows, and thus is able to represent the relationship between both rows and columns (car types and keywords) in a single space (map).

5.3. Relationship between Rule Analysis and Correspondence Analysis

The performing of both rule analysis and Correspondence Analysis is essential for mining the characteristics of analysis targets in open answers, as the former is designed to extract the characteristics of individual analysis targets (Function A), and the latter to extract the relationships among the targets (Function B). Rule analysis employs a conditional probability model represented as $P(Y \mid X)$, and Correspondence Analysis employs a joint probability model represented as $P(Y,X)$, where Y denotes analysis targets, and X words. Correspondence Analysis yields the *entire structure* (cf., Figure 3), while rule analysis provides the *facts in detail* (cf., Tables 2-5). Both are indispensable for the mining of open questions.

6. MINING RESULTS

Our survey analysis system, released as a product of NEC, has already been used by a number of large corporations in Japan to perform text mining on various types of survey data including open answers about brand images, open answers about company images, complaints about products, comments written on home pages, business reports, and help desk records. The confidential nature of the surveys, however, prevents us here from describing more than the outlines of some of the results.

6.1. Mining Results with Car Data

Hakuhodo[3], for example, the second largest advertising agency in Japan, has used **SA** to analyze questionnaire data, and have provided us with the questionnaire data on cars presented in Table 1. The mining results presented in Tables 2-5 and Figure 3 represent a part of the results of analyses we performed.

Table 7 Evaluation Results

	Association Rule		**Classification Rule**	
	Recall	**Precision**	**Recall**	**Precision**
Car A	0.80	0.90	0.70	0.80
Car B	0.90	0.90	0.90	0.90
Car C	0.90	0.90	0.80	0.80
Car D	0.50	1.00	0.40	1.00
Car E	0.80	0.80	0.70	1.00
Car F	0.60	0.90	0.60	1.00
Average	0.75	0.90	0.68	0.92

In a quantitative evaluation of the mining results, we extracted, on the basis of intuition, 10 keywords from image answers for each type of car and tested to what extent those keywords agreed with the words appearing in the first 10 classification and association rules output by **SA**. We evaluated these results in terms of precision and recall. Here, precision is defined as the ratio of the number of words correctly extracted to the total number of words extracted. Recall is defined as the ratio of the number of words correctly extracted to the total number of words that should have been extracted. Table 7, which shows the evaluation results in terms of precision and recall, indicates that **SA** is accurate in its extraction of keywords. (Although the evaluation cannot be viewed as completely objective, we can still observe certain tendencies in it.) Errors in classification and association rules were mainly due to the difficulty of dealing with varieties of linguistic expressions.

[3] http://www.hakuhodo.co.jp/

6.2. Mining Results with Eye-drop Data

Biglobe[4], the second largest Internet service provider in Japan, has used **SA** to analyze some of their survey data on eye-drops. In the survey, they asked users of eye-drops about their complaints. Viewing each of the eye-drop types as a category, they then conducted rule analysis. They discovered, for example, that the word 'yellow' appeared in classification rules for one type of eye-drop, i.e., the word 'yellow' appeared significantly often in complaints about that medicine. This motivates the search function of **SA** to identify complaints containing the word 'yellow,' such as "The yellow liquid stained my white shirt," "It is terrible if you spill the yellow eye drops." That is to say, **SA** has here successfully mined useful information from questionnaire data.

6.3. Mining Results with Beverage Data

Dentsu[5], the largest advertising agency in Japan, has also applied **SA** to survey results data. In one case, for example, they asked people to indicate whether they are heavy-consumers, moderate, or non-consumers of a certain beverage, and then asked heavy and moderate-consumers to describe the occasions on which they consumed it. Viewing heavy-consumer and moderate-consumer as categories, they conducted rule analysis, which revealed a number of unanticipated consumption patterns among heavy-consumers which were considered highly significant. For example, the word 'furoagari (after a bath)' appeared in classification rules, i.e., the word appeared significantly in the answers of heavy-users, which was an unanticipated pattern.

7. ADVANTAGES OF THE MINING SYSTEM

SA provides a *new* way of mining from open answers in questionnaire data., and it offers many advantages.

- The use of **SA** has an advantage over analysis by humans, because it eliminates the bias that can result from overly intuitive human analysis and also because it is a much faster and less costly way to summarize or mine from open questions.
- To the best of our knowledge, **SA** is the first system that can perform both Function A and Function B (described above) for mining from open questions. It does so by performing rule analysis and Correspondence Analysis.
- **SA** is also novel in its employment of a new statistical learning methodology based on Stochastic Complexity. It is thus able to provide *reliable* mining.
- **SA** has successfully been used in the mining of various types of questionnaire data and found to be useful in forming a basis for effective business decisions.

8. ACKNOWLEDGMENTS

We are very grateful to Hakuhodo for their providing of car data, and to Biglobe and Dentsu for their permission to describe results of their mining in this paper.

9. REFERENCES

[1] M.R. Anderberg, Cluster Analysis for Applications, Academic Press, 1973.

[2] J.P. Benzecri, Correspondence Analysis Handbook. Mercel Dekker, 1992.

[3] Jochen Dorre and Peter Gerstl and Roland Seiffert, Text mining: finding nuggets in mountains of textual data, Proceedings of the 5th International Conference on Knowledge Discovery and Data Mining, 398-401, 1999.

[4] Ronen Feldman and Ido Dagan, Knowledge discovery in textual databases (KDT), Proceedings of First International Conference on Knowledge Discovery and Data Mining, 1995.

[5] Fujitsu, Symfoware World http://www.fujitsu.co.jp/jp/soft/symfoware/index.html, 2001.

[6] Marko Grobelnik, Dunja Mladenic, and Natasa Milic-Fraling (Ed.) Proceedings of KDD-2000 Workshop on Text Mining, 2000.

[7] Marti Hearst, Untangling text data mining, Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics, 3-10, 1999.

[8] Komatsu Soft, Information Mining Tool VextSearch (in Japanese) http://www.komatsusoft.co.jp/develp/vxtsc/index.html, 2001.

[9] Brian Lent and Rakesh Agrawal and Ramakrishnan Srikant, Discovering trends in text databases, Proceedings of the Third International Conference on Knowledge Discovery and Data Mining, 227-230, 1997.

[10] Hang Li and Kenji Yamanishi, Text classification using ESC-based stochastic decision lists, Proceedings of the 8th International Conference on Information and Knowledge Management, 122-130, 1999.

[11] Hang Li and Kenji Yamanishi, Topic analysis using a finite mixture model, Proceedings of 2000 Joint ACL-SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora, 35-44, 2000.

[12] Jorma Rissanen, Fisher information and stochastic complexity, IEEE Transaction on Information Theory, 42(1):40-47, 1996.

[13] Russell Swan and James Allan, Extracting significant time varying features from text, Proceedings of the 8th International Conference on Information and Knowledge Management, 45, 1999.

[14] Mark Shewhart and Mark Wasson, Monitoring a newsfeed for hot topics, Proceedings of the 5th International Conference on Knowledge Discovery and Data Mining, 402-404, 1999.

[15] Kenji Yamanishi, A learning criterion for stochastic rules, Machine Learning, 9:165-203, 1992.

[16] Kenji Yamanishi, A decision-theoretic extension of stochastic complexity and its applications to learning, IEEE Transaction on InfortmationTheory.,44(4):1424-1439,1998.

[4] http://www.biglobe.ne.jp/

[5] http://www.dentsu.co.jp