

# Text analysis and knowledge mining system

---

by T. Nasukawa  
T. Nagano

*Large text databases potentially contain a great wealth of knowledge. However, text represents factual information (and information about the author's communicative intentions) in a complex, rich, and opaque manner. Consequently, unlike numerical and fixed field data, it cannot be analyzed by standard statistical data mining methods. Relying on human analysis results in either huge workloads or the analysis of only a tiny fraction of the database. We are working on text mining technology to extract knowledge from very large amounts of textual data. Unlike information retrieval technology that allows a user to select documents that meet the user's requirements and interests, or document clustering technology that organizes documents, we focus on finding valuable patterns and rules in text that indicate trends and significant features about specific topics. By applying our prototype system named TAKMI (Text Analysis and Knowledge Mining) to textual databases in PC help centers, we can automatically detect product failures; determine issues that have led to rapid increases in the number of calls and their underlying reasons; and analyze help center productivity and changes in customers' behavior involving a particular product, without reading any of the text. We have verified that our framework is also effective for other data such as patent documents.*

Since a textual format is a very flexible way to describe and store various types of information, large amounts of information are stored and distributed as text. Moreover, the amount of accessible textual data has been increasing rapidly. Such data may potentially contain a great wealth of knowledge. However, analyzing huge amounts of textual data requires a tremendous amount of work in reading all of the text and organizing the content. Thus, the in-

crease in accessible textual data has caused an information flood in spite of hope of becoming knowledgeable about various topics.

In order to overcome this counterintuitive situation, a text mining technology called TAKMI (Text Analysis and Knowledge Mining) has been developed to acquire useful knowledge from large amounts of textual data such as internal reports, various technical documents, messages from various individuals, and so on. It is described in this paper. In particular, we try to analyze what a large set of documents indicates as a whole rather than focus on the specific information in each document.

The most important issue for this text mining technology is how to represent the contents of textual data in order to apply statistical analysis. Since any information that cannot be retained within the representation cannot appear in the final output, we should carefully examine the expressive power of the representation system. Then, once the appropriate information is extracted from the text, the next issue is how to provide statistical analysis. We should then apply appropriate mining functions adapted to the representations of the original content of the text. Finally, since the content of the text varies greatly, it is essential to visualize the results and allow an interactive analysis to meet the requirements of analysts working from multiple points of view.

©Copyright 2001 by International Business Machines Corporation. Copying in printed form for private use is permitted without payment of royalty provided that (1) each reproduction is done without alteration and (2) the *Journal* reference and IBM copyright notice are included on the first page. The title and abstract, but no other portions, of this paper may be copied or distributed royalty free without further permission by computer-based and other information-service systems. Permission to *republish* any other portion of this paper must be obtained from the Editor.

Table 1 Comparison of document-handling technologies

Function	Purpose	Technology	Data Representation	Natural Language Processing	Output
Search documents	Focus on data related to some specific topics	Information retrieval	Character strings, keywords	Keyword extraction (conversion to base forms)	A set of documents
Organize documents	Overview of topics	Clustering, <sup>1</sup> classification <sup>2</sup>	Set of keywords (Vector Space Model) <sup>3</sup>	Analysis of keyword distribution	Sets (clusters) of documents
Discover knowledge	Extract interesting information from content	NLP, data mining, visualization	Semantic concepts	Semantic analysis, intention analysis	Digested information (trend patterns, association rules, etc.)

In the next section of this paper, we contrast our text mining technology against other document-handling technologies, such as document search and document classification, as well as other work in text mining. In the third section, we describe the technical details of our technology, especially focusing on natural language processing to extract concepts from text, and we also describe our data mining technology to find interesting patterns and features, as well as our visualization technology that allows interactive analysis. The fourth section is devoted to the successful application of our text mining technology to records in customer contact centers such as PC help centers. The expressive power of the representation system for concept extraction is evaluated by using the customer contact records in this section. Applications for other data such as patent documents are discussed in the fifth section, followed by concluding remarks.

### Text mining technology

We first describe various document-handling technologies and related work in text mining technology.

**Document-handling technologies.** In order to reduce the work in handling huge amounts of textual data, various technologies have been developed, and the features of some document-handling technologies are classified in Table 1.<sup>1-3</sup>

Information retrieval technology is probably the most common technology to use when we are faced with a very large number of documents. The term “text mining” (or “text data mining”) is sometimes used to indicate this technology because it detects and extracts documents that we want from mountains of documents, and it allows us to select data related to

some specific topics that we are interested in so that the amount of data we have to handle is reduced without losing the information we want. However, we have to specify what we want in the form of a query in order to use this technology. Thus, the use of this technology is limited when we do not have a clear intention about what to search for and knowledge of what can be retrieved from the database we are searching. Moreover, even when we have some specific topics to search for and successfully make some queries, the output we obtain is a list of documents that we still have to read to find the information, unless we are simply interested in such data as the number of documents that contain specific keywords or character strings.

Document organization technology can give us an overview of a document archive, either by classifying documents into predefined classes or by clustering documents with similar contents, where similarity is defined by the system. By using this technology, we can find out what kind of topics are contained in the document archive and the sizes of the classes or clusters for each topic. The term “text mining” is also used to indicate this technology, because it may break up and organize mountains of documents. Yet we still have to read each document in the class or the cluster that we want to analyze for details.

In contrast, “text mining” in the sense of knowledge discovery can be characterized as a novel approach to the data overload problem that the search and organization paradigms attempt to alleviate as in Hearst.<sup>4</sup> In our work, we are aiming at a technology to analyze more detailed information in the content of each document and to extract interesting information that can be provided only by multiple documents viewed as a whole, such as trends and sig-

nificant features that may be a trigger to useful actions and decision-making. For example, in a large number of documents related to customers' calls (as shown in later examples) we want to find what kinds of topics have recently been increasingly mentioned and which product is associated with specific topics, so that we can take appropriate actions for the improvement of call center productivity and product quality.

**Related work in text mining technology.** In terms of knowledge extraction, many kinds of knowledge can be extracted from textual data, such as linguistic knowledge for Natural Language Processing (NLP)<sup>5</sup> and domain-specific lexical and semantic information that may be stored in a database.<sup>6</sup> The technology of Information Extraction<sup>7</sup> is also related in terms of extracting meaningful items from textual data. However, Information Extraction, typically focused in Message Understanding Conferences (MUCs),<sup>8</sup> is intended to find a specified class of events, such as company mergers, and to fill in a template for each instance of such an event. Thus, this technology is almost the inverse of our text mining that aims to find novel patterns rather than predefined patterns in a specified class. In our research, we focus on extracting information that may be used to discover trends in the domain described in the textual database. Consequently, we are not concerned with providing analyses of documents but rather of a subset of the textual database viewed as a whole. At the same time, we focus on technologies that may also be used to discover relationships between separate categories. Thus, our "text mining" is a text version of generalized data mining, and it consists of NLP to extract concepts from each piece of text, statistical analysis to find interesting patterns among the concepts, and visualization to allow interactive analysis.

In spite of strong demand from individuals who deal with large amounts of text, there has been little practical work in applying data mining techniques to text. Among the small number of publications on using data mining technology for text, Feldman and Dagan<sup>9</sup> presented a framework for finding interesting patterns in the distributions of concepts in documents, and Feldman et al.<sup>10</sup> introduced a method for visualizing associations of concepts in a certain context. Lent et al.<sup>11</sup> presented a technique for finding trends in the use of words and phrases in documents. However, the output of these previous methods is limited by the shallow NLP. Each document is represented by a simple character string or a set

of keywords as in Mladenic.<sup>12</sup> When keywords are extracted from each document, linguistic knowledge may be applied to convert conjugated forms and plural forms into canonical forms and singular forms so that the number of variants of the keywords is reduced. Synonyms may also be unified into a canonical expression. However, a lot of information is conveyed with functional words, such as negations. Such words are usually treated as stop words, which are ignored. Also, the relationships among words are lost. Moreover, these technologies do not discriminate between information conveyed with higher-level sentential structures, such as the interrogative mood and the imperative mood. Such information is crucial to recognize communicative intentions such as questions, requests, complaints, and commendations, which is especially important to make a useful analysis of data related to customer relationship management (CRM) as described in a later section. Thus, we developed a framework for extracting concepts that consist of predicate argument pairs in association with information on modality, which often indicates intention and polarity, either negative or positive, which is also an important factor in analyzing intentions.<sup>13</sup>

## Framework of text mining

In this section, we discuss issues of concern for obtaining useful results in text mining, followed by our approach to the development of a practical text mining system, since these issues affect each of the following three components that constitute the framework.

1. Concept extraction based on robust natural language processing
2. Data mining for discovering rules and patterns
3. Visualization and interactive analysis

**Concept extraction for text mining.** In this paper, we use the term "concept" as a representation of the textual content in order to distinguish it from a simple keyword with the surface expression.

*Issues in representation of textual contents.* The first problem is that, because of the ambiguities in natural language, the same keyword may express entirely different meanings. For example, the word "Washington" may represent a person, place, or something else. The meaning of such polysemous words is normally determined according to their context.

The inverse problem is that different expressions may refer to the same meaning, for instance, “car” and “automobile” or “H/W” and “hardware.” Even when the meaning may not be exactly the same, it may be necessary to treat these expressions as denoting the same meaning for text mining, especially when some of the synonyms are used infrequently, in order to avoid data sparseness, since a small number of appearances compared to others tend to be ignored in the final output.

Thus, the surface expression of keywords is not a proper representation for text mining. And the representation of context should have the capability of reflecting the results of semantic disambiguation based on context analysis in order to deal with the polysemy and the synonymy.

Moreover, insertion of a single word such as “not” or alteration of word position may change the entire meaning of a sentence as in the following examples.

- (a) X did fail.
- (b) X did not fail.
- (c) Did X fail?

However, in the bag-of-words approach, these sentences are treated as the same content since they share the same set of keywords. Besides the negation and interrogative mood, some auxiliary verbs such as “can” and some verbs such as “want” often indicate the author’s communicative intentions.

Furthermore, the relationships among words in a document are only defined as co-occurrences in the bag-of-words approach, which, for example, may lead us to a misinterpretation of a relationship between “B” and “good” in the following sentence.

A is good, but B is bad.

According to Matsuzawa,<sup>14</sup> only 40 percent of the predicate and noun pairs within the same sentence have grammatical dependencies. Since such grammatical dependencies may convey useful meaning such as a description of the actions of the subject, the representation of the dependencies is important. In addition, the role of modifree or modifier, such as being subject or object, is also important, as in the following example.

- (a) A deleted B.
- (b) B deleted A.

*Concept extraction in TAKMI.* In order to deal with the above problems, we took the following approach to represent the content in textual data for our prototype text mining system named TAKMI.

- Assign semantic features to words and phrases expressed in the text and unify synonyms into a canonical form by referring to a semantic dictionary, in order to categorize them semantically so that they are treated as unique concepts.

Washington [person]  
 Washington [place]  
 fail [complaint]<sup>15</sup>

- Associate communicative intentions with predicates by analyzing grammatical features and lexical information so that “fail” in the previous example (a) to (c) may be represented as follows:

- (a) X did fail. → fail[complaint]
- (b) X did not fail. → not fail[commendation]
- (c) Did X fail? → fail[question]

instead of

X and fail

in the bag-of-words representation.

- Extract dependency pairs of words and phrases so that the content of the previous examples may be represented as follows:

Program A is good, but Program B is bad.  
 → Program A[software]... good[commendation],  
 Program B[software]... bad[complaint]

Program A deleted Program B.  
 → Program A[software]... delete[complaint],  
 delete[complaint]... Program B[software]

In order to generate these representations, in TAKMI we use the steps described in the following subsections.

*Creation of semantic dictionary.* Assuming that each domain has important terms for analysis, we make a list of words extracted from the textual database sorted by their frequency and ask domain experts to assign semantic categories to words and phrases that they consider important, as well as to assign the appropriate canonical forms to take care of synonymy.

mous expressions or variations in the expressions. This dictionary consists of entries with surface representations, parts of speech (POS), canonical representations, and semantic categories such as the following example for a PC help center.

batt	noun	battery	hardware
bty	noun	battery	hardware
card driver	noun	PCMCIA driver	software

Since the word distribution is not balanced (as we will describe it in a later section), the number of frequently appearing words is relatively limited in a textual database, especially when its content belongs to a narrow domain. The workload for this dictionary creation has been relatively small in our experience (as also described in a later section).

**Intention analysis.** Intention analysis is performed by matching patterns of grammatical forms or certain expressions and by searching in a semantic dictionary in the following manner.

1. Use a POS tagger to assign a POS and a base form to each word.
2. Assign semantic features to each word and phrase by looking it up in the semantic dictionary.
3. Cluster words into verb groups and noun groups by using a shallow parser.
4. Assign intentions to verb groups by matching POS patterns and their base forms using heuristic rules such as

want to + VERB  $\rightarrow$  VERB[request]  
 please + VERB  $\rightarrow$  VERB[request]

**Dependency analysis.** Dependency analysis is basically done by grammatical analysis but with the aim of focusing on important issues concerning the text mining application. The specification of the semantic category of words, including an indication of the intention, is effective. For example, to facilitate the analysis of problems in software, extraction of dependency pairs of a predicate indicative of problems and a noun with a semantic category [software] is generally a robust method. Thus, we apply dependency analysis to the results of intention analysis. The basic procedure of the dependency analysis is to check the local grammatical dependencies among verb groups and noun groups clustered in the intention analysis, and extract predicate and argument pairs such as a subjective noun and its predicative verb.

In order to sustain the robustness of the analysis, a noun group and a verb group within a sentence without a verb group or conjunctions in between may be considered to have a dependency.

**Data mining functions for text mining.** Once appropriate concepts are extracted from each piece of text, we can apply various statistical analysis methods in data mining to the set of concepts as well as to structured data. As a result, even a simple function that examines the increase and decrease of occurrences of each concept in a certain period may allow us to analyze trends in topics. Moreover, the semantic classification of concepts enables us to analyze the content of texts from the viewpoints of various semantic categories. An example of this function is shown later in Figure 2.

The semantic classification also enables us to find significant features for some topics by analyzing associations among concepts. Although finding association rules<sup>16</sup> is one of the most attractive functions in data mining in terms of discovering novel facts, analysis of association rules in very large text samples does not provide much useful information if we simply treat each piece of text as a set of items consisting of content words within the text. Because of the large number of variations in items (words) extracted from text, the number of association rules is usually too large to verify. In addition, most of them are elements of compound words that naturally occur together. However, by using the semantic categories, we can detect the significant associations based on the assumption that concepts in the same semantic category have a similar distribution of associations in comparison to other concepts.

Table 2 is an example of the highlighting of significant associations among the concepts of [liquid]s and [problem]s in the data of a PC help center. In this table, described as a two-dimensional map in TAKMI, each cell represents an association rule with a “support” number that indicates the number of textual records that contain both of the concepts in the leftmost cell and the top cell, and a “confidence” value within parentheses that indicates the percentage of the support number divided by the number of textual records that contain the concept in the leftmost cell. Those concepts are listed in order, according to the number of texts that contain them. This table tells us that “soda” in [liquid] is strongly associated with “sticking” in [problem]. The numbers in the cell, “12 (12.63%),” mean that there were 12 calls from customers who mentioned both “sticking”

Table 2 Analysis of association among [liquid]s and [problem]s

	Damage	Fail	Sticking	Dead	Bad	Freeze
Water	94 (11.1%)	27 (3.19%)	5 (0.59%)	21 (2.48%)	17 (2.01%)	16 (1.89%)
Coffee	31 (6.87%)	12 (2.66%)	13 (2.88%)	7 (1.55%)	6 (1.33%)	5 (1.11%)
Juice	3 (2.94%)	1 (0.98%)	7 (6.86%)	4 (3.92%)	5 (4.9%)	0 (0.0%)
Soda	7 (7.37%)	2 (2.11%)	12 (12.63%)	4 (4.21%)	1 (1.05%)	1 (1.05%)
Tea	3 (7.5%)	1 (2.5%)	1 (2.5%)	0 (0.0%)	0 (0.0%)	1 (2.5%)
Beer	2 (5.88%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	4 (11.76%)

and “soda,” and those 12 calls occupy 12.63 percent of the calls mentioning “soda.” Since this percentage, 12.63, is much higher than percentages for other items in [liquid], the cell is automatically highlighted.

In addition, a “topic extraction” function based on the work of Nomiyama<sup>17</sup> allows us to spot abnormal trend situations by analyzing the number of records in certain categories compared to variations in the background level of activity. This approach analyzes changes of topics by detecting concepts whose frequency has suddenly increased during a specific period, and it arranges the concepts in the order of their frequencies at the peak and by considering the duration of the peak. Real-world examples of this function are provided later in connection with Figures 7 and 11.

**Visualization and interactive analysis for text mining.** In order to deal with requirements from various viewpoints and to compensate for the low accuracy of linguistic analysis, it is indispensable to provide functions to visualize the results for intuitive understanding and to allow interactive analysis. We are using Information Outlining technology<sup>18</sup> for this purpose. Information Outlining enables a user to obtain an overview of the features and trends of a set of data by showing the distributions of items associated with the target set of data from various viewpoints indicated by categories, and by allowing the user to select or specify concepts to narrow or broaden the target sets of data. Mining functions in the previous subsection can be invoked from this visualization component for interactive analysis.

The GUI (graphical user interface) of TAKMI is shown in Figure 1. It consists of four main frames as follows:

- Frame A shows the number of records of the current analysis associated with the search criteria to

focus on the set of texts. In this example, we are analyzing 12883 records that contain “Windows\*\*98.”

- Titles of the texts are shown in Frame B, and the actual content of each text is displayed by clicking the title such as “CDROM” or “PLANR” in this frame.
- The distribution of concepts associated with the current set of texts is shown in Frame C in accordance with the categories of the semantic features or intentions. In this frame, concepts can be sorted according to their absolute frequency, relative frequency within the data, or alphabetical order of concepts. The relative frequency is calculated using the following formula.

Relative frequency =

$$\frac{\frac{\text{(Number of records referring to the concept among the selected records)}}{\text{(Number of records selected)}}}{\frac{\text{(Number of records in the complete set of records referring to the concept)}}{\text{(Total number of records)}}}$$

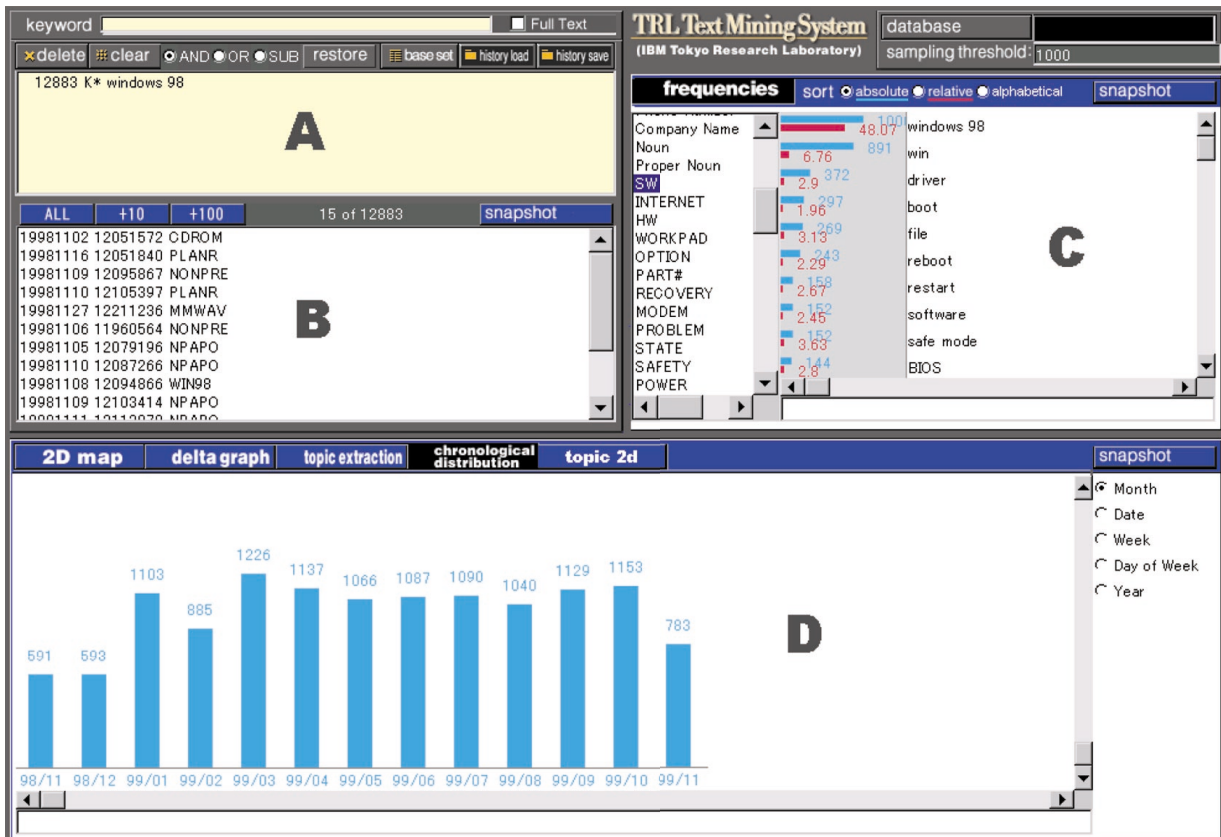
Thus, a concept with a higher relative frequency value may be strongly related to the current data set.

- The results of the mining functions are shown in Frame D.

Thus, by using TAKMI, analysts can:

1. Detect trends such as topics whose importance is increasing and find significant associations among concepts in Frame D
2. Analyze details of the trends or significant associations by browsing the relevant concepts in Frame C

Figure 1 GUI of TAKMI



3. See the basis of these results in the original texts accessed from Frame B

### Use of TAKMI to analyze records in customer contact centers

Many companies support their customers by phone, e-mail, fax, and so on. Data on customer contacts from such support centers are a good resource for obtaining knowledge about the customers, as well as knowledge about products and services that they have purchased and used.

Each contact record generally contains both structured and unstructured data. The fixed field, structured data might contain:

- Customer information (ID [identifier], name, phone number, etc.)
- Related products or services
- Contact type (question, request, query, etc.)

However, since there is a great variety of contact types and the content of each contact is complex, it is difficult to represent each interaction precisely with a limited set of predefined categories. Consequently, a free text field is used to capture details of each contact in natural language.

Although statistical analysis of the structured data provides valuable insights that help to improve business,<sup>19</sup> most of the unstructured data have been ignored because analysis by hand is prohibitively expensive. For example, one of the customer support centers that we dealt with was a PC help center that receives over 50000 calls each month. A number of analysts handle these calls in the help center, and they make reports every week by reading the customer contact records. However, they make the reports based on reading about 200 to 300 records out of tens of thousands of records. Therefore, they are using less than 5 percent of all the data they have.



Thus, analysis of reports on customer contacts was a good application for text mining, since it allows inspection of all of the data possible. The source data obviously contains useful information to detect product failures, to detect issues that have led to rapid increases in the number of customer contacts and the reasons for those increases, to analyze the productivity of help desk workers, and to analyze changes in customers' behaviors toward particular products over a period of time.

We have analyzed the records of customer contact information in a PC help center in order to apply our technology and noticed the following features:

- (a) Informal writing—The records have many indicators of informal writing:

- Various expressions are used for the same meaning, and there is a lack of consistency in expressions. Examples are:

customer = cu = cus = cust = end user =  
user = eu

Windows 95 = w95 = Win95

- Many sentences are ungrammatical.
- Many spelling mistakes are included.

- (b) Various content types—There are several important types of contents that should be recognized and handled differently in some sentences, including: requests, questions, complaints (trouble or problem), and commendations.
- (c) Multiple topics—One text may contain multiple problems and topics.
- (d) Importance levels—There are several levels of importance for the concepts. For example, [safety issues] such as smoke, spark, and injury are critical for the help center.
- (e) Multiple viewpoints—Different persons want to extract different information from different points of view. The information contained in the customers' contact information is so rich that different kinds of knowledge can be extracted from analyses with different viewpoints. For example:

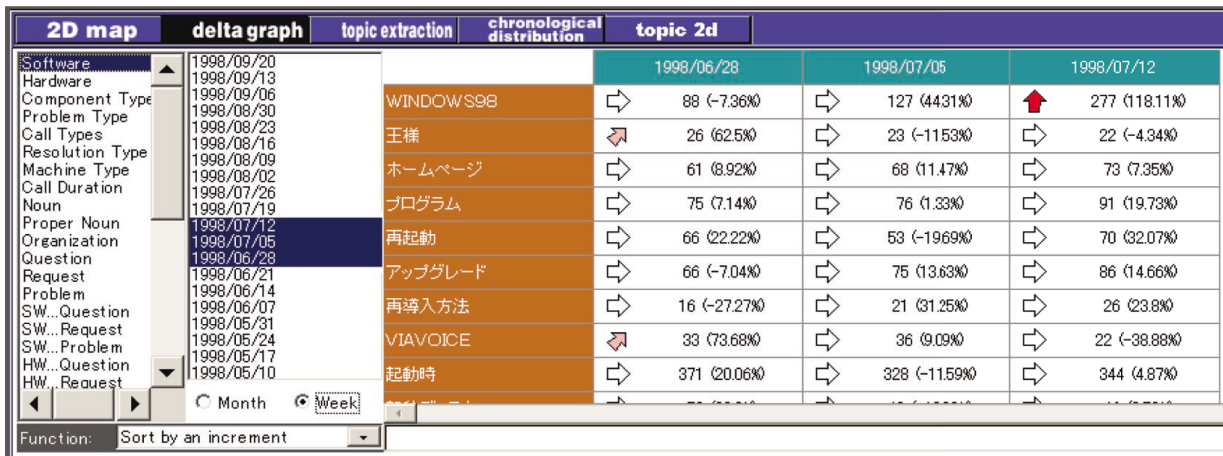
- The manager of a call center wants to improve productivity, reduce cost, improve customer satisfaction, etc. From this point of view, analysis of call-taker skills and creation of a FAQ (frequently asked question) database is important.
- An analyst for product development wants to improve product quality. Therefore, early detection of product failures is important.
- A planner for marketing may want to develop an attractive product to market, and analysis of customer buying behaviors is essential from this point of view.

- (f) Limited vocabulary—The vocabulary is fairly limited (as also described in a later section with Figure 10). For example, out of one month of data (about 40000 text entries) at a PC help center in Japan, we extracted a total of 1473444 noun phrases (including compound nouns and their component words) as keywords. However, since most of the noun phrases appeared in more than one text record, the number of different keywords was only 101987. And the number of keywords that appeared more than once was 36811, which covered 95.6 percent of the whole set of 1473444 keywords. Furthermore, for all of the keywords, 89.1 percent appeared more than nine times in the entire body of text entries. The keywords that appeared at least one thousand times covered 45 percent of the entire group of keywords.

Thus, this application is ideal for our text mining technology. First of all, the limited vocabulary means that a semantic dictionary with a relatively small number of entries, on the order of a few thousand words, may be quite effective in its coverage. This is especially true for statistical analysis that is concerned with frequently repeated events—we can ignore uncommon expressions. For instance, when we installed TAKMI at a PC help center in the United States, we made a list of words extracted from texts in the records sorted by their frequency and asked call analysts in the help center to assign semantic categories such as [software], [hardware], and [problem] to words that they consider important for call analysis as well as assigning their canonical forms to take care of synonym expressions and variations of inconsistency in expressions. As a result, it took two analysts a couple of days to make a semantic dictionary for over six thousand words, which included over 80 percent of the content words in the call records for these data.



Figure 2 Delta analysis for increasing topics in [software] from Japanese PC help center



**Results of application.** Figure 2 shows that Windows 98 was the most rapidly increasing topic in [software] from the middle of June to the beginning of July in 1998 (year, month, day) as a result of using TAKMI on the call records in a Japanese PC help center. From the monthly distribution of the number of reports that mentioned Windows 98 from January 1998 to February 1999, as shown in Figure 3, we can see that the number was rapidly increasing from July to August 1998. Figure 4 shows a list of [software . . . question] pairs in the reports that mentioned Windows 98 in July 1998. It shows the following messages:

Is it possible to install Windows 98?  
Does it support Windows 98?  
Can I use Windows 98?  
Can I upgrade?

This list tells us that most of the customers were asking whether they could install Windows 98 on their machine. In this case, the company prepared an answer to this question and put the information on their World Wide Web home page in order to reduce the number of calls from customers as well as the workload of the call takers for preparing answers.

At the same time, the effects of such actions can be examined by using our system. Figure 5 shows the monthly distribution of the number of reports on customers' calls in which customers asked whether they could install or use Windows 98 on their PCs.

Figure 6 shows a monthly distribution of 1400 calls in the same Japanese PC help center for VoiceType\*, the predecessor of ViaVoice\* (an IBM voice recognition product). Since this product was old at this time (July 1997–April 1998), the number of calls was decreasing.

The results of the topic extraction function applied to the same data for the category [Call type]<sup>20</sup> is shown in Figure 7. This figure indicates that the main focus of the calls shifts from “presale issues (guidance on purchasing)” (July–August) to “general guidance” (October–November), then to “request” (January). The system could not find any noteworthy items in the February–April period. This result seems to reflect the typical life cycle of a product and customers' behavior.

As a result, our system has been enthusiastically accepted by the help center staff since it relieves them of their routine work and improves the quality of their output. Furthermore, it made possible the use of information from large amounts of text data, not only by a limited number of analysts in the help centers, but also by nonanalysts in development and even by executives.

TAKMI has been employed in IBM PC help centers in Japan and the United States, supporting both Japanese and English. Despite the difference in languages, we could apply the same framework to both Japanese and English with a small modification in

Figure 3 Monthly distribution of calls on Windows 98 from Japanese PC help center

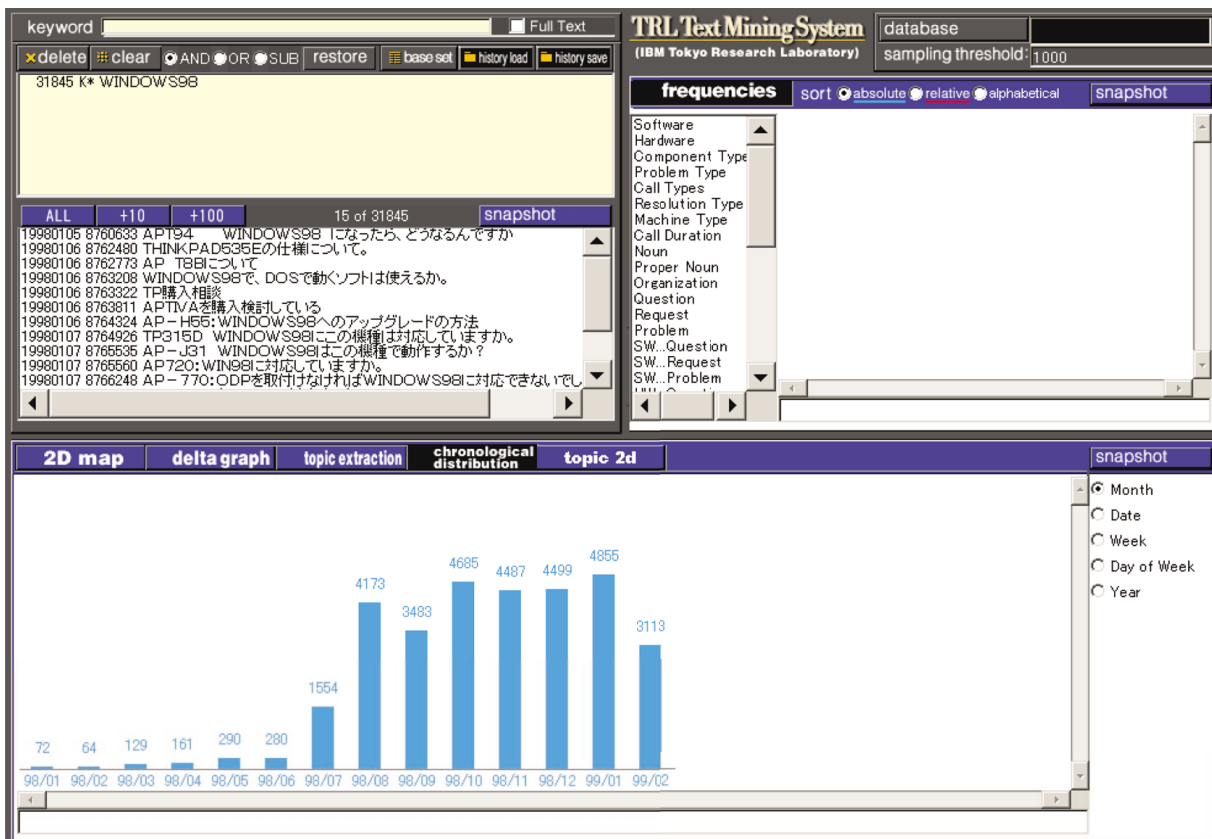
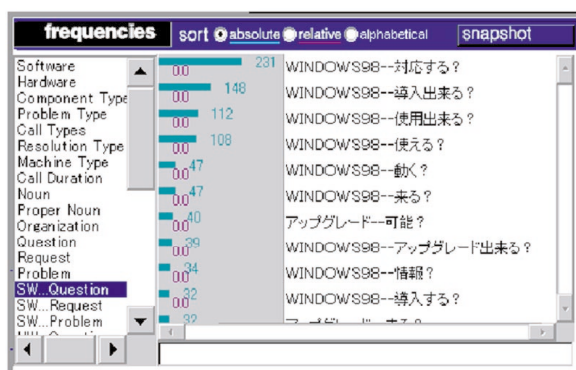


Figure 4 Frequency list of [software...question] pairs in the reports that mentioned Windows 98



the grammatical analysis for recognizing intentions and dependencies. In these help centers, TAKMI has

been evaluated as very effective for problem detection as well as for reducing the workload of analysts.

The two-dimensional association analysis proved quite effective for finding problems in products. For example, by checking associations among [machine type] and [problem] in one month of data from the Japanese help center, a specific machine was strongly associated with the concept “slow,” and in the data associated with the machine type and “slow,” “hard disk” had a high relative frequency value. As a result, the development team found a shielding problem in the hard disk of the specific type of machine that made them operate slowly.

**Evaluation of concept extraction in TAKMI.** We have evaluated the expressive power of the results of concept extraction by using the customer contact records.

Figure 5 Number of calls on the possibility of installing Windows 98

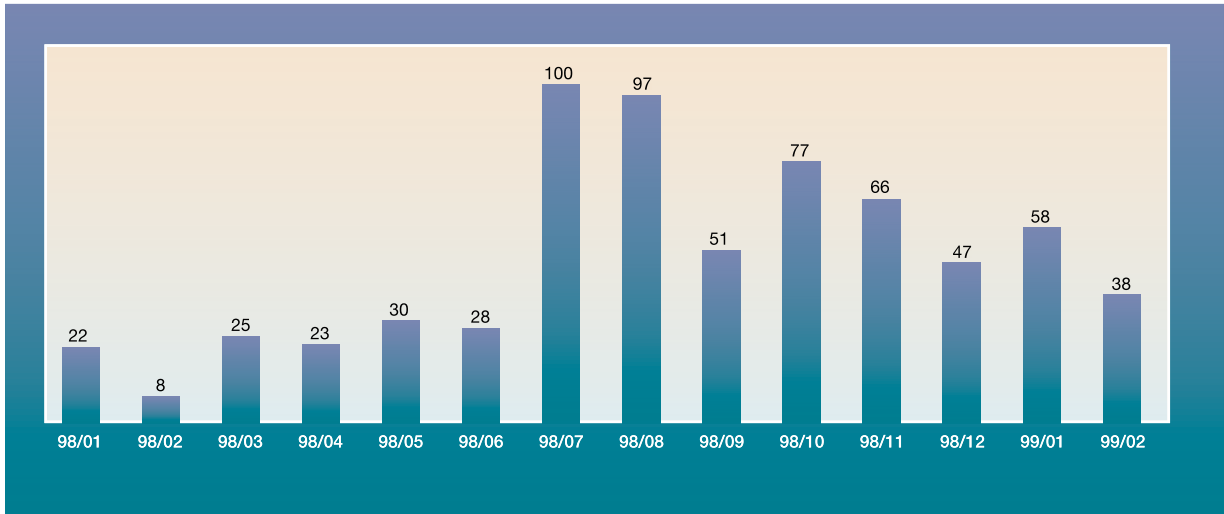
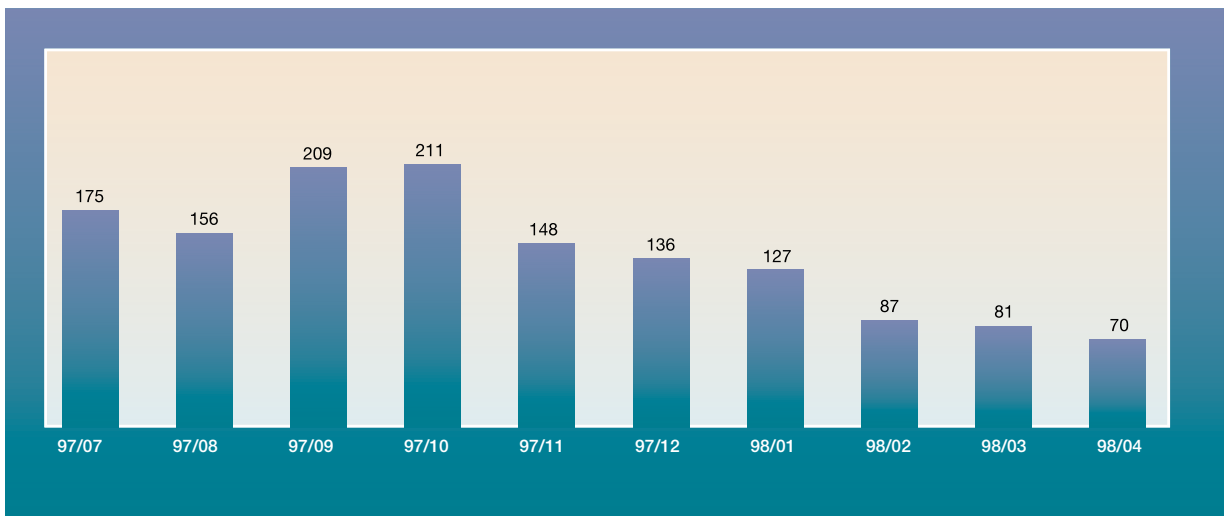


Figure 6 Monthly distribution of calls on VoiceType



First, in order to investigate the practical needs of intention analysis, we analyzed the use of verbs in call records. As shown in Table 3, the verb “use”<sup>21</sup> appeared with expressions that indicate some intentions in over 40 percent of the cases in the call records in the Japanese PC help center. Thus, disregarding this type of information may lead to a relatively high possibility of misinterpreting the textual content.

The complex concept representation based on intention analysis and dependency analysis was quite effective for representing sentential level information described in various expressions. For example, in one month of data from the Japanese PC help center, 55 cases contained “file . . . not found” in the [software . . . problem] category. Among the 55 cases, only 13 cases contained the same surface expression to describe that concept. The complex

Figure 7 Result of topic extraction from calls on VoiceType from Japanese PC help center

	1997/7	1997/8	1997/9	1997/10	1997/11	1997/12	1998/1
TC:購入相談	PRESALE						
TC:総合案内				GENERAL	GUIDANCE		
TC:要望							REQUEST

Table 3 Appearances of “use” in call records with their context of intention (originally in Japanese)

Typical Expression	Indication of Intention				Number of Appearances
	Possible	Negation	Request	Question	
Use	N	N	N	N	1998 (56.2%)
Not possible to use	Y	Y	N	N	637 (17.9%)
Possible to use	Y	N	N	N	297 (8.4%)
Want to use	N	N	Y	N	262 (7.4%)
Is it possible to use . . . ?	Y	N	N	Y	137 (3.9%)
Do/does not use	N	Y	N	N	137 (3.9%)
Does it use . . . ?	N	N	N	Y	57 (1.6%)
Isn't it possible to use . . . ?	Y	Y	N	Y	19 (0.5%)
Others					10 (0.3%)
Total					3554 (100%)

concept representation has been recognized as especially effective for searching for questions to include in FAQs as a result of its capability for representing sentential level concepts.

In addition, we conducted an experiment to determine whether these complex concepts are better indicators of the categories found in the fixed fields of the call center data. One month of data from the Japanese call center was used. Since categories are assigned by call takers per record, we examined the capability of each form of document representation (noun phrase and verb phrase versus predicate-argument) to classify each document according to the manually assigned categories. We calculated the entropy of each representation for distribution over each item in a category type by using the following formula.

$$H(w_x) = \sum_{i=1}^N P(w_x|C_i) \log_2 \frac{1}{P(w_x|C_i)}$$

where  $w_x$  is an instance of a representation such as “machine” (noun) or “program . . . delete” (noun-verb).  $C_i$  is an item in a category type that has  $N$  items such as “Technical QA” and “presale issues” in [Call type].  $P(w_x|C_i)$  is a probability that  $w_x$  appears in  $C_i$ . Thus, lower entropy indicates that the distribution of the representation is a better indicator of specific categories.

Figure 8 shows average entropies for [Call type] over the words with the same frequency within all of the data. Since frequency of items is an important factor in statistical analysis in text mining, we compared the expressive power of items with the same frequency. Figure 9 shows entropies for another category, [Component type], that contain items such as “monitor,” “modem,” and “hard drive.”

In both cases, the entropy of the predicate and argument pair (indicated with “noun-verb”) is much lower than just nouns or just verb phrases. Since [Component type] basically indicates objects, the

Figure 8 Average entropies for [Call type]



difference of the entropy between nouns and predicate-argument pairs is smaller than that for [Call type]. However, since it includes some items close to [Call type] such as “installation issues” and “general guidance,” predicate-argument pairs are effective for the classification of calls.

To summarize, these results are strong evidence supporting our claim regarding the effectiveness of our representation.

### Application of TAKMI to other data

We have verified that our framework is also effective for other data such as patent documents and Medline<sup>22</sup> data.

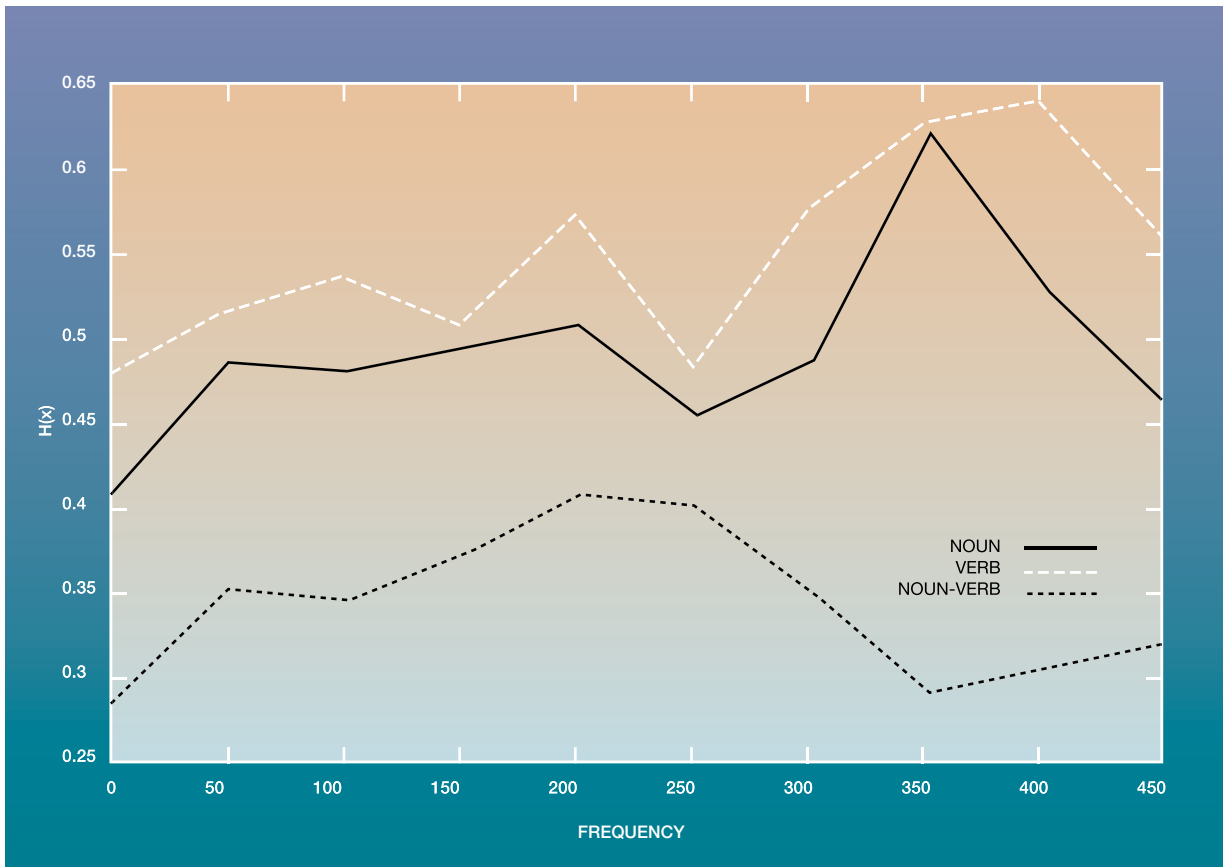
However, in patent and Medline data, the size of the vocabulary is quite large, unlike the records in con-

tact centers, unless focusing on data in a specific domain. For example, we applied a POS tagger to the following data sets:

- 31440 call records in a PC help center in the U.S. (total 2613215 words)
- 10070 abstracts of patent data (total 1267668 words)
- 15943 abstracts of Medline data (total 2600513 words)

We analyzed the distribution of words that the tagger assigned the POS “unknown” because it could not assign any other POS such as “noun” and “verb” based on grammatical information and its lexicon. Basically, most of the unknown words were technical terms and proper nouns. The unknown POS was assigned to

Figure 9 Average entropies for [Component type]



- 17.7 percent of all words in the call records (total 391013 words with 28463 distinct words)
- 2.5 percent of all words in the abstracts of patent data (total 29140 words with 6607 distinct words)
- 11.3 percent of all words in the abstracts of Medline data (total 293963 words with 38047 distinct words)

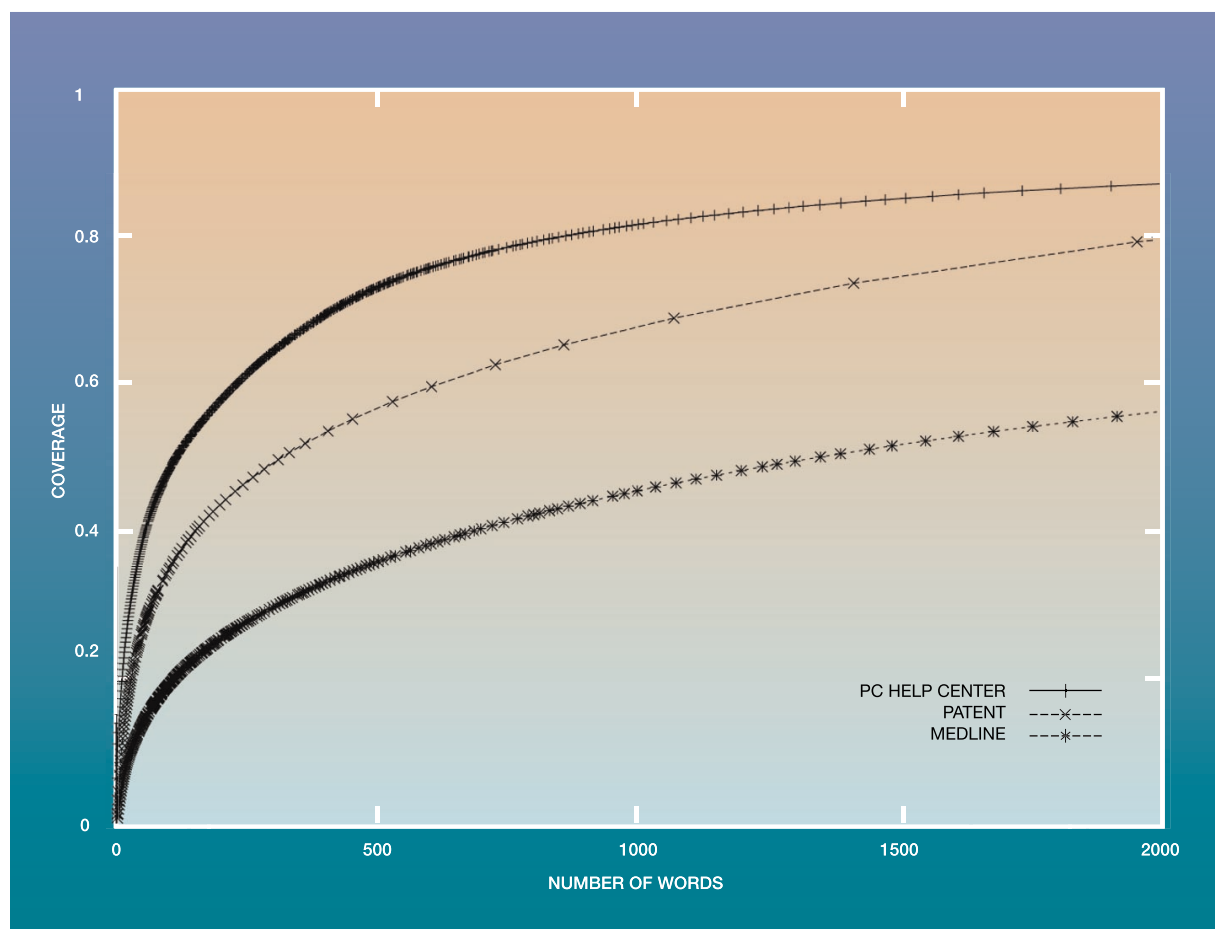
Even though the total number of unknown words is largest for the call records, those records contain only a small number of frequently appearing unknown words compared to the other two data sets. As shown in Figure 10, the most frequent 1000 words cover over 80 percent of the total word appearances of unknown words in the call records, whereas it takes more than 2000 words to cover 80 percent in the patent abstracts and 8000 words to cover 80 percent for the Medline data. This comparison shows the higher workload for creating semantic dictionaries

for patent and Medline data compared to call records.

Yet we can still apply TAKMI for these data sets without creating any new semantic dictionaries. Since the patent and Medline data have tagged information such as titles, the name of the author, and classification codes such as the IPC (International Patent Classification) code, we can use this information as well as using the grammatical category of each word assigned by a POS tagger.

**Analysis of patent documents.** Japanese patent documents are written in a markup language, and basic items such as titles, dates of issue, authors' names, and organization names are explicitly indicated by tags and can easily be extracted by applying simple pattern matching. Thus, from 15000 patent documents, we generated indexes by extracting terms in

Figure 10 Coverage of high frequent unknown words within each data set



the categories shown in Table 4 without human intervention. Among the terms for the categories shown in Table 4, the keywords in G through I were extracted by applying Japanese morphological analysis<sup>23</sup> based on Japanese grammar and lexicons in such a manner that nouns and compound words consisting of nouns were extracted as keywords.

By applying TAKMI to patent documents, we could easily draw a trend map of the patent submissions for a specific technology that shows changes for the entries of each company for the technology as well as specific features of the patent documents of each company.

*Analysis of patent strategy of a specific organization.* For example, by selecting the name of a company

Table 4 Example of categories of terms extracted from patent data

A	Date of issue
B	IPC (International Patent Classification) codes
C	Technology fields (derived from IPC)
D	Locations
E	Organization names
F	Patent holders
G	Keywords in title fields
H	Keywords in fields for the purpose of the inventions
I	Keywords in the bodies of a patent document

that is listed in the categorical viewer as category E in Table 4, a user can focus on the set of patent documents submitted by the company. Then, if the user chooses category G in Table 4 and invokes a topic



Figure 11 Topic extraction in [organization names] from 308 patent documents containing the word "inkjet"

	1991/6	1991/7	1991/8	1991/9	1991/10	1991/11	1991/12	1992/1	1992/2	1992/3
CANON										
FUJITSU										
FUJI XEROX										
ALPS										
RICOH										
SEIKO EPSON										
BROTHER IND.										
HITACHI IND.										
KAO										
FUJI FILM										
HITACHI										
NEC										
MURATA										
MATSUSHITA										
TOKYO ELECTRICITY										
SEIKOUSYA										

extraction viewer, terms that describe patented technologies are listed in accordance with the period in which the company submitted patents for each technology, showing the frequency of its submissions, since terms are listed according to their topic indexes in a topic extraction viewer, and the topic index is calculated in such a manner that a higher value is assigned to a term that appears frequently in a short period. Thus, this view reveals the patent strategy of the company.

*Analysis of competitive organizations for a specific technology.* By retrieving patent documents with a keyword that indicates a patented technology, a user can focus on a set of patents related to the technology. Then, if the user chooses category E in Table 4 and invokes a topic extraction viewer, organizations, mainly companies, that submitted patents related to the technology are listed in accordance with the period and frequency of their patent submissions. Figure 11 shows that "Canon" frequently submitted patents related to "inkjet" prior to other companies, and that other companies followed Canon's lead.<sup>24</sup> Thus, this figure shows the relative strength of companies with regard to a particular technology. By listing terms with higher relative frequency values as described earlier, we can clarify the features of each set of patented technologies as we select them.

As another example, by retrieving patent documents containing the word "inkjet" and focusing on the same set of 308 patent documents related to this topic, a user can examine technical terms typically related to "inkjet" by choosing category G in Table 4 and invoking a singularity analysis viewer. Then, by choosing the name of a company listed in a categorical viewer with category E, say, "Canon," the user can focus on a set of Canon patents related to "inkjet" technology. The user can also examine technical terms typically related to "inkjet" in patent applications submitted by Canon by invoking a singularity analysis viewer with category G in Table 4. According to our data extracted from the full set of 15000 patents, the top three terms in Canon patents are "carriage movement device," "movement device," and "jet," whereas the top three terms in Fujitsu patents related to "inkjet" are "inkjet head," "inkjet print head," and "inkjet."

Thus, we can apply TAKMI to any collection of documents even without creating a semantic dictionary. In fact, even in the PC help centers that are using TAKMI with their own semantic dictionaries, grammatical categories such as noun and proper noun (basically unknown words) are used in order to deal with unregistered words such as the name of a virus that suddenly appears. However, creation of a semantic

dictionary is effective in supporting deeper analysis. For example, the data from customer centers in broadcasting companies deal with a very wide range of topics. Nevertheless, we can register words or phrases that analysts are interested in, allowing them to ignore other words to avoid a mixture of data on irrelevant topics. In addition, application of dependency analysis to Medline data allows us to analyze interactions among objects and their effects.

### Concluding remarks

We have developed a framework of text mining to discover knowledge from very large amounts of textual data, especially focusing on NLP to extract concepts from each piece of text. Regarding technical enhancements in NLP, we believe that intention analysis and dependency analysis were the key features of our successful implementation. Intention analysis allowed us to classify predicates by analyzing functional words, which have often been ignored as stop words. Dependency analysis allowed us to capture higher-level sentential information effectively. By applying the results of concept extraction to statistical analysis functions that take advantage of semantic features of concepts, our system, TAKMI, provided practical results. Categorization of terms based on the attached semantic features is important in organizing the output knowledge as well as to facilitate interactive analysis to deal with multiple viewpoints.

Because of the interactive manner in which an analyst can easily confirm the results of analysis by checking the original documents, and the nature of the statistical analysis that tends to ignore minor patterns, occasional failures in NLP can be treated as noise and do not cause critical problems. Thus, even for the informal style of writing in the text entries of PC help center data and even for difficult grammatical analysis, TAKMI generated valuable results.

As a result, our system has been enthusiastically accepted by users who had been manually analyzing textual records, since it eases their routine work and improves the quality of their output.

### Acknowledgments

We would like to thank Jung-kook Hong, Matthew Hurst, Hiroshi Maruyama, Hirofumi Matsuzawa, Masayuki Morohashi, Akiko Murakami, Hiroshi Nomiya, Koichi Takeda, Hironori Takeuchi, the PC SW Help Center in IBM Japan, the Text Analysis

and Language Engineering Group of the IBM Thomas J. Watson Research Center, and the e-Business Technologies Group of the IBM Almaden Research Center for their collaboration and support of the TAKMI project, as well as the anonymous reviewers for their comments and suggestions. We would also like to thank Shannon Jacobs for invaluable help in proofreading an early version of this paper.

\*Trademark or registered trademark of International Business Machines Corporation.

\*\*Trademark or registered trademark of Microsoft Corporation.

### Cited references and notes

1. O. Zamir, O. Etzioni, and R. Karp, "Fast and Intuitive Clustering of Web Documents," *Proceedings of KDD-97* (1997), pp. 287–290.
2. W. Cohen and H. Hirsh, "Joins That Generalize: Text Classification Using WHIRL," *Proceedings of KDD-98* (1998), pp. 169–173.
3. G. Salton and M. J. McGill, *SMART and SIRE Experimental Retrieval Systems*, McGraw-Hill, Inc., New York (1983).
4. A. M. Hearst, "Untangling Text Data Mining," *Proceedings of ACL-99* (1999), pp. 3–10.
5. K. Night, "Mining Online Text," *Communications of the ACM* **42**, No. 11, 58–61 (1999).
6. U. Hahn and K. Schnattinger, "Deep Knowledge Discovery from Natural Language Texts," *Proceedings of KDD-97* (1997), pp. 175–178.
7. *Information Extraction*, Lecture Notes in Artificial Intelligence, M. T. Pazienza, Editor, Springer-Verlag, Rome (1997).
8. Message Understanding Conferences, see [http://www.itl.nist.gov/iad/894.02/related\\_projects/muc/index.html](http://www.itl.nist.gov/iad/894.02/related_projects/muc/index.html).
9. R. Feldman and I. Dagan, "Knowledge Discovery in Textual Databases," *Proceedings of KDD-95* (1995), pp. 112–117.
10. R. Feldman, W. Kloesgen, and A. Zilberstein, "Visualization Techniques to Explore Data Mining Results for Documents," *Proceedings of KDD-97* (1997), pp. 16–23.
11. B. Lent, R. Agrawal, and R. Srikant, "Discovering Trends in Text Databases," *Proceedings of KDD-97* (1997), pp. 227–230.
12. J. Mladenic, "Text-Learning and Related Intelligent Agents: A Survey," *IEEE Intelligent Systems* **14**, No. 4, 44–54 (1999).
13. V. Hatzivassiloglou and K. McKeown, "Predicting the Semantic Orientation of Adjectives," *Proceedings of ACL-97* (1997), pp. 174–181.
14. H. Matsuzawa and T. Fukuda, "Mining Structured Association Patterns from Databases," *Proceedings of the 4th Pacific and Asia International Conference on Knowledge Discovery and Data Mining* (2000), pp. 233–244.
15. The category of "fail" is very dependent on the domain. Thus, it should be defined in the semantic dictionary.
16. R. Agrawal, T. Imielinski, and A. Swami, "Mining Association Rules Between Sets of Items in Large Databases," *Proceedings of the ACM SIGMOD '93* (1993), pp. 207–216.
17. H. Nomiya, *Topic Analysis in Newspaper Articles*, Technical Report TR-0129, IBM Tokyo Research Laboratory, Tokyo (1996).
18. M. Morohashi, K. Takeda, H. Nomiya, and H. Maruyama, "Information Outlining—Filling the Gap Between Visualization and Navigation in Digital Libraries," *Proceedings of the International Symposium on Digital Libraries* (1995), pp. 151–158.

19. P. Xia, "Knowledge Discovery in Integrated Call Centers: A Framework for Effective Customer-Driven Marketing," *Proceedings of KDD-97* (1997), pp. 279–282.
20. This category is contained in structured data, whereas calls on VoiceType were collected based on information in unstructured text.
21. The verb "use" is "tsukau" in Japanese.
22. Information on Medline can be found at <http://www.nlm.nih.gov/>.
23. H. Maruyama, *A Formal Approach to Japanese Analysis in Japanese-to-English Machine Translation*, Dissertation, Kyoto University, Kyoto, Japan (1995).
24. This is a result of analysis in a small set of sample data to demonstrate the capability of the system.

*Accepted for publication June 26, 2001.*

**Tetsuya Nasukawa** *IBM Research Division, Tokyo Research Laboratory, 1623-14, Shimotsuruma, Yamato, Kanagawa, Japan (electronic mail: nasukawa@jp.ibm.com).* Dr. Nasukawa joined the IBM Tokyo Research Laboratory in 1989, after receiving a master's degree from Waseda University. He was involved with English-to-Japanese machine translation projects and digital library projects before he started the text mining project in 1998. He received a Ph.D. degree from Waseda University in 1998 for his work on natural language processing. His research interests include natural language understanding.

**Tohru Nagano** *IBM Research Division, Tokyo Research Laboratory, 1623-14, Shimotsuruma, Yamato, Kanagawa, Japan (electronic mail: tohru3@jp.ibm.com).* Mr. Nagano joined the IBM Tokyo Research Laboratory and text mining project in 1998, after receiving a master's degree in computer science at the University of Tsukuba in Japan. Currently he is a member of the text mining project of the Internet Technology Division, where he is developing applications for the text mining system. His research interests include natural language processing, machine learning, and statistical analysis.