Data Mining for Knowledge Organization

By Kenji YAMANISHI* and Satoshi MORINAGA*

ABSTRACT It becomes increasingly important to automatically discover business knowledge from large databases in order to drastically reduce operators' costs in the areas of CRM (Customer Relationship Management), knowledge management, Web marketing, etc. This paper introduces NEC's technology concept of Knowledge Organization and data mining engines designed for it. They include text mining tool SurveyAnalyzer, key semantics mining, and topic analysis engine TopicAnalyzer. We briefly overview the principles of these engines and illustrate their applications to real domains.

KEYWORDS **Knowledge Organization, Data mining, Text mining, Topic analysis**

1. INTRODUCTION

Data mining is a technology of extracting valuable knowledge from a large amount of data sets. For example, it can be applied to the extraction of marketing knowledge from customers' survey data. Such a technology has recently received vast attentions in a wide range of business areas including CRM (Customer Relationship Management), security, network monitoring, etc., in which it contributes to giving additional values to existing solution business or platform businesses and differentiate them from the competitors' ones.

Typical data mining methods include classification, clustering, association-rule discovery, time series analysis, graphical dependency analysis, anomaly detection, so on. The heart of all of these methods is machine learning. This is intended to automatically learn a model or patterns from data in order to understand the nature of information sources and make a reasonable decision for future data. In addition, the overall process of discovering useful knowledge from large databases must include steps; data selection, preprocessing, transformation, interpretation and evaluation. This is called KDD (Knowledge Discovery from Databases) process[1] (**Fig. 1**).

The purpose of this paper is to introduce NEC's research activities on data mining focusing on its application to business knowledge creation, which we may call Knowledge Organization. Specifically, give a brief overview of three text mining technologies: Text mining tool SurveyAnalyzer, key semantics mining, and topic analysis and illustrate how they are applied to real domains.

The rest of this paper is organized as follows: Section 2 introduces the notion of knowledge organization. Sections 3, 4 and 5 show the technologies of SurveyAnalyzer, key semantics mining, and topic analysis, respectively.

2. KNOWLEDGE ORGANIZATION

The data mining research group in NEC has been developing fundamental data mining engines and focused on two application domains: Security Intelligence and Knowledge Organization[2] (**Fig. 2**).

Security Intelligence means the technology system whose goal is to realize secure infrastructures or autonomic control systems. In it, we have developed SmartSifter (outlier detection engine), ChangeFinder (change-point detection engine), AccessTracer (anomalous behavior detection engine) and have applied them to intrusion/virus detection, network/computer failure detection, etc.

Meanwhile, Knowledge Organization means the technology system whose goal is to discover useful business knowledge mainly from text data. In it, data mining contributes to drastic reduction of operators' cost for classifying or clustering a large volume of texts data. For this purpose we have developed SurveyAnalyzer (text mining engine), TopicAnalyzer (topic trend analysis engine), key semantics extraction engine, so on and have applied them to CRM, knowledge management, and marketing.

In the scenario of Knowledge Organization, which is of our main concern in this paper, SurveyAnalyzer

*Internet Systems Research Laboratories

has been designed to conduct basic functions of text mining such as association rule discovery from static text database. It offers only at most two-word relation analysis but cannot handle complicated "contexts" which are relations among more than two words. We have recently developed the technology of key semantics mining to handle them. Further note that SurveyAnalyzer does not work in an on-line fashion. In real-time management process, which increasingly becomes important in the scenarios of BAM (Business Activity Monitoring), text streams must be processed in real-time. TopicAnalyzer has been designed to satisfy this requirement to realize real-time text-clustering and dynamic topic trend analysis.

3. TEXT MINING TOOL: SurveyAnalyzer

3.1 What is SurveyAnalyzer for?

Surveys are an important part of marketing and CRM, and open answers in particular may provide an important basis for making business decisions. The question of how to automatically mine useful information from open answers has become an important issue. Further, manual handling of it all has become costly. Computer systems that can automatically analyze open answers are sorely needed.

For example, let us consider the questionnaire data in **Table I**, referring to automobile brand images. Each row consists of a car type, which we call a category and brand image texts.

We are concerned with the following four kinds of analyses for this type of data:

1) Association-rule Induction: For a specified target category, we extract characteristic words, which are indicative of the category.
2) Correspondence Analysis: We generate a two dimensional positioning map that visually displays the correspondence relationships among target categories and their characteristic words.
3) Co-occurrence analysis: For each characteristic word, we further extract a list of words which significantly co-occur with it.
4) Typical sentence analysis: For a specified category, we assign to each text in it a score that indicates how typical that opinion is in that category and sort texts according to their scores.

We have developed a text mining tool SurveyAnalyzer (for short, SA) that conducts the four functions as above. It is released as a product of NEC in the name of TopicScope[3] and 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, complaints, business reports, and help desk records.

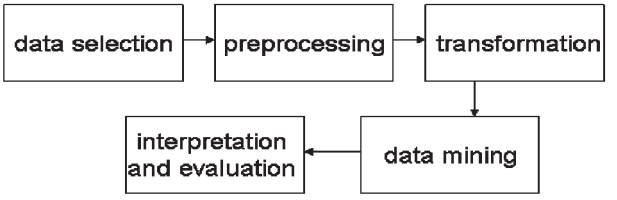


Fig. 1 KDD process[1].

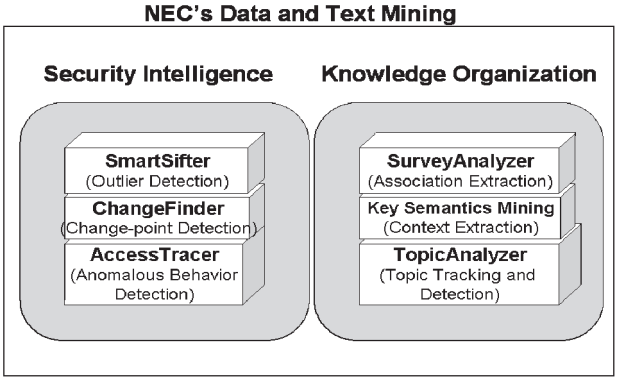


Fig. 2 NEC's data and text mining.

3.2 Association-Rule Induction

The first step is to extract keywords that appear significantly more frequently in answers for a target category than other ones. In order to extract them, we learn association rules from examples (see Reference [4]). The learned rules consist of words which have strong associations with the target category, and these words represent the characteristic of the category. Each rule is given a score measured in terms of information gain. Rules are sorted according to their information gain.

Information gain intuitively means to what extent

Table I Example questionnaire data.

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

the information is increased by dividing the original text sequences into the one which contains the keyword and that which does not. Here we further measure the information in terms of information-theoretic measure called stochastic complexity or extended stochastic complexity (see Reference [5]).

Table II shows association rules of 10 highest information gain for the brand images of Car A. Here Score shows information gain of a keyword, 'Freq.' shows its frequency in the specified category and 'Total Freq.' shows its frequency over all categories. 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 there are three occurrences of 'X, LTD,' and all of them appear in answers about Car A.

Table II 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

3.3 Correspondence Analysis

We conduct correspondence analysis in order to get a two-dimensional positioning map over the set of several categories and their keywords. The map visually shows their relationships with distance on the map being a representation of closeness. It helps the user to understand what categories are close one another and what keywords are shared in common by different categories.

The positioning map is constructed by applying the technique similar to Singular Value Decomposition into the frequency table for extracted words for each of the target categories. See Reference [4] for technical details.

Figure 3 shows the positioning map for the car data in Table I. 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.'

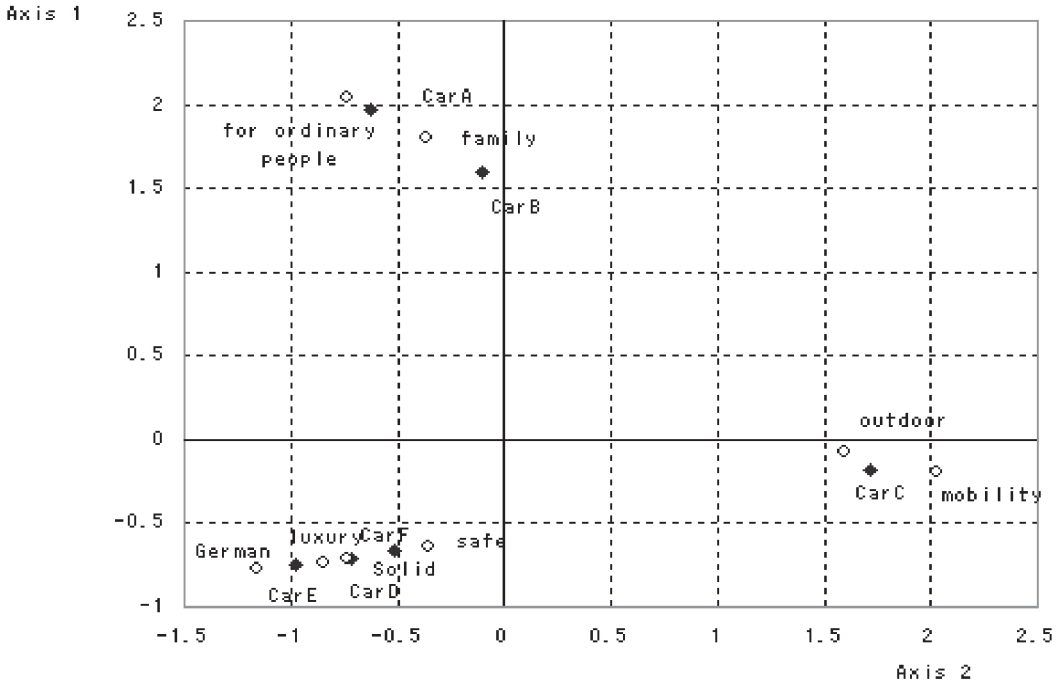


Fig. 3 Positioning map.

3.4 Co-occurrence Analysis

For each characteristic word belonging to a specific category, we extract a list of words or phrases that co-occur with that word or phrase. Through this list we are better able to understand the contexts in which the characteristic keywords appear. This analysis is often conducted for drilling down from the rule analysis or correspondence analysis.

The degree of co-occurrence is measured in terms of the information gain as used in association rule induction. **Table III** shows an example of co-occurring word list for 'popularity' and 'simplicity.'

3.5 Typical Sentence Analysis

For a set of texts belonging to a specific category, we give a score to each of them, with a high score indicating a high possibility of its being a typical opinion for the category. This gives the user a simple overview of the tendencies of that category.

A score for each sentence is measured in terms of its Bayesian posterior probability given the category. See Reference [4] for technical details.

Table IV shows a list of typical sentences in the category Car A. The characteristic words extracted above (such as "for ordinary people," "tradition," etc.) appear with high frequency in typical high scoring sentences.

4. KEY SEMANTICS MINING

In SurveyAnalyzer introduced above, keywords are extracted as the characteristics of given documents. However, a word is generally too small as a unit of extraction to catch semantic information about the target category, and a user might miss the point of characteristics. We defined key semantics as characteristic sub-structures of syntactic dependencies in the given documents, and have developed a key semantics mining system to supply more semantic information to a user[7].

Figure 4 shows examples of dependency trees and their sub-tree. (d) is a sub-tree of (a) as an ordered tree*, a sub-tree of (a) and (b) as an unordered tree†, and a sub-tree of (a) to (c) as a free tree‡, respectively. The sub-structure (d) carries much more semantic information than mere words "data" or "record" or "careful."

The key semantics mining system conducts the following three tasks:

1) Key semantics extraction: extracting characteristic syntactic dependency structures efficiently as ordered trees or unordered trees or free trees according to mining mode specified by a user,
2) Redundancy reduction: from the result of extraction, deleting redundant dependency structures such as sub-structures or equivalent structures of the others,
3) Phrase/sentence reconstruction: generating a

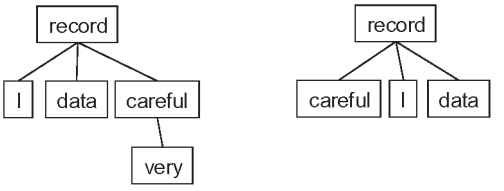
Table III Co-occurrence analysis.

Condition	Co-occurrig Word	Score	Freq.
popularity	Widely	0.14	1
popularity	Spread	0.14	1
simplicity	Image	0.14	1
simplicity	Relaxation	0.14	1

Table IV Typical sentence analysis.

Typical sentence	Score
It's for ordinary people.	2.982
I feel something traditional.	2.646
Image is 'popularity'.	2.646

(a) I recorded data very carefully. (b) Carefully, I recorded data.



(c) I have data recorded carefully. (d) ... record data carefully ...

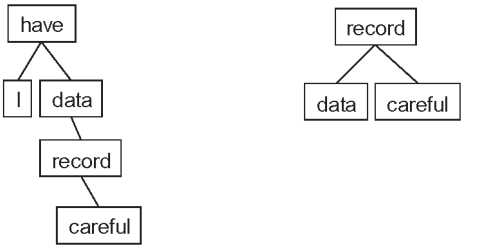


Fig. 4 Syntactic dependency trees and a sub-tree.

*A tree with sibling order relations
†A tree without sibling order relations
‡A graph with undirected edges

phrase or sentence in a natural language corresponding to the extracted structure.

The system consists of the following five units: syntactic dependency analysis unit, input filters, characteristic subtree extraction unit, output filters, and phrase/sentence reconstruction unit. The processing procedure is 1) Dependency analysis of the given documents, 2) Translation of the dependency trees by the input filters, 3) Extraction of characteristic ordered subtrees from the translated trees, 4) Redundancy reduction of the extraction result by the output filters, and 5) Reconstruction of phrases or sentences from the extracted results in a natural language.

As for the task 1, we introduce several input filters which translates syntactic dependency trees into some normalized trees, in order to divert an existing efficient algorithm[8] for ordered-subtree enumeration into unordered and free tree mining. In our system, a syntactic dependency tree is constructed for each sentence by morphological and syntactic analysis[9] at first. And after the trees are translated by the filter(s), characteristic subtrees are extracted from the translated trees as ordered trees. Then the information gains for the subtrees are calculated in the same way of SurveyAnalyzer. Although characteristics as ordered trees are extracted from the translated trees, they fall into characteristics as unordered trees or free trees of the dependency trees before translation, depending on which of the input filters was/were applied. A user specifies a mode from ordered or unordered or free tree mining mode depending on what the user wants to extract, and appropriate filters are selected to be applied.

As for the task 2, we introduce several output filters. These filters check the inclusion relations (and the information gains) among the extracted structures and delete those which are regarded as redundant for the sake of efficiency in knowledge discovery process. A user selects filters to be applied among them.

As for the task 3, for each extracted structure, we recover relevant information lost in the above tasks from the original documents, and construct a phrase or a sentence corresponding to the structure. Reconstruction candidates are generated by connecting the original words in the documents, and one with highest likelihood is selected based on the bigram language model.

Let us look at the behaviors of the system, using contact data of a help desk for an internal e-mail service as an example. Here, we extracted key semantics for questions/requests data field with contact duration (answered data/time minus contact date/time) less than five minutes as free trees. Original data are in Japanese. We translate the result into English here.

Table V shows the output of the sub-tree extraction unit, and **Table VI** shows the final output of the system after the redundancy reduction by deleting trees included by others with higher characteristic ranking and the phrase/sentence reconstruction.

In Table V, large structures such as the fourth tree are extracted. They contain much richer semantic contents than mere single words. However, the tree data are not easy to read. Moreover, it contains many similar results.

In Table VI, key semantics are output in a natural language, and almost all of the similar results are deleted from the list. The readability is much higher than Table I. Also, the redundancy in Table I was reduced sufficiently. The efficiency of knowledge discovery process will extensively improve.

5. TOPIC TREND DISCOVERY: TopicAnalyzer

Both SA and the key semantics mining system analyze data given collectively. However, in a wide range of business areas including CRM, knowledge management, and Web monitoring services, text data streams must be dealt with, and it is an important issue to discover topic trends and analyze their dynamics in real-time. For example, it is desired in the CRM area to grasp a new trend of topics in customers' claims every day and to track a new topic as soon as it emerges. Here a topic is defined as a burst of texts referring to a single event or activity.

Specifically we consider the following three tasks in topic trend analysis:

1) Topic Structure Identification; identifying what kinds of main topics exist and how important they are,
2) Topic Emergence Detection; detecting the emergence of a new topic and recognizing how it grows,
3) Topic Characterization; identifying the characteristics for each of main topics.

For real topic analysis systems, we may require that these three tasks be performed in an on-line fashion rather than in a retrospective way, and be dealt with in a single framework. We have proposed a new topic analysis system TopicAnalyzer (**Fig. 5**) which satisfies this requirement from a unifying viewpoint that a topic structure is modeled using a

finite mixture model (a model of the form of a weighted average of a number of probabilistic models) and that any change of a topic trend is tracked by learning the finite mixture model dynamically[10]. Here each topic corresponds to a single mixture component in the model.

All of the tasks 1)-3) are formalized in terms of a finite mixture model as follows:

Table V Extracted characteristic sub-trees.

Characteristic subtree	Score	Freq./ Total freq.
'set up'	7.98	11/15
'when'	6.68	18/30
password know	5.57	21/37
know password	5.57	21/37
how do execute Service XXX stop use	5.32	4/4
use Service XXX stop execute how do	5.32	4/4
how do execute use stop Service XXX	5.32	4/4
how do execute use Service XXX stop	5.32	4/4
use Service XXX stop how do execute	5.32	4/4
use Service XXX stop execute do how	5.32	4/4
...	...	...

As for the task 1), the topic structure is identified by statistical parameters of a finite mixture model. They are learned using our original time-stamp based discounting learning algorithm, which incrementally and adaptively estimates statistical parameters of the model by gradually forgetting out-of-date statistics, making use of time-stamps of data. This makes the learning procedure adaptive to changes of the nature of text streams.

As for the task 2), any change of a topic structure is recognized by tracking the change of main components in a mixture model. We apply the theory of dynamic model selection[11] to detecting changes of the optimal number of main components and their organization in the finite mixture model. We may recognize that a new topic has emerged if a new mixture component is detected in the model and remains for a while.

As for the task 3), we classify every text into the cluster for which the posterior probability is largest, and then we characterize each topic using feature terms characterizing texts classified into its corresponding cluster. These feature terms are extracted as those of highest information gain, which are computed in real-time.

The overall flow is illustrated in **Fig. 6**. A text is sequentially input to the system. We prepare a number of finite mixture models, for each of which we learn statistical parameters using the time-stamp based learning algorithm to perform topic identification. These tasks are performed in parallel. On the basis of the input data and learned models, we conduct dynamic model selection for choosing the optimal finite mixture model. We then compare the new optimal model with the last one to conduct topic emergence detection. Finally for each component of the optimal model, we conduct topic characterization.

We conducted an experiment on the same contact data used in Section 4. It has the field of contact date/time, question/request, answered date/time, answer, and so on.

We input contact dates as the time-stamps, and

Table VI Output of key semantics mining.

Key semantics	Score	Freq. / Total Freq.
'set up'	7.98	11/15
'from when'	6.68	18/30
'do not know password'	5.57	21/37
'How to stop use of Service XXX'	5.32	4/4
...	...	...

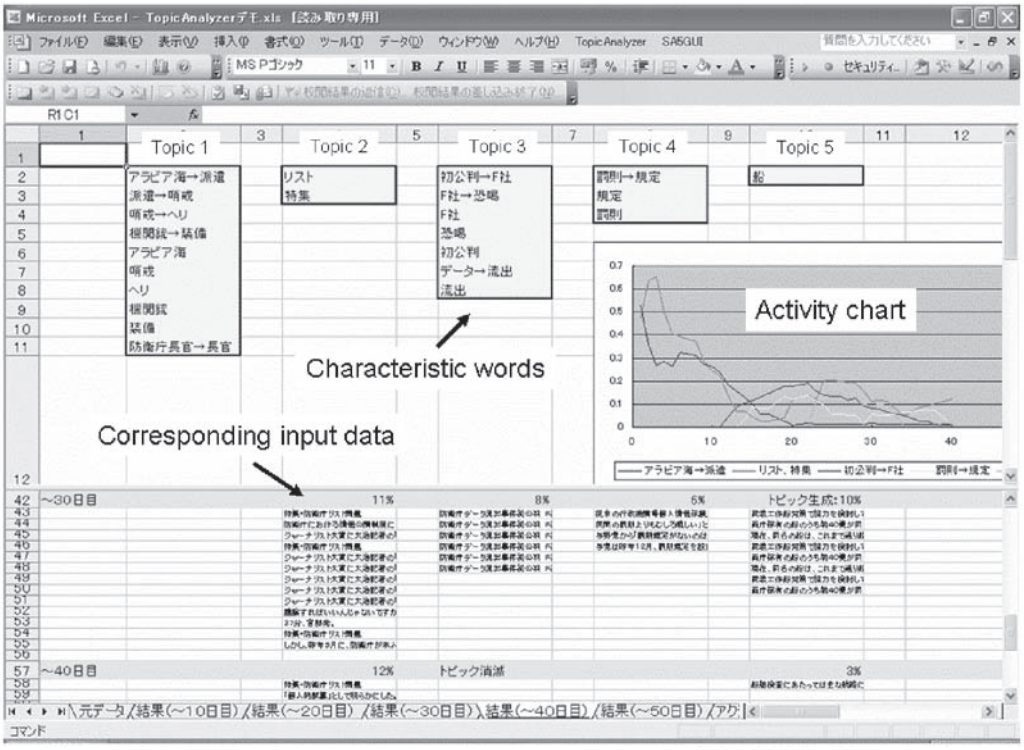


Fig. 5 TopicAnalyzer.

questions/requests as the text data to our system.

Figure 7 shows the number of topics detected by our system as active. The number increases at the beginning of March, and has a peak in the middle of April. Since a fiscal year begins at April in Japan, we can suppose that the number of topics at the help desk is increasing around the first day of April.

Let us look into a few of the components detail. **Figure 8** shows the activities and life periods of topics which have characteristic word(s) of "transfer" and "Service ZZZ," "failure," respectively. The activity of the former increases in the beginning of April and has the first peak at April 12. Then it repeats increase and decrease until the middle of May. The latter is active only during April. The lines indicate the relative input frequency of corresponding text data, and show how important the corresponding topics are in each time. From the figure, we can observe how the emerged topics grow and disappears.

Texts classified corresponding to the topic "transfer" are questions like "is it possible to use Service XXX after I am transferred to YYY?" That kind of questions may increase around the beginning of a fiscal year. In terms of the topic "Service ZZZ" and "failure," Service ZZZ actually failed in the beginning of April, when the topic became active, and the

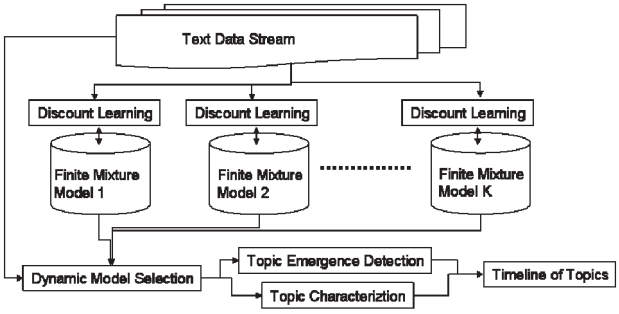


Fig. 6 Overall flow of the system.

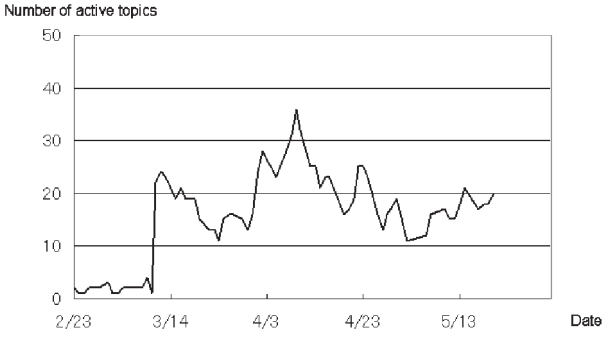


Fig. 7 Number of active topics.

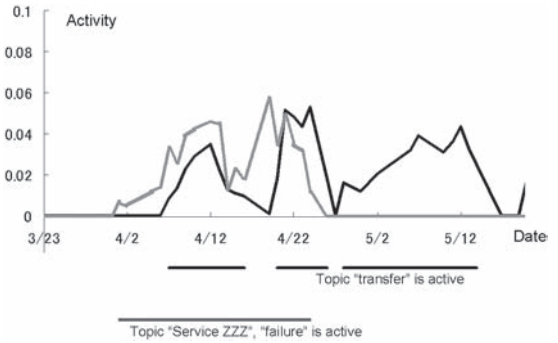


Fig. 8 Activity of topics.

corresponding texts consisted of related complaints and questions.

All of these results can be output in real time. In this way we can recognize the emergence, growth, and decay of each topic from the system. Through this example it has turned out that our framework for topic trend analysis are very effective for tracking dynamics of topic trends in contact data at a help desk.

6. CONCLUSION

We have introduced NEC's research activities on data mining focusing on Knowledge Organization. It means the technology system whose goal is to discover useful business knowledge mainly from text data. In it, data mining has turned out to contribute to drastic reduction of operators' cost for classifying or clustering a large volume of text data. We gave a brief overview of three text mining technologies: Text mining tool SurveyAnalyzer, key semantics mining, and topic analysis, and showed that how effective they were in business knowledge creation, in the areas of CRM, knowledge management, and marketing. We expect that they would further contribute to business process optimization, real-time management etc. in future.

REFERENCES

[1] U. Fayyad, P. Shapiro and P. Smyth, "From Data Mining to Knowledge Discovery: An Overview," *Advances in Knowledge Discovery and Data Mining*, AAAI Press/The MIT Press, Menlo Park, CA, pp.1-34, 1996.
[2] http://www.labs.nec.co.jp/DTmining/
[3] http://www.topicscope.com/index.html
[4] K. Yamanishi and H. Li, "Mining Open Answers in Questionnaire Data," *IEEE Intelligent Systems*, pp.58-63, 2002.
[5] K. Yamanishi, "A decision-theoretic extension of stochastic complexity and its applications to learning," *IEEE Trans. on Information Theory*, **44**, pp.1424-1439, 1998.
[6] S. Morinaga, K. Yamanishi, et al., "Mining Product Reputations on the Web," in Proc. of KDD 2002, pp. 341-349, 2002.
[7] S. Morinaga, H. Arimura, et al., "Characteristic Unordered/Free Tree Structure Extraction Using Enumeration of Ordered Trees," in Proc. of Workshop on Information-Based Induction Sciences (IBIS2004), (in Japanese), 2004.
[8] T. Asai, K. Abe, et al., "Efficient Substructure Discovery from Large Semi-structured Data," in Proc. of SDM'02, pp.158-174, SIAM, 2002.
[9] K. Satoh, T. Ikeda, et al., "Japanese Processing Middleware for Customer Relationship Management," in Proc. of the 9th Annual Meeting of the Association for NLP, pp.109-112, (in Japanese), 2003.
[10] S. Morinaga and K. Yamanishi, "Tracking Dynamics of Topic Trends Using a Finite Mixture Model," in Proc. of KDD 2004, pp.811-816, 2004.
[11] Y. Maruyama and K. Yamanishi, "Dynamic Model Selection with Its Applications to Computer Security," in Proc. of 2004 IEEE Information Theory Workshop, 2004.

Received April 28, 2005

* * * * * * * * * * * * * * * *

Kenji YAMANISHI received his M.E degree from the University of Tokyo in 1987. He joined NEC Corporation in 1987, and is now a Research Fellow of Internet Systems Research Laboratories. He worked for NEC Research Institute in U.S. as a visiting scientist, from 1992 to 1995. He received Doctorial degree from the University of Tokyo in 1992. He is engaged in research and development of data mining technologies.

Satoshi MORINAGA received his M.E degree from the University of Tokyo in 1994. He joined NEC Corporation in 1994, was transferred to the financial supervisory agency of Japanese government, and is now an Assistant Manager of Internet Systems Research Laboratories. He received Doctorial degree from the University of Tokyo in 1999. He is engaged in research and development of data mining technologies.

* * * * * * * * * * * * * * * *