World Academy of Science, Engineering and Technology
International Journal of Computer and Information Engineering
Vol:2, No:11, 2008

A Sequential Pattern Mining Method based on Sequential Interestingness

Shigeaki Sakurai, Youichi Kitahara, and Ryohei Orihara

***Abstract*—**Sequential mining methods efficiently discover all frequent sequential patterns included in sequential data. These methods use the support, which is the previous criterion that satisfies the Apriori property, to evaluate the frequency. However, the discovered patterns do not always correspond to the interests of analysts, because the patterns are common and the analysts cannot get new knowledge from the patterns. The paper proposes a new criterion, namely, the sequential interestingness, to discover sequential patterns that are more attractive for the analysts. The paper shows that the criterion satisfies the Apriori property and how the criterion is related to the support. Also, the paper proposes an efficient sequential mining method based on the proposed criterion. Lastly, the paper shows the effectiveness of the proposed method by applying the method to two kinds of sequential data.

***Keywords*—**Sequential mining, Support, Confidence, Apriori property

I. Introduction

Owing to the progress of computer and network environments, it is easy to collect data with time information such as daily business reports, web log data, and physiological information. This is the context in which methods of analyzing data with time information have been studied. Some previous studies dealt with numerical sequential data and other studies dealt with discrete sequential data. The paper focuses on analysis of discrete data.

J. Ayres et al. [3], J. Pei et al. [12], R. Srikant et al. [17], and M. J. Zaki [21] proposed methods that efficiently discover frequent sequential patterns from sequential data. The data is composed of rows of item sets and the patterns are frequent sub-rows of the item sets. These methods regard the frequent patterns as characteristic patterns. However, the patterns do not always correspond to the interests of analysts, because the patterns are common and are not a source of new knowledge for the analysts.

For this problem, M. N. Garofalakis et al. [6] proposed a method that uses user-specified regular expressions as background knowledge. The method applies sequential patterns to the regular expressions and extracts only sequential patterns that satisfy the regular expressions. Pei et al. [13] presented 7 kinds of constraints, including an item constraint, a super-pattern constraint, and a regular expression constraint. Here, the item constraint can extract sequential patterns that include or do not include specific items, and the super-pattern constraint can extract sequential patterns that include specific sequential sub-patterns. Pei et al. [13] also investigated characterization of the constraints and proposed a new framework that characterizes the constraints. On the other hand, S. -J. Yen [20] proposed an SQL-like data mining language. The language is used to discover sequential patterns including items corresponding to the interests of analysts. Sakurai et al. [14] proposed a method that introduces sub-pattern constraints and time constraints. The method extracts sequential patterns that include user-specified sub-patterns and satisfy time constraints between items. Using these methods, analysts can discover characteristic sequential patterns according to their interests. However, these methods require background knowledge depending on analysis tasks. If the analysts have insufficient background knowledge, the methods cannot discover characteristic sequential patterns.

Thus, this paper proposes a method that discovers sequential patterns corresponding to the interests of analysts without using background knowledge. The paper defines a new criterion called the sequential interestingness. The criterion is different from many previous criteria for data mining [7] [11]. It is possible for the criterion to discover sequential patterns with relatively high frequency and high confidence. That is, the sequential patterns with high sequential interestingness can predict remaining item sets in the case that the sequential sub-patterns are given. The patterns are regarded as kinds of characteristic sequential patterns. Also, the criterion satisfies the Apriori property, where the property is indispensable to efficient discovery of sequential patterns. The paper shows some theoretical properties of the criterion and proposes a new sequential mining method based on the criterion. Lastly, the paper applies the method to two kinds of sequential data and verifies the effectiveness of the proposed method through numerical experiments.

II. Sequential Interestingness

A. Sequential pattern

This paper deals with sequential patterns composed of rows of item sets. Here, each item set has some items that occur at the same time, but each item set does not have multiple identical items. Formally, a sequential pattern s_x is described as $(l_{x1},\ l_{x2},\ \cdots,\ l_{xn_x})$, where l_{xi} is an item set and n_x is the number of the item sets included in the sequential pattern. The number n_x is called length and the pattern s_x is called n_x-sequential pattern. Also, each l_{xi} is described as

Shigeaki Sakurai is with the System Engineering Laboratory, Corporate Research & Development Center, Toshiba Corporation, Kawasaki, e-mail: shigeaki.sakurai@toshiba.co.jp

Youichi Kitahara is with the System Engineering Laboratory, Corporate Research & Development Center, Toshiba Corporation, Kawasaki, e-mail: youichi.kitahara@toshiba.co.jp

Ryohei Orihara is with the HumanCentric Laboratory, Corporate Research & Development Center, Toshiba Corporation, Kawasaki, e-mail: ryohei.orihara@toshiba.co.jp

$(v_{xi1}, v_{xi2}, \cdots, v_{xin_{xi}})$, where v_{xij} is an item, $v_{xik_1} \neq v_{xik_2}$ $(k_1 \neq k_2)$, and n_{xi} is the number of items included in l_{xi}. For example, ({"beer", "diaper"}, {"beer", "snack", "milk"}, {"snack", "diaper"}) is an example of the sequential pattern. The pattern is a 3-sequential pattern and is composed of three item sets: {"beer", "diaper"}, {"beer", "snack", "milk"}, and {"snack", "diaper"}. "beer", "diaper", "milk", and "snack" are items. The pattern shows that a person buys "beer" and "diaper" on the first day, buys "beer", "snack", and "milk" on the second day, and buys "snack" and "diaper" on the third day.

On the other hand, when two sequential patterns s_1 and s_2 are given, their inclusion $s_2 = (l_{21}, l_{22}, \cdots, l_{2n_2}) \subseteq s_1 = (l_{11}, l_{12}, \cdots, l_{1n_1})$ is defined: $\exists \{z_1, z_2, \cdots, z_{n_2}\}$, $l_{21} \subseteq l_{1z_1}, l_{22} \subseteq l_{1z_2}, \cdots, l_{2n_2} \subseteq l_{1z_{n_2}}$. Figure 1 shows an example of the inclusion. In this figure, each circle shows an item and the same items have the same pattern on the circle. The concept of inclusion is used in evaluating the frequency of sequential patterns. The frequency is the number of sequential data including the patterns.

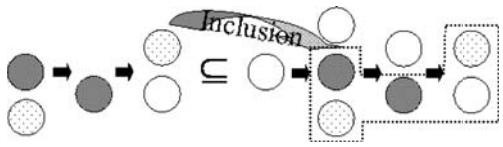


Fig. 1. Inclusion of sequential patterns

In the following sections, the indicator of each symbol is appropriately deleted to simplify the notation.

B. Previous criteria

The support and the confidence [2] are the most popular criteria for sequential patterns. The support evaluates frequencies of the patterns and the confidence evaluates frequencies of patterns in the case that sub-patterns are given. These criteria are defined by Formula (1) and Formula (2), respectively.

$$supp(s) = \frac{f_s(s)}{N} \tag{1}$$

$$conf(s|s_p) = \frac{f_s(s)}{f_s(s_p)} \tag{2}$$

Here, s is a sequential pattern, s_p is a sequential sub-pattern of the pattern s, $f_s(s)$ is frequency of the pattern s, and N is the total number of sequential data.

The support satisfies the Apriori property. That is, values of the support of any sequential patterns are smaller than or equal to values of the support of their sub-patterns. It is possible for the property to judge whether values of the support of sequential patterns are frequent or not by evaluating only sequential patterns composed of frequent sequential sub-patterns. The mining method based on the support can efficiently discover all frequent sequential patterns. However, the patterns do not always correspond to the interests of analysts, because the patterns are common and the analysts already know about the patterns. The analysts cannot always discover sequential patterns corresponding to their interests by the method.

The confidence can evaluate the relationships between sequential patterns and their sequential sub-patterns. The method based on the confidence discovers sequential patterns whose frequencies are close to the frequencies of their sequential sub-patterns. The analysts are interested in the patterns, because the analysts can regard the patterns as probable inference rules. That is, the analysts can predict remaining item sets with high probability by referring to the patterns when their sequential sub-patterns are given. However, the confidence does not satisfy the Apriori property. The method cannot efficiently discover sequential patterns with high confidence. Usually, sequential mining methods discover sequential patterns with high support and extract sequential patterns with high confidence from the patterns. These methods do not assure the discovery of all sequential patterns with high confidence. We have to set a smaller threshold for the support to avoid missing sequential patterns with high confidence. The methods tend to discover many sequential patterns. The analysts have to spend much time examining the patterns.

On the other hand, in regard to the study of associative rules R. Agrawal et al. [1], J. Blanchard et al. [4], S. Brin et al. [5], K. Shimazu et al. [15], A. Silberschatz et al. [16], and E. Suzuki et al. [18] proposed other criteria. K. Shimazu et al. [15] and E. Suzuki et al. [18] proposed criteria that discover exceptional patterns where the patterns have low support and high confidence. A. Silberschatz et al. [16] proposed a criterion that measures the interestingness of a pattern in terms of the belief system the analysts have. J. Blanchard et al. [4] proposed a criterion that is based on a probabilistic model and measures the deviation from the maximum uncertainty of the consequent given that the antecedent is true. In addition, S. Brin et al. [5] proposed a criterion that measures significance of associations via the χ^2 test for correlation from classical statistics. The patterns discovered by the criteria are not always frequent but are characteristic of viewpoints. We may be able to apply the criteria to sequential pattern mining methods. However, the criteria do not satisfy the Apriori property. The methods based on the criteria have a problem similar to the methods based on the confidence. It is necessary to define a new criterion in order to efficiently discover sequential patterns corresponding to the interests of analysts.

C. Definition of sequential interestingness

We note that a sequential pattern includes sequential sub-patterns whose frequencies are not always high and are close to the frequency of the pattern. The pattern is tied to the sub-patterns with high confidence, despite the fact that the frequency of the sub-pattern is not high. We can use the pattern as a probable inference rule. The analysts would be interested in such patterns. Thus, we define a new criterion that discovers the pattern by Formula (3), where α (≥ 0) is a parameter which represents how important the frequency of the pattern is. The first term of the criterion evaluates that the frequencies of the sub-patterns are not frequent and the second term of the criterion evaluates that the frequency of the pattern is frequent. The criterion is called the *sequential interestingness*, the parameter α is called the *confidence priority*, and the

pattern that is bigger than or equal to the minimum sequential interestingness given by the analysts is called the *interesting pattern* in the following. In particular, the pattern is called the *interesting set* when the length is 1, and the interesting set is called the *interesting item* when the set is composed of an item.

$$inst(s) = \min_{s_p \subseteq s} \{(\frac{1}{f_s(s_p)})^{\alpha}\} \times \frac{(f_s(s))^{(1+\alpha)}}{N} \quad (3)$$

Formula (3) helps us discover the pattern whose frequency is relatively high, despite the fact that the frequencies of the sub-patterns are not always high. The formula corresponds to the definition of the support in the case that α is equal to 0. The formula shows that the sequential interestingness corresponds to the support in the case that the number of items included in the pattern is 1, because $\min_{s_p \subseteq s}(\frac{1}{f_s(s_p)}) = \frac{1}{f_s(s)}$. In addition, the formula satisfies the Apriori property. The proof is given in the following.

[Proof] Let s_1 and s_2 be such interesting patterns that s_1 is included in s_2.

$$\begin{aligned} inst(s_2) &= \min_{s_p \subseteq s_2} \{(\frac{1}{f_s(s_p)})^{\alpha}\} \times \frac{(f_s(s_2))^{(1+\alpha)}}{N} \\ &\leq \min_{s_p \subseteq s_2} \{(\frac{1}{f_s(s_p)})^{\alpha}\} \times \frac{(f_s(s_1))^{(1+\alpha)}}{N} \end{aligned}$$

Here, we note that the set of sub-sequences included in s_2 is equal to the union of two sets of sub-sequences. One is the set of sub-sequences included in s_1, and the other is the set of sub-sequences included in s_2 and not included in s_1.

$$\begin{aligned} &= \min[\min_{s_p \subseteq s_1} \{(\frac{1}{f_s(s_p)})^{\alpha}\}, \\ &\quad \min_{(s_p \subseteq s_2) \cap (s_p \not\subseteq s_1)} \{(\frac{1}{f_s(s_p)})^{\alpha}\}] \times \frac{(f_s(s_1))^{(1+\alpha)}}{N} \\ &\leq \min_{s_p \subseteq s_1} \{(\frac{1}{f_s(s_p)})^{\alpha}\} \times \frac{(f_s(s_1))^{(1+\alpha)}}{N} = inst(s_1) \end{aligned}$$

By the above transformation, the sequential interestingness satisfies the Apriori property. □

Next, we investigate the relationships among the sequential interestingness, the support, and the confidence. We can get the Formula (4) by transforming Formula (3). Therefore, we can regard the sequential interestingness as the support adjusted by the minimum value of the confidence of sequential sub-patterns included in a pattern. That is, the analysts can decide the sequential interestingness based on the support and the confidence that the analysts require. Also, the analysts can decide the importance of the confidence by adjusting the confidence priority. The importance of the confidence increases as the confidence priority increases. Typically, the analysts can set 1.0 to the confidence priority, because the value evaluates sequential patterns such that the importance of the confidence corresponds to the importance of the support.

$$\begin{aligned} inst(s) &= \min_{s_p \subseteq s} \{(\frac{1}{f_s(s_p)})^{\alpha}\} \times \frac{(f_s(s))^{(1+\alpha)}}{N} \\ &= \min_{s_p \subseteq s} \{(\frac{f_s(s)}{f_s(s_p)})^{\alpha}\} \times \frac{f_s(s)}{N} \\ &= \min_{s_p \subseteq s} \{(conf(s|s_p))^{\alpha}\} \times supp(s) \end{aligned} \quad (4)$$

We also get the Formula (5) by transforming Formula (3). Here, $f_s(v)$ is frequency of item v. Now, we can regard the sequential interestingness as the support adjusted by the minimum values of the confidence of the items included in a pattern.

$$\begin{aligned} inst(s) &= \min_{s_p \subseteq s} \{(\frac{1}{f_s(s_p)})^{\alpha}\} \times \frac{(f_s(s))^{(1+\alpha)}}{N} \\ &= \frac{1}{\max_{s_p \subseteq s} \{(f_s(s_p))^{\alpha}\}} \times \frac{(f_s(s))^{(1+\alpha)}}{N} \\ &= \frac{1}{\max_{v \in s} \{(f_s(v))^{\alpha}\}} \times \frac{(f_s(s))^{(1+\alpha)}}{N} \\ &= \min_{v \in s} \{(\frac{1}{f_s(v)})^{\alpha}\} \times \frac{(f_s(s))^{(1+\alpha)}}{N} \\ &= \min_{v \in s} \{(\frac{f_s(s)}{f_s(v)})^{\alpha}\} \times \frac{f_s(s)}{N} \\ &= \min_{v \in s} \{(conf(s|v))^{\alpha}\} \times supp(s) \end{aligned} \quad (5)$$

We note how the sequential interestingness can discover the patterns whose frequencies are smaller than the frequencies of sequential patterns that are not interesting patterns. The discovery of the patterns requires satisfying the following condition.

$$0 \leq \alpha \quad (6)$$

$$0 \leq m_1 \leq m_2 \quad (7)$$

$$0 \leq m_1 \leq y \quad (8)$$

$$0 \leq m_2 \leq x \quad (9)$$

$$(\frac{m_1}{y})^{\alpha} \frac{m_1}{N} \geq (\frac{m_2}{x})^{\alpha} \frac{m_2}{N} \quad (10)$$

But, m_1 is the frequency $f_s(s_1)$ of a sequential pattern s_1, m_2 is the frequency $f_s(s_2)$ of a sequential pattern s_2, and y and x correspond to $\max_{v \in s_1}\{f_s(v)\}$ and $\max_{v \in s_2}\{f_s(v)\}$, respectively. We note that m_1 and y depend on each other due to the sequential pattern s_1. m_2 and x have a similar relationship. However, y and x can be any values for any m_1 and m_2 in the range that satisfies the conditions (8) and (9), m_1 and y are regarded as independent of each other, and m_2 and x are dealt with similarly. On the other hand, the condition (11) is given by the conditions (6) and (10).

$$\text{If } \alpha \neq 0, \text{then } (\frac{m_1}{m_2})^{\frac{\alpha+1}{\alpha}} x \geq y \quad (11)$$

Thus, the shadowed range $S_{\alpha}(x)$ in Figure 2 corresponds to the range in which the frequencies of the interesting patterns are smaller than the frequencies of the sequential patterns that are not interesting patterns when $\alpha \neq 0$. We calculate the area of the range from 0 to T: $S_{\alpha}(x)$. Here, $T > \frac{m_2^{\frac{\alpha+1}{\alpha}}}{m_1^{\frac{1}{\alpha}}}$ and α is a confidence priority. The area is given by Formula (12).

$$S_{\alpha}(x) = \int_{\frac{m_2^{\frac{\alpha+1}{\alpha}}}{m_1^{\frac{1}{\alpha}}}}^{T} \{(\frac{m_1}{m_2})^{\frac{\alpha+1}{\alpha}} x - m_1\} dx \quad (12)$$

Next, we calculate the difference of the areas for two confidence priorities α_1 and α_2, where $(\alpha_1 > \alpha_2 > 0)$. The

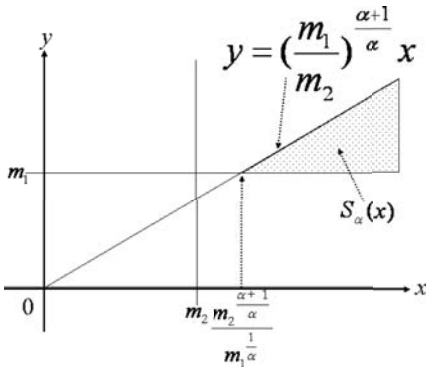


Fig. 2. Range with higher sequential interestingness

difference is given by Formula (13).

$$
\begin{aligned}
S_{\alpha_1} - S_{\alpha_2} &= \int_{m_2^{\frac{\alpha_1+1}{\alpha_1}}}^{T} \{(\frac{m_1}{m_2})^{\frac{\alpha_1+1}{\alpha_1}} x - m_1\}dx \\
&\quad - \int_{m_2^{\frac{\alpha_2+1}{\alpha_2}}}^{T} \{(\frac{m_1}{m_2})^{\frac{\alpha_2+1}{\alpha_2}} x - m_1\}dx \\
&= [\frac{1}{2}(\frac{m_1}{m_2})^{\frac{\alpha_1+1}{\alpha_1}} x^2 - m_1 x]_{m_1^{\frac{1}{\alpha_1}}}^{T} \\
&\quad - [\frac{1}{2}(\frac{m_1}{m_2})^{\frac{\alpha_2+1}{\alpha_2}} x^2 - m_1 x]_{m_1^{\frac{1}{\alpha_2}}}^{T} \\
&= \{\frac{T^2}{2}(\frac{m_1}{m_2})^{\frac{\alpha_1+1}{\alpha_1}} - m_1 T + \frac{m_2^{\frac{\alpha_1+1}{\alpha_1}}}{2m_1^{\frac{1-\alpha_1}{\alpha_1}}}\} \\
&\quad - \{\frac{T^2}{2}(\frac{m_1}{m_2})^{\frac{\alpha_2+1}{\alpha_2}} - m_1 T + \frac{m_2^{\frac{\alpha_2+1}{\alpha_2}}}{2m_1^{\frac{1-\alpha_2}{\alpha_2}}}\} \\
&= \frac{T^2}{2}\{(\frac{m_1}{m_2})^{\frac{\alpha_1+1}{\alpha_1}} - (\frac{m_1}{m_2})^{\frac{\alpha_2+1}{\alpha_2}}\} \\
&\quad + \frac{{m_1}^2}{2}\{(\frac{m_2}{m_1})^{\frac{\alpha_1+1}{\alpha_1}} - (\frac{m_2}{m_1})^{\frac{\alpha_2+1}{\alpha_2}}\}
\end{aligned}
\tag{13}
$$

Here, the condition $\frac{\alpha_2+1}{\alpha_2} > \frac{\alpha_1+1}{\alpha_1} > 0$ is given by the condition $\alpha_1 > \alpha_2 > 0$. Also, the condition $\frac{m_1}{m_2} \leq 1$ is given. The first term in Formula (13) is always positive. In addition, the second term in Formula (13) is constant. Therefore, the difference is always positive, when the condition $T \to \infty$ is satisfied. On the other hand, $S_\alpha(x) = 0$ when $\alpha = 0$. These results show that interesting patterns with small frequency can be discovered more frequently as the confidence priority is bigger. That is, the sequential interestingness more easily discovers the patterns that are different to the patterns discovered by the support as the confidence priority is bigger.

D. Discovery of interesting sequential patterns

We try to compose an algorithm that discovers all interesting patterns based on the sequential interestingness. The criterion satisfies the Apriori property. Thus, we consider an efficient algorithm that is similar to AprioriAll [2] [17]. The algorithm is composed of three processes: the interesting item discovery, the interesting set discovery, and the interesting pattern discovery. In the following, each process is explained in detail.

At first, the interesting item discovery process picks up an item from sequential data. The process calculates the number of sequential data including the item as the frequency of the item. We note that the sequential interestingness corresponds to the support when the number of the items included in the sequential pattern is equal to 1. The process can calculate the sequential interestingness of the item by dividing the frequency with the total number of sequential data. The process judges the item to be an interesting item, if the sequential interestingness of the item is bigger than or equal to the minimum sequential interestingness. The process stores the frequencies of the interesting item. The process discovers all interesting items by evaluating all items included in the sequential data.

Next, the interesting set discovery process generates a candidate item set with two interesting items. The process calculates the frequency of the item set by applying it to the sequential data. The process calculates the sequential interestingness of the item set based on their frequencies, the frequencies of the item set, the total number of the sequential data, and the confidence priority. The process judges the item set to be an interesting set with two items, if the sequential interestingness of the item set is bigger than or equal to the minimum sequential interestingness. The process discovers all interesting sets with two items by evaluating all combinations of the interesting items. We note that the process can discover the interesting sets with two items, even if the process evaluates only the combinations of the interesting items. This is because the sequential interestingness satisfies the Apriori property. The process also generates a candidate item set with three items by combining two interesting sets with two items. That is, the process picks up two interesting sets with two items, where the sets have a common item. Then, the items are sorted with a specific order such as an alphabetic order in the interesting sets. It is easy for the process to pick up the two interesting sets by checking whether the first item of one set corresponds to the first item of the other set. The process generates the candidate by arranging three different items included in the sets. The process similarly evaluates the candidate item set and judges whether the item set is interesting or not. The process discovers all interesting sets with three items. In general, the process generates a candidate item set with $(i+1)$ items V_{i+1} $(= \{v_1, \cdots, v_{i-1}, v_i, v_{i+1}\})$ from two interesting sets with i items $V_{i,1}$ $(= \{v_1, \cdots, v_{i-1}, v_i\})$ and $V_{i,2}$ $(= \{v_1, \cdots, v_{i-1}, v_{i+1}\})$ as shown in Figure 3. But, the former $(i-1)$ items included in the two interesting sets with i items correspond to each other. The process evaluates whether the candidate item set is interesting or not. Then, it is necessary for the process to calculate the frequency $f_s(V_{i+1})$ and the maximum frequency of items in the set $\max_{v_k \in V_{i+1}}\{f_s(v_k)\}$. We note that the maximum frequency is divided into two parts as shown in Formula (14). Each part corresponds to the maximum frequency of items included in the interesting sets with i items $V_{i,1}$ and $V_{i,2}$, respectively. Therefore, it can easily calculate the maximum frequency of the candidate item set with $(i+1)$ items by storing the maximum frequencies of the interesting sets with i items.

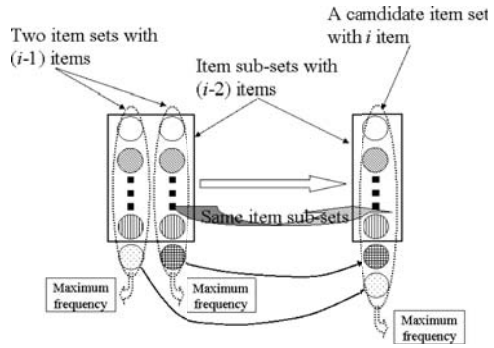


Fig. 3. Generation of a candidate item set

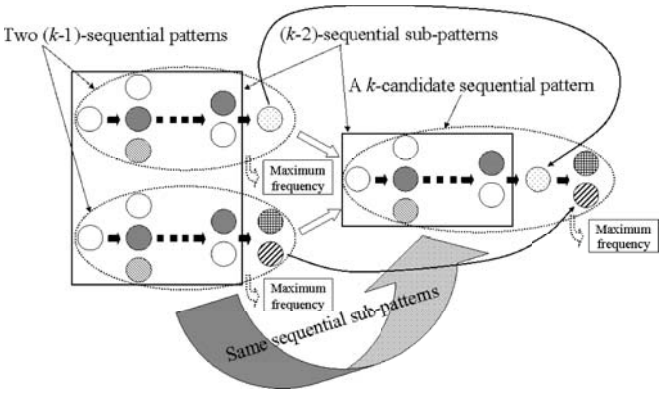


Fig. 4. Generation of a candidate sequential pattern

The process is repeated until all interesting sets are discovered.

$$\max_{\substack{v_k \in \{v_1,\cdots,\\ v_{i-1},v_i,v_{i+1}\}}} \{f_s(v_k)\} = \max[\max_{\substack{v_k \in \{v_1,\cdots,\\ v_{i-1},v_i\}}} \{f_s(v_k)\}, \max_{\substack{v_k \in \{v_1,\cdots,\\ v_{i-1},v_{i+1}\}}} \{f_s(v_k)\}] \quad (14)$$

The discovered interesting items and the discovered interesting sets are regarded as 1-interesting patterns.

Next, the interesting pattern discovery process generates 2-candidate sequential patterns by combining two 1-interesting patterns. But, the combination reflects the order of the interesting patterns. This is because 2-candidate sequential patterns (l_1,l_2) and (l_2,l_1) are different candidate sequential patterns. The process calculates the frequency of the 2-candidate sequential patterns. The process calculates the sequential interestingness based on their frequencies, the maximum frequency of the items included in the candidate sequential pattern, the total number of the sequential data, and the confidence priority. The process judges the candidate sequential pattern to be a 2-interesting pattern if the sequential interestingness is bigger than or equal to the minimum sequential interestingness. The process discovers all 2-interesting patterns by evaluating all combinations of 1-interesting patterns. Here, we note that only the combinations are evaluated. This is because the sequential interestingness satisfies the Apriori property. The process also picks up two 2-interesting patterns whose first items correspond to each other. The process generates 3-candidate sequential patterns by combining the two 2-interesting patterns. The process similarly evaluates whether the candidate sequential pattern is interesting or not. The process discovers all 3-interesting patterns by combining the two 2-interesting patterns. In general, the process generates a $(k+1)$-candidate sequential pattern from two k-interesting patterns as shown in Figure 4. But, the former $k-1$ item sets of the k-interesting patterns correspond to each other. We describes the k-interesting pattern as (s_p, l_1) and (s_p, l_2). That is, $(k+1)$-candidate sequential pattern is described as (s_p, l_1, l_2). The process evaluates whether the $(k+1)$-candidate sequential pattern is interesting or not. Then, it is necessary for the process to calculate the frequency $f_s((s_p, l_1, l_2))$ and the maximum frequency of items included in the pattern $\max_{v\in(s_p,l_1,l_2)}\{f_s(v)\}$. We note that the maximum frequency is divided into two parts as shown in Formula (15). Each part corresponds to the maximum frequency of items included in the respective interesting patterns (s_p, l_1) and (s_p, l_2). Therefore, it can easily calculate the maximum frequency of the $(k+1)$-candidate sequential pattern by storing the maximum frequencies of the k-sequential patterns. The process is repeated until all interesting sets are discovered.

$$\max_{v\in(s_p,l_1,l_2)} \{f_s(v)\} = \max[\max_{v\in(s_p,l_1)} \{f_s(v)\}, \max_{v\in(s_p,l_2)} \{f_s(v)\}] \quad (15)$$

Based on the above discussions, we can efficiently discover all interesting patterns by expanding the number of items and the length of the sequential pattern from discovery of 1-interesting patterns with an item. The algorithm incorporating the three processes is shown in Figure 5. The sequential data SeqDB, the minimum sequential interestingness MinInst, and the confidence priority α are input to the algorithm. The algorithm outputs all k-interesting patterns Q_k.

In this algorithm, calc_freq() is a function that calculates frequencies of sequential patterns, sf[] is a storage area that stores the maximum frequency of the items included in an interesting pattern, subset() is a function that picks up items with the number of items of an input value from an item set in alphabetic order, subseq() is a function that picks up a sequential sub-pattern with the number of item sets of an input value from the top of an interesting pattern, and $\bowtie_{seq}$ is an operator that generates a $(k+1)$-candidate sequential pattern from two k-interesting patterns.

In this paper, we described an AprioriAll-like sequential mining algorithm. However, these techniques of the sequential interestingness can be applied to a projection-based sequential mining method such as PrefixSpan [12].

III. Numerical Experiments

A. Data for experiments

We used two kinds of sequential data in order to evaluate the effect of the proposed method. One is daily business report data and the other is medical examination data.

The former data is 27,731 daily business reports written by salespersons. The reports were collected by an SFA (Sales Force Automation) system that was introduced in 5 departments. We classify the reports into clusters with respect to

```
//Interesting item discovery;
Q_11 = φ;
For each item v ∈ l, l ∈ d, d ∈ SeqDB
  freq=calc_freq(v, SeqDB, 1);
  inst=freq/|SeqDB|;
  If inst ≥ MinInst;
  Then store freq to sf[v];
       add v to Q_11;
//Interesting set discovery;
For(i=1; Q_1i!=φ; i++)
  Q_1i+1 = φ;
  N_1i = φ;
  For each interesting set V_1 ∈ Q_1i
    add V_1 to N_1i;
    For each interesting set V_2 ∈ (Q_1i − N_1i)
      If subset(V_1, i − 1)==subset(V_2, i − 1);
      Then V = V_1 ∪ V_2;
           freq=calc_freq(V, SeqDB, 1);
           tinst = max(sf[V_1], sf[V_2]);
           inst=(freq/tinst)^α × freq/|SeqDB|;
           If inst ≥ MinInst;
           Then store tinst to sf[V];
                add V to Q_1i+1;
Q_1 = ⋃_i Q_1i;
//Interesting pattern discovery;
For(k=1; Q_k!=φ; k++)
  Q_k+1 = φ;
  For each interesting pattern q_1 ∈ Q_k
    For each interesting pattern q_2 ∈ Q_k
      If subseq(q_1, k − 1)==subseq(q_2, k − 1);
      Then q = q_1 ⋈_seq q_2;
           freq=calc_freq(q, SeqDB, k + 1);
           tinst = max(sf[q_1], sf[q_2]);
           inst=(freq/tinst)^α × freq/|SeqDB|;
           If inst ≥ MinInst;
           Then store tinst to sf[q];
                add q to Q_k+1;
```

Fig. 5. A sequential mining algorithm based on the sequential interestingness

customer names and product names, and arrange the reports included in each cluster in order of time. But, if reports included in a cluster have the same time information, the reports are merged. We generate 6,434 sequences of sequential reports. Also, we extract items from the reports by using a key concept dictionary [8] and generate a row of item sets from each sequence of the reports. Here, the key concept dictionary is a kind of thesaurus. The dictionary is composed of 3 layers: the concept class, the key concept, and the expression. The expression describes important words or important phrases in the reports using regular expressions. The key concept is a set of expressions with the same meaning. The concept class is a set of relevant key concepts. Each key concept is regarded as an item. In the experiments, we use a key concept dictionary composed of 3 concept classes, 61 key concepts, and 835 expressions. The dictionary is created by a human expert for the analysis task of the daily business reports. For example, the dictionary extracts "inquiry regarding system specifications", "demonstration of a system", "difficult problem", "acceptance of an order", and so on as items. The sequential data is composed of 57,133 items.

The latter data is the data related to the medical examination that all employees of our company undergo every year. Most Japanese companies perform similar medical examinations. The data is composed of 50 attributes such as results of the medical examination, age, sex, and identification number of employees. The data is also encrypted to avoid identification of individual persons and the encryption key is controlled by the doctor. Specifically, this experiment deals with 28,995 data sets of males ranging in age from 20 to 29. We collect data sets of the same person and arrange the collected data sets according to the date of the medical examination. We generate 10,412 rows of item sets from the 28,995 data sets. We delete age, sex, identification number of employees, and the date from the sequential data sets. That is, 46 attributes remain. On the other hand, each attribute is composed of three or more attribute values. We generate an item corresponding to each attribute value. For example, if an attribute "blood pressure" is composed of three attribute values, "high", "normal", and "low", then "blood pressure/high", "blood pressure/normal", and "blood pressure/low" are generated as items. Therefore, each item set of the sequential data is composed of 454 items.

B. Experimental method

We perform two kinds of experiments. In these experiments, we use 7 confidence priorities: 0.0, 0.25, 0.5, 1.0, 2.0, 5.0, and 10.0. Here, we note that results of the confidence priority 0.0 correspond to the results of the support. We can compare the sequential interestingness with the support. Also, in the case of the medical examination data, we use two additional constraints in order to decrease the number of interesting patterns. One is the number of items included in item sets and the other is the time interval of continuous items. The number is 2 and the time interval is a year.

The first experiment uses 1.00% and 2.00% as the minimum sequential interestingness. The experiment investigates the numbers of the discovered sequential patterns. Also, the experiment evaluates the relationships between the numbers and confidence priorities.

The second experiment adjusts the minimum sequential interestingness. The experiment investigates the contents of discovered sequential patterns. Also, the experiment evaluates the relationships between the contents and the confidence priorities. In this experiment, the minimum support is 3.00%. That is, the minimum sequential interestingness is 3.00% in the case that the confidence priority is 0.0. The minimum sequential interestingness is adjusted such that the number of 1-interesting patterns in the case of the sequential interestingness is nearly equal to that in the case of the minimum support of 3.00%. Here, we note that the numbers are not completely equal because multiple 1-interesting patterns can have the same sequential interestingness. Table I shows the adjusted minimum sequential interestingness for each confidence priority. In this table, D_1 and D_2 refer to the daily business report data and the medical examination data, respectively. The minimum sequential interestingness decreases, according to the increase of the confidence priorities. This is because the

TABLE I
ADJUSTED MINIMUM SEQUENTIAL INTERESTINGNESS

α	0.0	0.25	0.5	1.0	2.0	5.0	10.0
D_1	3.00%	1.85%	1.19%	5.07×10^{-1}%	9.10×10^{-2}%	3.84×10^{-4}%	4.34×10^{-8}%
D_2	3.00%	1.54%	7.94×10^{-1}%	2.13×10^{-1}%	1.64×10^{-2}%	7.89×10^{-6}%	2.27×10^{-11}%

sequential interestingness is composed of the support and the confidence as shown in Formula (5). That is, the confidence has a large impact on the sequential interestingness in the case that the confidence priorities are big, and the number of sequential patterns with high confidence is small.

C. Experimental results

Figure 6 (a) $\sim$ Figure 6 (d) show results in the case of the daily business report data. Figure 6 (a) and Figure 6 (b) show the number of interesting patterns in the case of 1.00% and 2.00%, respectively. Also, Figure 6 (c) and Figure 6 (d) show the number of common interesting patterns and the number of different interesting patterns. That is, the common interesting patterns are the patterns discovered by both the sequential interestingness and the support. The different patterns are the patterns only discovered by the sequential interestingness. In these figures, x axis shows the confidence priorities and y axis shows the number of the patterns. Each line shows results corresponding to the length of the patterns.

Similarly, Figure 7 (a) $\sim$ Figure 7 (d) show results in the case of the medical examination data. But, in the case that the confidence priorities are small, there are insufficient points to discover many interesting patterns.

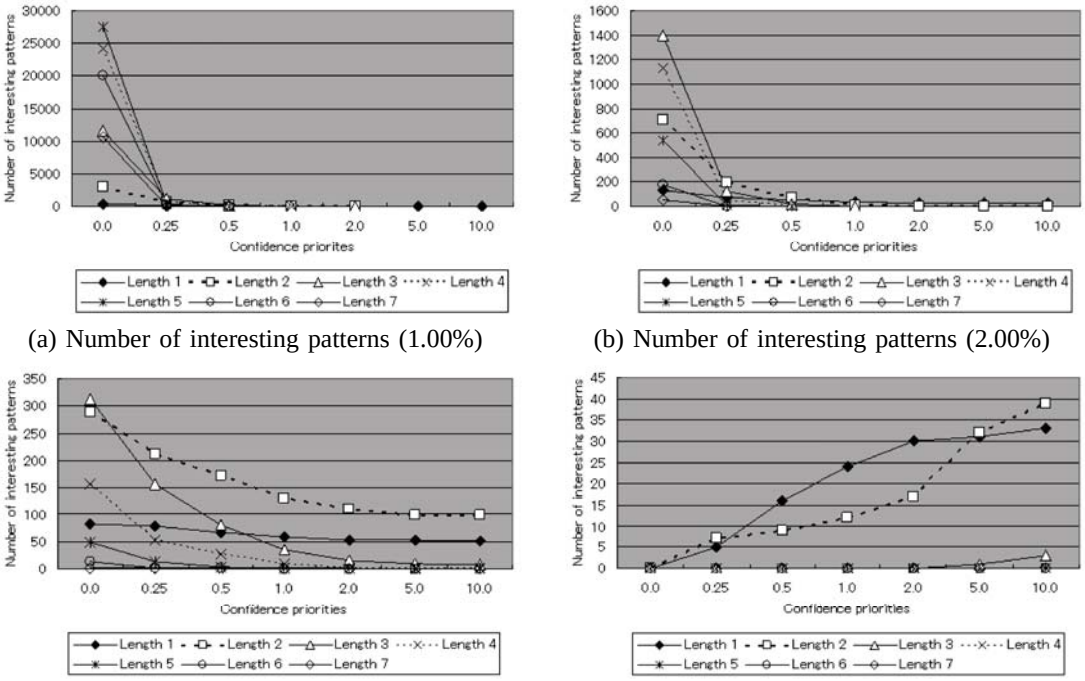
D. Discussions

Feature of sequential interestingness: The sequential interestingness corresponds to the support revised by the minimum confidence as shown in Formula (5). The sequential interestingness is smaller than the support, because the support and the minimum confidence range from 0.0 to 1.0. Formula (5) also shows that the first term in Formula (5) decreases as the confidence priority increases. Thus, the number of interesting patterns discovered by the sequential interestingness is smaller than the number discovered by the support when the minimum values are equal to each other. Figure 6 (a), Figure 6 (b), Figure 7 (a), and Figure 7 (b) correspond to this property. That is, the number of the patterns decreases as the confidence priorities increase. So, it is necessary to set the smaller minimum sequential interestingness, if we try to discover the patterns whose number is similar to the number of the patterns based on the support. Here, we note the relationships between the minimum support of 3.00% and the minimum sequential interestingness in the second experiment. The degrees of the minimum sequential interestingness are related to the confidence priorities as shown in Table I. We can set the confidence priorities by referring to the degree to some extent. In future work, we will try to devise a method that decides the best confidence priority.

Feature of discovered interesting patterns: We note Figure 6 (c), Figure 6 (d), Figure 7 (c), and Figure 7 (d). These figures show that the common patterns decrease and the different patterns increase as the confidence priorities increase. This is because the minimum confidence is more strongly evaluated than the support as the confidence priorities increase. The results correspond to the theoretical result that the patterns with the smaller frequency are more easily extracted. The different interesting patterns include the items that are not included in the patterns based on the support, because the items have small frequencies. On the other hand, the patterns only discovered by the support are apt to be composed of the items included in the common interesting patterns. The patterns are similar to the common interesting patterns. Therefore, the patterns based on the sequential interestingness are more attractive, because the patterns are composed of various items and different types of interesting patterns are discovered. The analysts can easily discover sequential patterns corresponding to their interests. For example, the big confidence priorities discover some attractive sequential patterns such as "the difficult problem $\rightarrow$ the good impression" and "the good impression $\rightarrow$ the difficult problem" in the case of the daily business report data. Also, the sequential interestingness can discover interesting patterns, even if the support cannot discover interesting patterns owing to the restriction of our computer environment in the case of the medical examination data.

Next, we note the total number of the discovered interesting patterns. The number of the patterns decreases as the length increases, even if the numbers of 1-interesting patterns are nearly equal to each other. The trend is clearer as the confidence priority increases and the length increases. This is because the sequential interestingness has two factors that lead to decrease of the patterns. One factor is the first term in Formula (5) and the other factor is the second term in Formula (5). The first term decreases monotonically as the length increases, because the number of items included in the patterns increases and the maximum frequency increases. The first term also decreases monotonically as the confidence priority increases. In addition, the second term decreases monotonically as the length increases. The sequential interestingness is drastically smaller than the support. The number of the patterns discovered by the sequential interestingness is drastically smaller. On the other hand, the operation based on the sequential interestingness is similar to the operations in which the confidence squeezes the patterns after the support squeezes the patterns. Therefore, the sequential interestingness can discover more attractive sequential patterns than the support.

Calculation time: It is necessary for the sequential interestingness to calculate the maximum frequency of the items included in the patterns. Therefore, its calculation time is larger than the calculation time of the support. The mining algorithm of sequential patterns requires additional calculation time in the case of the sequential interestingness. However, the



(a) Number of interesting patterns (1.00%)

(b) Number of interesting patterns (2.00%)

(c) Number of interesting patterns (common)

(d) Number of interesting patterns (different)

Fig. 6. Experimental results in the case of the daily business report data

mining algorithm can easily calculate the maximum frequency by storing the maximum frequencies of interesting patterns and the time required is not long. On the other hand, the mining algorithm requires a large amount of calculation time to calculate the frequencies of the patterns. The latter calculation time is much larger than the former calculation time. The calculation time of the mining algorithm does not nearly increase, even if the sequential interestingness is calculated instead of the support. In this experiment, the AprioriAll-like mining algorithm was the experimental system. We could not discover any notable difference of the calculation time between the sequential interestingness and the support. We think the increase of the calculation time is small, even if the sequential interestingness is incorporated into the other mining algorithms such as PrefixSpan.

Based on the above discussions, the mining algorithm based on the sequential interestingness can efficiently discover more attractive sequential patterns than the method based on the support. We believe that the algorithm can easily discover sequential patterns corresponding to the interests of analysts.

IV. Summaru and Future Work

This paper has proposed a new criterion: the sequential interestingness. The paper theoretically has shown some properties of the criterion and the relationships among the criterion, the support, and the confidence. The paper has also proposed an efficient sequential mining method based on the new criterion. In addition, the paper has verified the effectiveness of the proposed method by applying the method to daily business report data and medical examination data.

In the future, by listening to the views of analysts, we will try to verify in more detail whether or not the discovered interesting patterns give them new knowledge. This is because the patterns based on the interesting patterns may not give them new knowledge, even if the patterns are composed of many kinds of items, the patterns and their specific sub-patterns are related to each other with high confidence, and some attractive sequential patterns are extracted from the daily business report data. We will also consider how to decide the confidence priority, because it is not easy for the analyst to decide the importance of the confidence. In addition, we will also try to deal seamlessly with discrete data, textual data, and numerical data in order to deal with them in many application fields.

References

[1] R. Agrawal and R. Srikant, "Fast Algorithms for Mining Association Rules," *in Proc. of the 20th Int. Conf. Very Large Data Bases*, 1994, Santiago de Chile, Chile, pp. 487-499.

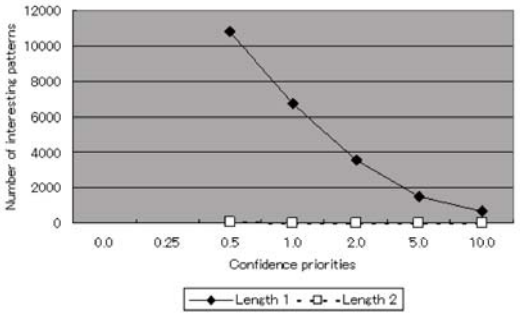
[2] R. Agrawal and R. Srikant, "Mining Sequential Patterns," *in Proc. of the 11th Int. Conf. Data Engineering*, 1995, Taipei, Taiwan, pp. 3-14.

[3] J. Ayres, J. E. Gehrke, T. Yiu, and J. Flannick, "Sequential PAttern Mining Using Bitmaps," *In Proc. of the 8th Int. Conf. on Knowledge Discovery and Data Mining*, 2002, Edmonton, Alberta, Canada, pp. 429-435.

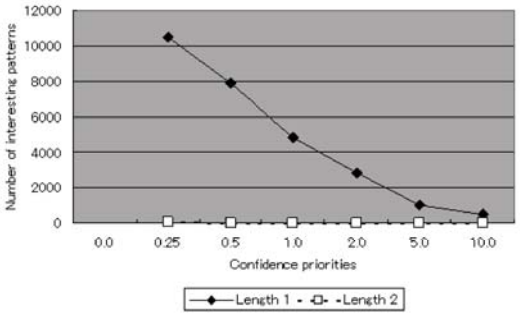
[4] J. Blanchard, F. Guillet, H. Briand, and R. Gras, "Assessing Rule Interestingness with a Probabilistic Measure of Deviation from Equilibrium," *in Proc. of the 11th Int. Sympo. on Applied Stochastic Models and Data Analysis*, 2005, Brest, France, pp. 191-200.

[5] S. Brin, R. Motwani, and C. Silverstein, "Beyond Market Baskets: Generalizing Association Rules to Correlations," *in Proc. of the 1997 ACM SIGMOD Int. Conf. on Management of Data*, 1997, Tucson, Arizona, USA, pp. 265-276.

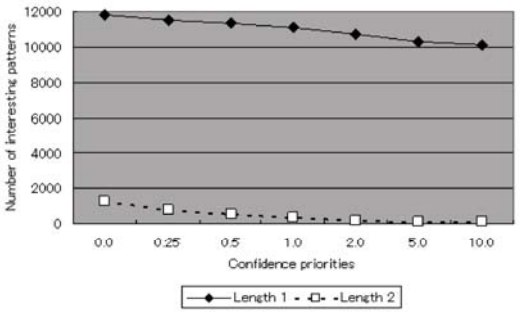
[6] M. N. Garofalakis, R. Rastogi, and K. Shim, "SPIRIT: Sequential Pattern Mining with Regular Expression Constraints," *in Proc. of the Very Large Data Bases Conf.*, 1999, Edinburgh, Scotland, UK, pp. 223-234.



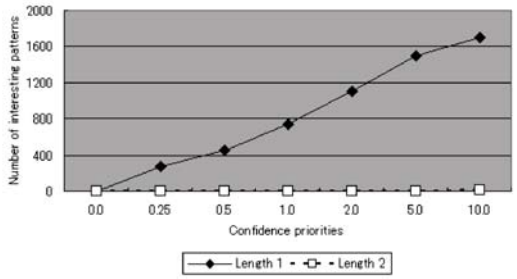
(a) Number of interesting patterns (1.00%)



(b) Number of interesting patterns (2.00%)



(c) Number of interesting patterns (common)



(d) Number of interesting patterns (different)

Fig. 7. Experimental results in the case of the medical examination data

[7] L. Geng and H. J. Hamilton, "Interestingness measures for data mining: A survey," *ACM Computing Surveys*, vol. 38, no. 3, article 9, 2006.
[8] Y. Ichimura, Y. Nakayama, M. Miyoshi, T. Akahane, T. Sekiguchi, Y. Fujiwara, "Text Mining System for Analysis of a Salesperson's Daily Reports," *in Proc. of Pacific Association for Computational Linguistics 2001*, 2001, Kitakyushu, Japan, pp. 127-135.
[9] V. Lavrenko, M. Schmill, D. Lawrie, P. Ogilvie, D. Jensen, J. Allan, "Mining of Concurrent Text and Time-Series," *in Proc. of the KDD-2000 Workshop on Text Mining*, 2000, Boston, Massachusetts, USA, pp. 37-44.
[10] B. Lent, R. Agrawal, R. Srikant, "Discovering Trends in Text Databases," *in Proc. of the 3rd Int. Conf. on Knowledge Discovery and Data Mining*, 1997, Newport Beach, California, USA, pp. 227-230.
[11] K. McGarry, "A Survey of Interestingness Measures for Knowledge Discovery," *the Knowledge Engineering Review*, vol. 20, no. 1, pp.39-61, 2005.
[12] J. Pei, J. Han, B. Mortazavi-Asl, H. Pinto, Q. Chen, U. Dayal, M. Hsu, "PrefixSpan: Mining Sequential Patterns Efficiently by Prefix-Projected Pattern Growth," *in Proc. of the 2001 Int. Conf. Data Engineering*, 2001, Heidelberg, Germany, pp. 215-224.
[13] J. Pei, J. Han, W. Wang, "Mining Sequential Patterns with Constraints in Large Databases," *in Proc. of the 11th ACM Int. Conf. on Information and Knowledge Management*, 2002, McLean, Virginia, USA, pp. 18-25.
[14] S. Sakurai, K. Ueno, R. Orihara, "Discovery of Time Series Event Patterns based on Time Constraints from Textual Data," *Int. J. of Computational Intelligence*, vol. 4, no. 2, pp. 144-151, 2008.
[15] K. Shimazu, A. Momma, and K. Furukawa, "Discovering Exceptional Information from Customer Inquiry by Association Rule Miner," *in Proc. of the 6th Int. Conf. on Discovery Science 2003*, 2003, Sapporo, Japan, pp. 269-282.
[16] A. Silberschatz and A. Tuzhilin, "What Makes Patterns Interesting in Knowledge Discovery Systems," *IEEE Trans. on Knowledge and Data Engineering*, vol. 8, no. 6, pp. 970-974, Dec., 1996.
[17] R. Srikant and R. Agrawal, "Mining Sequential Patterns: Generalizations and Performance Improvements," *in Proc. of the 5th Int. Conf. Extending Database Technology*, 1996, Avignon, France, pp. 3-17.
[18] E. Suzuki and J. M. Zytkow, "Unified Algorithm for Undirected Discovery of Exception Rules," *Int. J. of Intelligent Systems*, vol. 20, no. 7, pp. 673-691, July, 2005.
[19] R. Swan and D. Jensen, "TimeMines: Constructing Timelines with Statistical Models of Word Usage," *in Proc. of the KDD-2000 Workshop on Text Mining*, 2000, Boston, Massachusetts, USA, pp. 73-80.
[20] S. -J. Yen, "Mining Interesting Sequential Patterns for Intelligent Systems," *Int. J. of Intelligent Systems*, vol. 20, no. 1 , pp 73-87, Jan., 2005.
[21] M. J. Zaki, "Sequence Mining in Categorical Domains: Algorithms and Applications," *in Sequence Learning: Paradigms, Algorithms, and Applications, Lecture Notes in Computer Science*, vol. 1828, pp. 162-187, 2001.

Shigeaki Sakurai received an MS degree in mathematics and a Ph.D. degree in industrial administration from Tokyo University of Science, Japan, in 1991 and 2001, respectively. He was a Professional Engineer of Japan in the field of information engineering in 2004.

He is a research scientist at the System Engineering Laboratory, Corporate Research & Development Center, Toshiba Corporation. His research interests include data mining, soft computing, and web technology.

Dr. Sakurai is a member of IEICE, SOFT, and JSAI.

Youichi Kitahara received an MS degree in earth system science and technology from Kyushu University, Japan, in 2003.

He works at the System Engineering Laboratory, Corporate Research & Development Center, Toshiba Corporation. His research interests include data mining and machine learning.

Ryohei Orihara received a BS degree, an MS degree, and a Ph.D. degree in engineering from the University of Tsukuba, Japan, in 1986, 1988 and 1999, respectively.

He is the laboratory leader at the HumanCentric Laboratory, Corporate Research & Development Center, Toshiba Corporation. He is also a part-time associate professor at Tokyo Institute of Technology, Japan. His research interests include machine learning, creativity support systems, analogical reasoning, metaphor understanding, data mining and text mining.

Dr. Orihara is a member of IPSJ, JSAI, and JSSST.