Semantic Indexing and Temporal Rule Discovery for Time-series Satellite Images

Rie Honda
Kochi University
Akebono-cyo 2-5-1
Kochi, JAPAN 780-8520
honda@is.kochi-u.ac.jp

Hirokazu Takimito
Kochi University
Akebono-cyo 2-5-1
Kochi, JAPAN 780-8520
takimoto@is.kochi-u.ac.jp

Osamu Konishi
Kochi University
Akebono-cyo 2-5-1
Kochi, JAPAN 780-8520
konishi@is.kochi-u.ac.jp

ABSTRACT

Feature extraction and knowledge discovery from a large amount of image data such as remote sensing images have become highly required recent years. In this study, we present a framework for data mining from a set of time-series images including moving objects using clustering by self-organizing mapping(SOM) and extraction of time-dependent association rules. We applied this method to weather satellite cloud images taken by GMS-5 and evaluated its usefulness. The images are classified automatically by two-stage SOM. The results were examined and the cluster addresses were described in regard to season and prominent features such as typhoons or high-pressure masses. Sequential images are then transformed into a data series expressed by cluster addresses and time of occurrence, from which time-dependent association rules (simple serial rules) are extracted using a method for finding frequently co-occurring term-pairs from text. Semantic indexed data and extracted rules are stored in the database, which allows high-level queries by entering SQL through user interface, and thus supports knowledge discovery for domain-experts. We believe that this approach can be widely useful and applicable to knowledge discovery from an enormous amount of multimedia data, which includes unknown sequential patterns.

Categories and Subject Descriptors

H.2.8 [**Information Systems**]: Database Applications—*data mining, image database, scientific databases*
; H.3.3 [**Information Systems**]: Information Search and Retrieval—*clustering*; J.2 [**Computer Applications**]: Physical Science and Engineering—*earth and atomspheric science*

General Terms

DESIGN, EXPERIMENTATION, PERFORMANCE

Keywords

Satellite image database, clustering, self-organizing feature map, time dependent association rule, R-tree, content-based image retrieval, SQL query

1. INTRODUCTION

A huge amount of data has been stored in databases in the areas of business or science. Data mining or knowledge discovery from database(KDD) is a method for extracting unknown information such as rules and patterns from a large-scale database. The well-known data mining methods include decision tree, association rules[3], classification, clustering, and time-series analysis[1][2].

The process of the data mining is composed of the following six parts: (1) acquisition of input data, (2) selection of input data, (3) preprocessing, (4) transformation, (5) extraction of patterns, rules, etc., and (6) interpretation and evaluation of the results.

There are two main areas of in the data mining: one focused on business data and one focused on scientific data.

One of well-known cases of scientific data mining is the Sky Image Cataloging and Analysis tool (SKICAT) developed for the second Palomar Observatory Sky Survey[6]. They extracted astronomical body candidates from enormous raw images and classified them using a decision tree. In this process the researchers discovered both the classification rules and the novel bodies. Smyth et al.[8] and Burl et al.[7] have also reported a discovery system for venusian volcanoes based on synthetic aperture radar images taken by the spacecraft Magellan, which are very effective as recognition guides.

Image data such as satellite images and medical images often amount to several Tera bytes, thus manual and detailed analysis of these data becomes impractical[5]. Therefore an automated(or semi-automated) procedure

Proceedings of the International Workshop on Multimedia Data Mining (MDM/KDD'2000), in conjunction with ACM SIGKDD conference. Boston, USA, August 20, 2000
(O.R. Zaïane, S.J. Simoff, eds.)

to extract knowledge from these data should be included in the data mining from the image database.

In our recent studies[14][15], we have applied data mining methods such as clustering and association rules to a large number of the satellite weather images over the Japanese islands taken by Japanese stationary satellite GMS-5. These weather images are accumulated everyday and form a large amount of raw database.

Metrological events are considered to be chaotic phenomena in that an object such as a mass of cloud changes its position and form frequently. Furthermore they are time-sequential data such as video images.

Features of our studies applied to the weather images are summarized as follows:

(1) The application of data mining method to image classification and retrieval.

(2) Feature description from time-series data.

(3) Implementation of the result of classification as the user retrieval interface.

(4) Construction of the whole system as a domain-expert supporting system.

We describe an overview of the system in Section 2. A clustering algorithm for time-sequential images and its experimental results are described in Section 3. Section 4 describes the algorithm of extraction of time-dependent association rules and its experimental results. Section 5 describes details of the construction of the database by using R-tree and the results of its implementation. Section 6 provides a conclusion.

2. SYSTEM OVERVIEW

We constructed a weather image database that gathers the sequential changes of cloud images and the domain-expert analysis support system for these images. We characterize the system's images using clustering method (Section 3) and describe the image changes in terms of the sequential cluster numbers. Then we derive the time-dependent association rules from the sequential data (Section 4) and index them. The flow of this system is shown in Figure 1 and described as follows:

step 1 Clustering using a self-organizing map.

step 2 Generation of time-sequential data from a series of cluster addresses.

step 3 Extraction of time-dependent rules from the time-sequential data.

step 4 Indexing of rules and a series of cluster addresses by using R-tree, and construction of the database.

step 5 Searching for time-sequential variation patterns and browsing for the retrieved data in the form of animation.

The above-described process enables us to characterize enormous amount of images acquired at a certain time interval semi-automatically, and to retrieve the images by using the extracted rules. For example, this process enables queries like ”search for frequent events that occur between one typhoon and the next typhoon”, or ”search for a weather change such that a typhoon occurs within 10 days after a front and high pressure mass developed within the time interval of 5 days”.

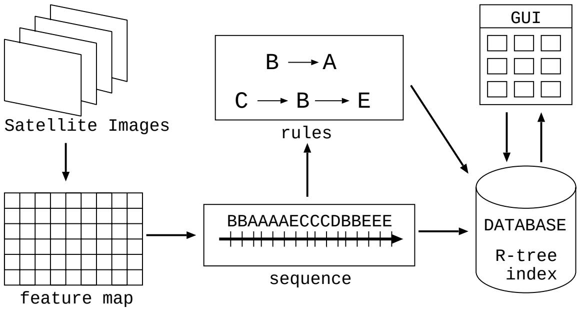


Figure 1: Overview of the system.

3. TIME-SEQUENTIAL DATA DESCRIPTION BY USING CLUSTERING

3.1 Data set description

Satellite weather images, taken by GMS-5 and received at the Institute of Industrial Science, Tokyo University, are archived at the Kochi University weather page (http://weather.is.kochi-u.ac.jp). The images used in this study are infrared band(IR3: moisture band, wavelength of 6.5-7 μm) images taken Japanese islands, which are of 640-pixels in width and 480-pixels in height. Each image is taken every hour, and about 9000 images are archived every year. Figure 2 shows an example of the image sequence.

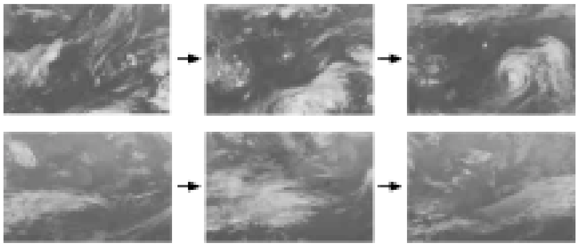


Figure 2: Example of weather image(GMS-5 IR3 band) sequence.

We considered that conventional image processing methods might be unable to detect moving objects such as the cloud masses that change their position and form during the time sequence. Thus we used the following SOM-based method for the automatic clustering of images by using the raster-like scanned image intensity vectors as the inputs.

3.2 Clustering and Kohonen's self-organizing map

Similarity analysis from full-text databases or image databases use sophisticated retrieval methods based on the indexing of space or the indexing of feature spaces. Clustering based on similarity is one of the extensions of these methods. When standard feature patterns are not given for the object data set, distance criteria in the feature spaces are used to divide the object set into the subset. This provides the rough structure to the given non-structured information.

Kohonen's self-organizing map (SOM)[9] is a paradigm which was suggested in 1990. The SOM is a two layer network that organizes a feature map by discovering feature relations based on input patterns through iterative non-supervised learning.

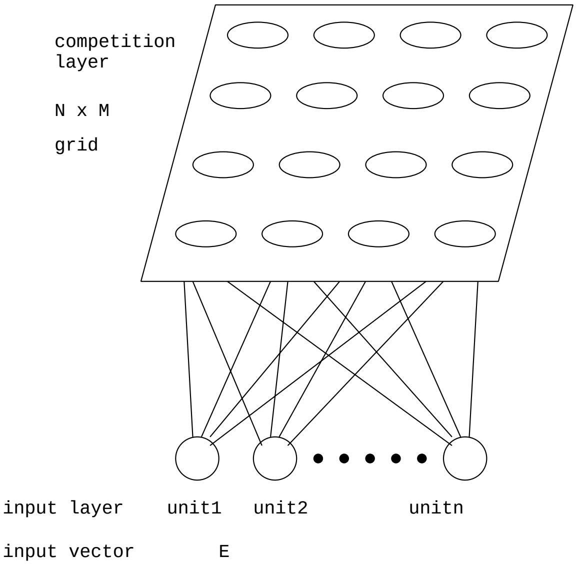


Figure 3: Basic structure of Kohonen's self-organizing map

Figure 3 presents basic schematic structure of Kohonen's self-organizing map. The network, a combination of the input layer and the competition layer, is trained through non-supervised learning. Each unit of the input layer has a vector whose components correspond to the input pattern elements.

The algorithm of the SOM is described as follows:

step 1 Let the input pattern vector $E \in R^n$ as,

$$E = [e_1, e_2, e_3, \cdots, e_n] \qquad (1)$$

step 2 Assume the weight of union from the input vector the to a unit i as

$$U_i = [u_{i1}, u_{i2}, u_{i3}, \cdots, u_{in}]. \qquad (2)$$

Initial values of u_{ij} are given randomly.

step 3 E is compared with all U_i, and the best matching node which has the smallest Euclidean distance $|E - U_i|$ is determined and signified by the subscript c,

$$c = \text{argmin}_i |E - U_i|. \qquad (3)$$

step 4 Weight vectors of the best matching node c and its neighbors,N_c, are adjusted to increase the similarity as follows,

$$u_{ij}^{new} = u_{ij}^{old} + \Delta u_{ij} \qquad (4)$$

where

$$\Delta u_{ij} = \begin{cases} \alpha(e_j - u_{ij}) & (i \in N_c) \\ 0 & (i \notin N_c) \end{cases} \qquad (5)$$

$$\alpha_t = \alpha_0 \left(1 - \frac{t}{T}\right) \qquad (6)$$

The α_t is the learning rate at the time of t iterations, α_0 is the initial leaning rate, and T is the total number of iterations.

step 5 The learning rate and the size of neighbor decreases as the learning proceeds.

The input signals E are classified into the activated (nearest) unit U_c of the input layer and projected onto the competition grids. The distance on the competition grids reflects the similarity between the patterns. After the training is completed, the obtained competition grids. i.e., the feature map, represents a natural relationship between the patterns of input signals entered into the network.

3.3 Clustering by two-stage SOM

Figure 4 represents the problem of clustering of weather images. Two images in Figure 4(a) are considered to have features similar to those of typhoon and a front, although their forms and positions are changed. When we take the input vectors simply as the raster-like scanned intensity vectors, these images are classified into the different groups based on the spatial variations of intensity. We considered that this difficulty is avoided by dividing the images into blocks as shown in Figure 4(b).

The procedure adopted here, named two-stage SOM, is described as follows:

stage 1 **Clustering of pattern cells**

step 1 All Images are divided into N×M blocks.

step 2 SOM generates the feature map, taking the each block's raster-like scanned intensity vectors as the input vectors.

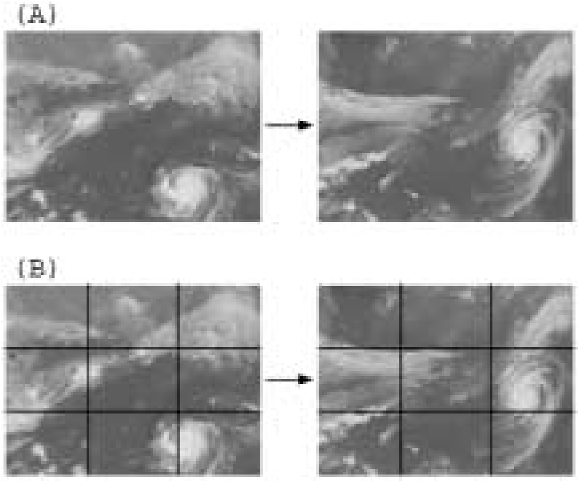


Figure 4: Problem for clustering of weather images.

- step 3 The SOM map cluster address is used to describe blocks of the original images. We refer to this characterized blocks as the pattern cells.
- stage 2 **Clustering of the images by using frequency histograms of pattern cells.**
- step 1 Each image is represented as the frequency histogram of the pattern cells.
- step 2 The feature map of SOM is generated by taking the frequency histogram of each image's pattern cell as the input.

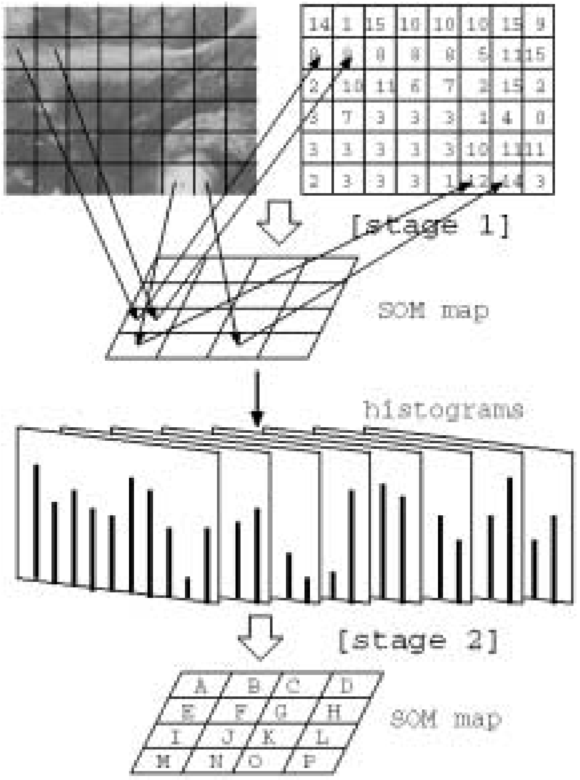
Extraction of frequency histogram of pattern cells in step 1 of stage 2 reduces the spatial information of blocks included in the images. Thus this process enables to classify time-series images which have similar objects at different positions as the same type of images.

Figure 5 schematically shows the above-described process of the two-stage SOM. The images that have similar objects are clustered into similar cells on the second stage feature map. Note that the difference in seasons is not distinguished at this point.

3.4 Result of experiments on clustering

In our experiments, we sampled GMS-5 IR3 images with 8 hour time intervals obtained between 1997 and 1998, and composed two data set for 1997 and 1998 which include 1044 and 966 images, respectively. We defined number of blocks for each image to be 12 × 16. The sizes of feature maps of both first stage SOM and second stage SOM are defined to be 4 × 4. Learning processes are iterated 8000-10000 times.

The results of the experiment show that images with similar features are classified into similar cells. To evaluate the accuracy of clustering quantitatively, we defined the following parameters,



14	1	15	10	10	10	15	9
0	0	0	0	0	5	11	15
2	10	11	6	7	2	15	2
3	7	3	3	3	1	4	0
3	1	3	3	3	10	11	11
2	3	3	3	1	12	14	3

A	B	C	D
E	F	G	H
I	J	K	L
M	N	O	P

Figure 5: Clustering of weather images by SOM.

$$Precision = B/(B+C), \qquad (7)$$

$$Recall = B/(A+B), \qquad (8)$$

where A is the number of the nonrelevant images that are classified into the cells, B is the number of the relevant images that are classified into the valid cell, and C is the number of the relevant images that are classified into the invalid cell.

Table 1 show the precision values for 1997 and 1998 to be 86.0% and 86.7%, respectively, and that the values of recall are 84.6% and 86.7%, respectively. These values indicate that the clustering of weather images by two-stage SOM can successfully learn the features of images and can classify them with a high degree of accuracy.

Table 1: Accuracies of clustering

year	Recall	Precision
1997	86.0%(876/1022)	84.6%(876/1044)
1998	86.7%(838/945)	86.7%(838/966)

Furthermore, we describe the semantic representation of clusters by specifying the season in which the clusters are observed, based on the frequency of each cluster every month, and by describing the representative object such as front or typhoon by means of visual observation of images in the cluster from a domain-expert like view. Table 2 shows the semantical descriptions of 1997 and 1998. The distribution of similar clusters for 1997 is different from 1998 since we performed the SOM leaning for these datasets independently. However, most of the groups are observed in both maps, thus the obtained result is meaningful even in the view of the domain-expert knowledge.

The obtained map is considered to be dependent on the block size of the original images and size of SOM map. Hierarchical division of each block in the original image by using standard deviation of intensity will be a solution to the determination of block sizes. The algorithm of Growing Hierarchical SOM[16], which is capable of growing both in terms of map size as well as the three-dimensional tree structure, will be effective for the adaptation of map size.

4. SEQUENTIAL ANALYSIS AND EXTRACTION OF TIME-DEPENDENT ASSOCIATION RULES

4.1 Association rules

Association rules are one of the key concepts of data mining[4]. An item i is defined to be a minimum element for extraction of rules. We define the set of items I and transaction database D as

$$I = [i_1, i_2, \cdots, i_m], D = [T_1, T_2, \cdots, T_n], (T_i \subseteq I), \quad (9)$$

where T_i is an element of the transaction database. A combination of k items is referred to as the item set with the length of k.

Then association rule is represented as

$$X \Rightarrow Y (X, Y \subset I, X \cap Y = \phi). \quad (10)$$

Evaluating parameters of the association rule $X \Rightarrow Y$, support and onfidence, are defined by

$$support(X \Rightarrow Y) = \frac{N(T_i \mid T_i \supseteq X \cup Y)}{N(D)}, \quad (11)$$

$$confidence(X \Rightarrow Y) = \frac{N(T_i \mid T_i \supseteq X \cup Y)}{N(T_i \mid T_i \supseteq X)}, \quad (12)$$

where N is the number of transactions in each condition. These parameters reflect the processing time and effectiveness of the rule.

Rule extraction is defined to find all rules that have larger confidence and support than the minimum threshold defined by users. The following process describes the extraction of association rules.

1. The item set that has larger support than the threshold is selected (referred to as the large item set).
2. The rules that have larger confidence than the threshold are selected from the large item set.

Table 2: Semantical description of each cluster. Cluster address is represented by the character of A, B, C, $\cdots$, P for the raster-like cells scanned from the upper left corner to the lower right corner.

1997

cluster address	season	prominent characteristics
A	spring summer	front, typhoon
B,C	spring autumn	high pressure in the west and low pressure in the east
D,H	spring autumn	band-like high-pressure
E	autumn	migratory anticyclone
F	spring autumn	front
G	autumn winter	linear clouds
I	summer	Pacific high pressure, front
J	spring summer	rainy season's front, typhoon
K,L	winter	winter type, whirl-like cloud
M	summer	Pacific high pressure, typhoon
N	spring summer	high pressure, typhoon
O	winter	cold front
P	spring autumn	migratory anticyclone

1998

cluster address	season	prominent characteristics
A,F,O	spring summer	front, typhoon
B	spring autumn	front, migratory anticyclone
C	summer	Pacific high-pressure
D	autumn	migratory anticyclone
E	spring autumn	band like high-pressure
G	spring summer	Pacific high pressure, front
H	spring summer	rainy season's front
I,K,N	winter	winter type, linear clouds(high pressure in the west and low pressure in the east)
J	summer	Pacific high pressure, front
L,M	winter	cold front
P	autumn winter	linear clouds

4.2 Time-series pattern analysis

Time-sequential data analysis is the method used to extract unknown patterns from time-sequential information, is related to the association rules, and is remarkable in the area of data mining. Episode rule[10][11] are known as one of those methods.

Episodes are defined as the event pairs in a certain time window. Events in time sequence are represented by (e, t), where e is the class of the event and t is its occurrence time. In Figure 4, an event sequence given by a string are represented by (E,31)(F,34)(A,35)(B,37)(C,38) $\cdots$ (D,49), where A, B, C are the event classes, and the number is the time of occurrence.

Figure 6 represents simple examples of episode rules such as those regarding serial episodes as "event B occurs after event A", parallel episodes such as "both events E and F occurs", or a combination of serial episodes and parallel episodes such as "event C occurs after event E and F".

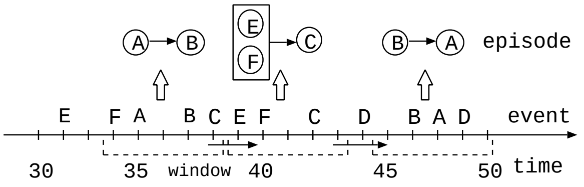


Figure 6: Example of event sequence and episode.

In order to define how closely these events occur, Mannila et al.[10] considered the time window that is shifted in an orderly manner in the sequence. Candidates of episodes are extracted as the co-occurring events in the time window. And combinations of events that have larger frequencies than the threshold frequency are determined to be episodes. A more flexible method that uses the minimal occurrence interval has also been suggested in [11].

4.3 Time-dependent association rule

In this study we present time-dependent association rules which modify the episode rules using the concept of cohesion, and represent local association rules such as "weather pattern B occurs after weather pattern A".

First we generate the sequential data of a weather pattern using cluster addresses as $(A, 1), (A, 2), (C, 3), \cdots$. We define the event as continuously occurring clusters. The event e_i in the sequence is then represented by

$$e_i =< C_i, S_{if}, T_{is}, T_{ie} > (i = 1, \cdot, \cdot, n), \qquad (13)$$

where C_i is the cluster addresses, S_{if} is the continuity, T_{is} is the starting time, and T_{ie} is the ending time. The sequence S is then represented by

$$S =< e_1, e_2, \cdots, e_n >, \qquad (14)$$

where n is the total number of the events in the sequence. Figure 7 shows a representation of event sequence in the case of $S_{if} \geq 2$.

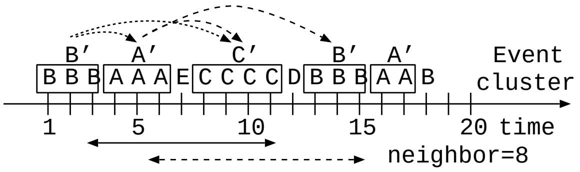


Figure 7: Example of description of cluster sequence, event sequence, and extraction of time-dependent association rules.

We extract the event pairs that occur closely in the sequence by introducing the neighborhood distance. The pattern change E is then represented by

$$E = \langle [e_i, e_j], neighbor \rangle (i = 1, \cdots, n-1, j = 1, \cdots, n), \qquad (15)$$

where $[e_i, e_j]$ represents a combination of the two events of e_i and e_j which satisfies $i < j$, and *neighbor* is the neighborhood distance.

Although *neighbor* is an idea similar with a time window in episode rules[10][11], we use this concept as the time interval necessary to extract only serial episodes such as $A \Rightarrow B$. We exclude parallel episode rules and combination of serial/parallel episode rules which are included in [10][11].

Furthermore we use the method of co-occurring term-pair [13] to evaluate the set of combinations of events which occurs closely and frequently in the local time window, and to extract them. The cohesion of the event e_i and e_j in a local time window is represented by

$$cohesion(e_i, e_j) = \frac{E_f(e_i, e_j)}{\sqrt{[f(e_i) \times f(e_j)]}}, \qquad (16)$$

where $f(e_i)$ and $f(e_j)$ are the frequencies of e_i and e_j, respectively, and $E_f(e_i, e_j)$ is the frequency of the co-occurrence of both e_i and e_j. The time-dependent association rules are extracted when the event pair has larger cohesion than the threshold.

The procedure of extraction of time-dependent association rule is shown schematically in Figure 8, and is described in the following:

step 1 The frequency of each event is determined (Fig. 8(1)).

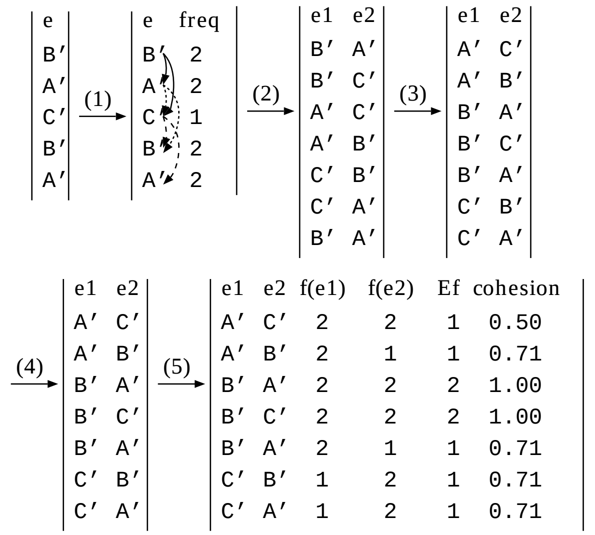
step 2 A combinational set of event pairs are determined as the candidates of rules, assuming the neighborhood distance (Figure 8(2)).

step 3 Event pairs are sorted lexicographically in regard to the first event (Fig. 8(3)).

step 4 Event pairs are sorted lexicographically in regard to the following event (Fig. 8(4)).

step 5 The candidates' frequency of co-occurrence and cohesion are calculated(Fig. 8(5)).

step 6 The event pairs that have larger cohesions than the threshold are extracted.



e
B′
A′
C′
B′
A′

e	freq
B′	2
A′	2
C′	1
B′	2
A′	2

e1	e2
B′	A′
B′	C′
A′	C′
A′	B′
C′	B′
C′	A′
B′	A′

e1	e2
A′	C′
A′	B′
B′	A′
B′	C′
B′	A′
C′	B′
C′	A′

e1	e2
A′	C′
A′	B′
B′	A′
B′	C′
B′	A′
C′	B′
C′	A′

e1	e2	f(e1)	f(e2)	Ef	cohesion
A′	C′	2	2	1	0.50
A′	B′	2	1	1	0.71
B′	A′	2	2	2	1.00
B′	C′	2	2	2	1.00
B′	A′	2	1	1	0.71
C′	B′	1	2	1	0.71
C′	A′	1	2	1	0.71

Figure 8: Procedure of a extraction of time-dependent association rule, where $e1$ and $e2$ are the first event and the following event, respectively, $f(e1)$ and $f(e2)$ are the frequencies of $e1$ and $e2$, respectively, E_f is the frequency of co-occurrence, and *cohesion* is the strength of cohesion between $e1$ and $e2$. The neighborhood distance is taken to be 8 in this case.

Strongly correlated event pairs in *neighbor* have large *cohesion* even if each event occurs less frequently. Inversely, weakly correlated event pairs have small *cohesion* even if each event occurs very frequently.

4.4 Result of experiments regarding time-dependent association rules

We performed the experiment by applying the above-described time-dependent association rule to the result of the clustering described in 3.4. Here we take the threshold of cohesion as 0.4, and *neighbor* ranging from 10 to 50. Since we sampled data every 8 hours, the virtual length of *neighbor* is between 3.3 days and 16.7 days.

Table 3 shows the relationship between *neighbor* and the number of extracted rules. Although the assessment of the context of the extracted rules is ongoing, the result suggests the similar numbers of rules are extracted from the different year's data set, which indicates that our present method is useful and robust.

Table 3: Relationship between *neighbor* and number of rules.

neighbor	10	20	30	40	50
number of rules(1997)	17	63	116	165	207
number of rules(1998)	7	50	98	166	218

5. CONSTRUCTION OF WEATHER IMAGE DATABASE

We constructed a weather image database which retrieves the above-described characteristics of weather images, visualizes time-dependent variation pattern, and supports the analysis and scientific discovery by domain-experts.

First we indexed the sequential time data by using events and time-dependent association rules, and constructed a weather satellite image database which contains index information regarding patterns such as time variations in weather.

5.1 Definition of attributes

We stored weather patterns extracted in the experiments in the following three tables: "series", "date_id", and "e_series" that represent contexts of time-dependent rules, the relationship between the observation date and image ID, and the contents of time-dependent rule candidates (those obtained in step 5 in 4.3), respectively.

Table 4: List of three table "series", "data_id", "e_series"

(a) "series" that represents time-dependent rule in which l_term is the cluster number of left term, r_term is the right term, location is the reference to the R-tree data(rectangular), and first and last are the image ID of the l_term starting point and r_term ending point, respectively .)

l_term	r_term	cohesion	location	first	last
int	int	float	box	int	int

(b) "date_id" that indicates the relationship between the observation date *date* and image ID *id*.

id	date
int	int

(c) "e_series" that indicates the candidate of time-dependent event rules, where *term* is the cluster number, *first* and *last* are the image ID of the starting point and ending point of *term*, respectively.

term	first	last
int	int	int

5.1.1 *Indexing by using R-tree*

We indexed the image IDs at the starting and ending point of the obtained pattern using R-tree. As shown in Figure 9, spring encloses March to May, and summer encloses June to August. Taking note of this relation, we index the enclosure relation between seasons and months, and index the starting and the ending times of variation patterns. This allows each variation pattern to contain an index which includes the enclosure relations by month or season as keys.

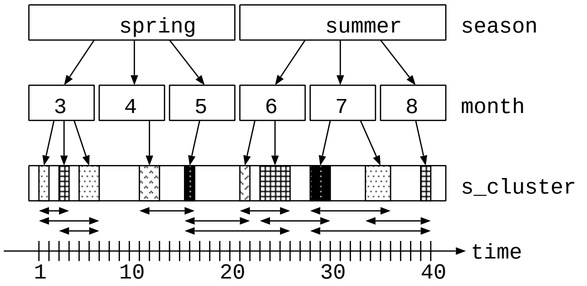


Figure 9: Indexing by using R-tree, remarking at the continuing sequence.

5.2 Query by SQL

Rule storage in the database enables the retrieval of the various queries by using SQL statements. We show examples of the queries and corresponding SQL statement[1] in the following:

”Search for typhoons that occurred within 20 days after July 16th, 1997.”

```
select first, last, t1.date, t2.date
from series, date_id t1, date_id t2
where t1.date = 97071617 and t1.id = first
and t2.id = last and(r_term = 0 or r_term =
9 or r_term = 12 or r_term = 13) and location
@ '((570, 0),
(630,15))'::box
```

”Search for weather changes between one typhoon and the another.”

```
select first, last, t1.date, t2.date
from series, date_id t1, date_id t2
where(l_term = 0 or l_term = 9 or l_term =
12 or l_term = 13) and(r_term = 0 or r_term
= 9 or r_term = 12 or r_term = 13) and t1.id
= first and t2.id = last
```

or

```
select t1.first, t2.last, t1.date, t2.date
from e_series t1, e_series t2, date_id t1,
date_id t2
where(t1.term = 0 or t1.term = 9 or t1.term
= 12 or t1.term = 13) and(t2.term = 0 or t2.term
= 9 or t2.term = 12 or t2.term = 13) and t1.id
= t1.first and t2.id = t2.last and t1.first
< t2.first order by t1.first
```

”Search for weather patterns in which typhoon occurs within 10 days after the development of front and typhoon during 5 days.”

```
select t1.first, t2.last, t1.date, t2.date
from series t1, e_series t2, date_id t1,
date_id t2
where (t1.l_term = 0 or t1.l_term = 5 or
t1.l_term = 8 or t1.l_term = 9 or t1.l_term
= 14)
and (t1.r_term = 1 or t1.r_term = 2 or t1.r_term
= 3
or t1.r_term = 4 or t1.r_term = 7 or t1.r_term
= 8
or t1.r_term = 12 or t1.r_term = 13 or t1.r_term
= 15)
and t1.first >=(t1.last - 15) and t1.id = first
and
t2.id = last and(t2.term = 0 or t2.term = 9
or
t2.term = 12 or t2.term = 13) and
t1.first >=(t2.last - 30) and t1.last <= t2.last
```

5.3 Result of implementation

Figure 10 shows the browse page[2] of the system which retrieves weather images using R-tree index. Entering the SQL in the upper frame performs retrievals. This example shows the results of query: ”Is there any weather pattern in which a typhoon occurred in 10 days after the development of front and typhoon during 5 days”. Seven periods are retrieved and listed in the lower left frame as the result, and the weather variation in these periods is shown as an animation in the lower right frame.

The problem of this method is that the accuracy of clustering and the semantical description of clusters changes the retrieval results significantly. Interactive processing interface, such as adjustment of the sample data or assumed parameters with metrological experts who are potential users, are required to solve this problem.

6. CONCLUSION

We applied clustering and time-dependent association rules to a large-scale content-based image database of weather satellite images. Each image is divided into $N \times M$ blocks and automatically classified by two-stage SOM. We also extracted unknown rules from time- sequential data expressed by a sequence of cluster addresses by using time-dependent association rules. Furthermore, we developed a knowledge discovery support system for domain experts, which retrieves image sequences using extracted events and association rules.

[1] Here cluster addressees are represented by numbers ranging from 0 to 15 instead of characters A-P in Table 2.

[2] http://zeus.is.kochi-u.ac.jp/~takimoto/java/servlets/index6e.
html

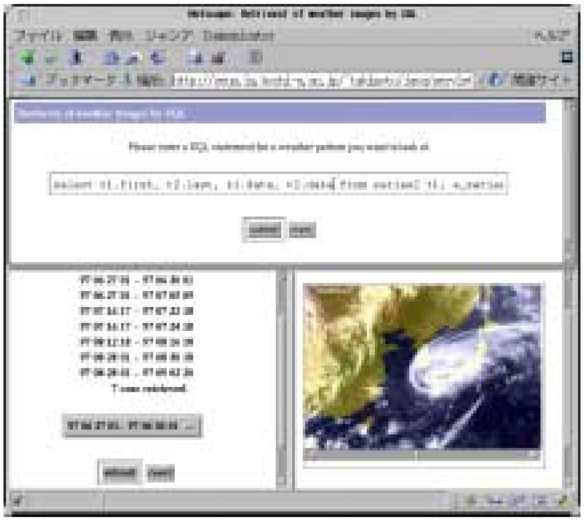


Figure 10: Example of the result of retrieval result from sequential image data.

From the perspective that high-level queries make the analysis easier, we stored the extracted rules in the database to admit sophisticated queries described by SQL. The retrieval responses to various queries shows the usefulness of this approach.

The framework presented in this study, clustering ⇒ transformation into time-sequential data ⇒ extraction of time-dependent association rules, is considered to be useful in managing enormous multimedia datasets which include sequential patterns such as video information and audio information.

7. ACKNOWLEDGEMENTS

The authors are grateful to Prof. T. Kikuchi for offering us well-prepared GMS-5 images. This research is partly supported by grants-in-aid for intensive research(A)(1)(Project 10143102) from the Ministry of Education, Science, and Culture of Japan.

8. REFERENCES

[1] J. Fomg,(Edt.). Data mining, data warehousing & client/server databases. In *Proceedings 8th International Database Workshop*, Springer, 1997.

[2] A. F. Alex and H. L. Simon. *Mining very large databases with parallel processing*. Kluwer Academic Publishers, 1998.

[3] R. Agrawal, T. Imelinski and A. Swani. Mining in association rules between sets of items in large database. In *Proc. ACM SIGMOD International Conference*, pages 207–216, 1993.

[4] R. Agrawal and R. Srikant. Fast Algorithms for mining association rules. In *Proceedings of 20th International Conference on VLDB*, pages 487–499, 1994.

[5] O. R. Zaiane, J. Han, Z. N. Li, J. Y. Chiang and S. Chee. Multimedia-miner : a system prototype for multimedia data mining. At *Proceedings ACM-SIGMOD Conference on Management of Data*, system demo, 1998.

[6] U. M. Fayyad, S. G. Djorgovski and N. Weir. Automatic the analysis and cataloging of sky surveys. *Advances in Knowledge Discovery and Data Mining*, pages 471–493 , AAAI Press/MIT Press, 1996.

[7] M. C. Burl, L. Asker, P. Smyth, U. M. Fayyad, P. Perona, L. Crumpler and J. Aubele. Learning to recognize volcanoes on venus. *machine learning*, 30(2/3):165–195, February, 1998.

[8] P. Smyth, M. C. Burl and U. M. Fayyad. Modeling subjective uncertainty in image annotation. In *Advances in Knowledge Discovery and Data Mining*, pages 517–539. AAAI Press/MIT Press, 1996.

[9] T. Kohonen. *Self-organizing maps*. Springer, 1995.

[10] H. Mannila, H. Tovinen and A. I. Verkano. Discovering frequent episodes in sequences. In *First International Conference on Knowledge Discovery and Data Mining(KDD'95)*, pages 210–215 , AAAI Press, 1995.

[11] H. Mannila, H. Tovinen. Discovering generalized episodes using minimal occurrences. In *Proceeding of the Second International Conference on Knowledge Discovery and Data Mining(KDD'96)*, pages 146–151, AAAI Press, 1996.

[12] T. N. Raymond, J. Han. Efficient and effective clustering methods for spatial data mining. In *Proceeding of 20th VLDB Conference*, Santiago, Chile, 1994.

[13] O. Konishi. A statistically build knowledge based terminology construction system (in Japanese). *Transaction of Information Processing Society of Japan*!$30(2):179–189, 1989.

[14] K. Katayama and O. Konishi. Construction satellite image databases for supporting knowledge discovery(in Japanese). *Transaction of Information Processing Society of Japan*, 40(SIG5(TOD2)):69–78, 1999.

[15] K. Katayama. and O. Konishi. Discovering co-occurencing patterns in event sequences (in Japanese). *DEWS'99*, 1999.

[16] D. Merkl and A. Rauber. Uncovering the hierarchical structure of text archives by using unsupervised neural network with adaptive architecture. In *PAKDD 2000*, pages 384–395, 2000.