Temporal Rule Discovery for Time-Series Satellite Images and Integration with RDB

Rie Honda[1] and Osamu Konishi[1]

Department of Mathematics and Information Science
Kochi University, Akebono-cyo 2-5-1 Kochi, 780-8520, JAPAN
{honda,konishi}@is.kochi-u.ac.jp
http://www.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, a framework for data mining from a set of time-series images including moving objects was presented. Time-series images are transformed into time-series cluster addresses by using clustering by two-stage SOM (Self-organizing map) and time-dependent association rules were extracted from it. Semantically indexed data and extracted rules are stored in the object-relational database, which allows high-level queries by entering SQL through the user interface. This method was applied to weather satellite cloud images taken by GMS-5 and its usefulness was evaluated.

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 [1] [2], classification, clustering, and time-series analysis [3], and there are some successful application studies for astronomical images such as SKICAT [5] and JARtool [4].

In our recent studies [7] [8], we have applied data mining methods such as clustering and association rules to a large number of time-series satellite weather images over the Japanese islands. Meteorological events are considered to be chaotic phenomena in that an object such as a mass of cloud changes its position and form continuously, and thus appropriate for experiments of spatial temporal data mining.

Features of our studies applied to the weather images are summarized as follows: application of data mining method to image classification and retrieval, feature description from time-series data, implementation of the result of classification as the user retrieval interface, and construction of the whole system as a domain-expert KDD supporting system.

We describe an overview of the system in Sect. 2. A clustering algorithm for time-sequential images and its experimental results are described in Sect. 3.

L. De Raedt and A. Siebes (Eds.): PKDD 2001, LNAI 2168, pp. 204–215, 2001.

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 aimed to construct the domain-expert analysis support system for these images. The flow of this system is shown in Fig. 1 and described as follows:

1 Clustering of frame images using a self-organizing map.
2 Transformation of time-series images into cluster address sequence.
3 Extraction of events and time-dependent rules from the time-sequential cluster addresses.
4 Indexing of events and rules by R-tree, and integration with the database.
5 Searching for time-sequential variation patterns and browsing of the retrieved data in the form of animation.

The above-described framework enables us to characterize enormous amount of images acquired at a regular time interval semi-automatically, and to retrieve the images by using the extracted rules. For example, this framework 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".

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 of Tokyo University, are archived at the Kochi University weather page (http://weather.is.kochi-u.ac.jp). In this study, we used infrared band (IR1) images around Japanese islands, which reflect the cloud distribution very well. The size of image is 640-pixels in width and 480-pixels in height. Each image is taken every hour, and about 9000 images are archived every year.

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

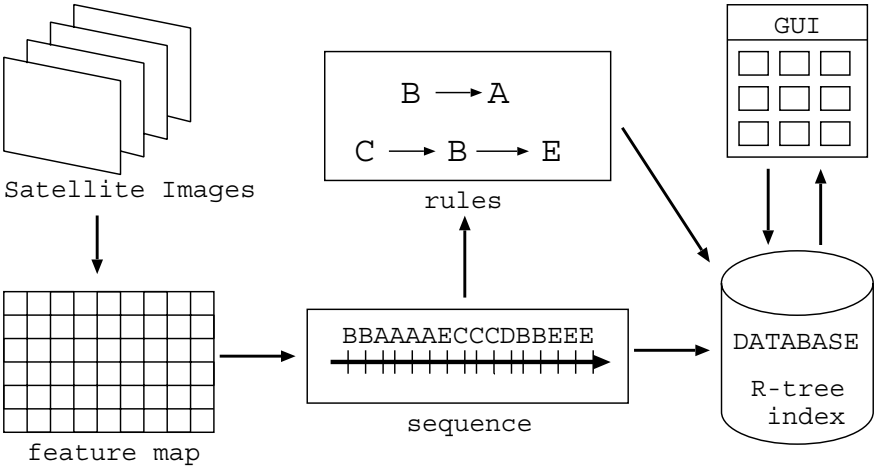

Fig. 1. Overview of the system.

3.2 Clustering and Kohonen's Self-Organizing Map

Kohonen's self-organizing map (SOM) [9] is a paradigm which was suggested in 1990, which has been widely used to provide a rough structure to a given non-structured information. The SOM is a two-layer network composed of a combination of the input layer and the competition layer that is trained through iterative non-supervised learning.

Each unit of the input layer and the competition layer has a vector whose components correspond to the input pattern elements. The algorithm of the SOM is described as follows: Let the input pattern vector $V \in R^n$ as $V = [v_1, v_2, v_3, \cdots, v_n]$, and the weight of union from the input vector to a unit i as $U_i = [u_{i1}, u_{i2}, u_{i3}, \cdots, u_{in}]$. Initial values of u_{ij} are given randomly. V is compared with all U_i, and the best matching unit which has the smallest Euclidean distance is determined and signified by the subscript c.

$$c = \text{argmin}|V - U_i|. \tag{1}$$

Weight vectors of the unit c and its neighbors N_c, which is the area of $N \times N$ units around the unit c, are adjusted to increase the similarity as follows,

$$u_{ij}^{new} = \begin{cases} u_{ij}^{old} + \alpha(t)(v_j - u_{ij}) & (i \in N_c) \\ u_{ij}^{old} & (i \notin N_c), \end{cases} \tag{2}$$

where

$$\alpha(t) = \alpha_0 \left(1 - t/T\right), \tag{3}$$

$$N(t) = N_0 \left(1 - t/T\right). \tag{4}$$

The $\alpha(t)$ $N(t)$ are the learning rate and the size of neighbors at the time of t iterations, respectively, α_0 and N_0 is the initial learning rate and the size of

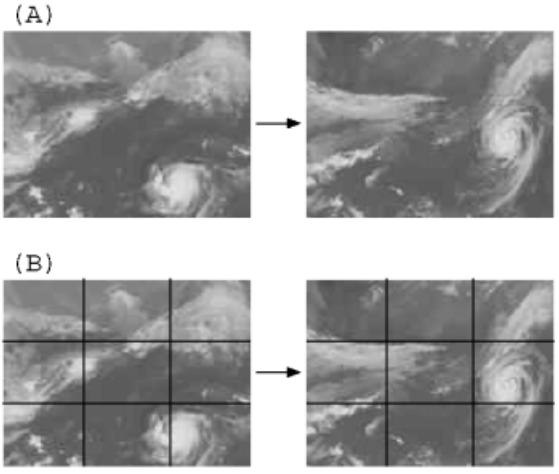

Fig. 2. Problem for clustering of weather images.

neighbors, and T is the total number of iterations of learning. The learning rate and the size of neighbor decreases as the learning proceeds to stabilize it.

The input signals V are classified into the activated (closest) unit U_c 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 represent a natural relationship between the patterns of input signals entered into the network. Hereafter we call the competition grids obtained after the learning as the feature map.

3.3 Clustering by Two-Stage SOM

Figure 2 represents the problem of clustering of weather images. Two images in Fig. 2(A) are considered to have the same features of a typhoon and a front, although their forms and positions change as time proceeds. When we take the input vectors simply as the raster image 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 Fig. 2(B).

The procedure adopted here, named two-stage SOM, is shown in Fig. 3 schematically and described as follows:

stage 1 **Clustering of pattern cells**
- step 1 All Images are divided into $N \times M$ blocks.
- step 2 Learning by SOM is performed by entering the each block's raster image intensity vector as the input vector successively.
- step 3 Each block of the original images are projected onto the first SOM feature map and characterized with the closest unit address. We refer to this characterized blocks as the pattern cells.

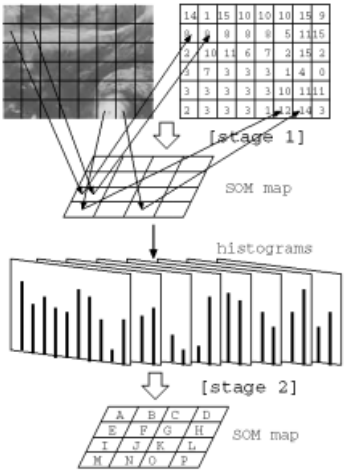

Fig. 3. Clustering of weather images by SOM.

stage 2 **Clustering of the images by using frequency histograms of pattern cells.**

step 1 Each image is transformed into the frequency histogram of the pattern cells.

step 2 The feature map of SOM is learned by entering each image's pattern cell frequency histogram as the input vector.

step 3 Each images are projected onto the second SOM feature map and characterized with the closest unit address.

Although the information of spatial distribution of pattern is lost by transforming images into frequency histograms of pattern cells (in step 1 of stage 2), this enables flexible classification of time-series images which have similar objects at different positions as shown in Fig. 2b as the same type of images.

Hereafter we refer the unit as the cluster, and express the cluster addresses by the characters of A, B, C, $\cdots$, P in the raster-scan order from the upper left corner to the lower right corner of 4×4 feature map.

3.4 Result of Experiments on Clustering

In the experiments, we sampled GMS-5 IR1 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, considering the typical size of cloud masses. The sizes of feature maps of both the first stage SOM and the second stage SOM are defined to be 4×4, which are determined by trial and error. Learning processes are iterated 8000-10000 times.

The result of the experiment shows that images including similar features are distributed into similar clusters. We describe clusters semantically by specifying

Table 1. Semantical description of each cluster

cluster address		season	prominent characteristics
1997	1998		
A	A,F,O	spring, summer	front, typhoon
J	H	spring, summer	rainy season's front, typhoon
N		spring, summer	high pressure, typhoon
B,C		spring autumn and low pressure in the east	high pressure in the west
D,H	E	spring, autumn	band-like high-pressure
F	B	spring, autumn	front
P	B	spring, autumn	migratory anticyclone
I	J	summer	Pacific high pressure, front
M		summer	Pacific high pressure, typhoon
	C	summer	Pacific high-pressure
E	D	autumn	migratory anticyclone
G	P	autumn, winter	linear clouds
K,L	L,M	winter or linear cloud	winter type, whirl-like
O	I,K,N	winter	cold front

Table 2. Accuracies of clustering

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

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 in a domain-expert like view. Table 1 shows the semantical descriptions of clusters for 1997 and 1998. The description of each clusters for 1998 is different from that for 1997 since we performed the SOM learning for these data sets 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.

To evaluate the accuracy of clustering quantitatively, we defined the following parameters,

$$Recall = A/(A+B), Precision = A/(A+C), \tag{5}$$

where A is the number of the relevant images classified into the cluster, B is the number of the relevant images classified into the other cluster, and C is the number of the nonrelevant images classified into the cluster. Relevance of images are evaluated by classifying the images visually.

Table 2 shows the values of recall for 1997 and 1998 to be 86.0% and 86.7%, respectively, and that the values of precision are 84.6% and 86.7%, respectively.

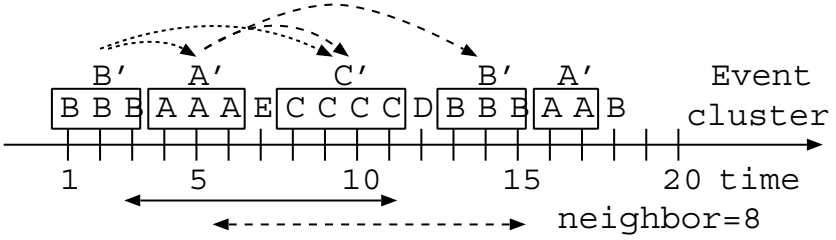

Fig. 4. Example of description of cluster sequence, event sequence, and extraction of time-dependent association rules.

These values indicate that two-stage SOM can successfully learn the features of weather images and can classify them with a high accuracy.

4 Time-Sequential Analysis and Extraction of Time-Dependent Association Rules

4.1 Time-Dependent Association Rule

In this study we extract time-dependent association rules such as "weather pattern B occurs after weather pattern A", which modify the episode rules [11] [12] using the concept of cohesion to evaluate its significance.

First we express the sequence of a weather pattern by $(A,1),(A,2),(C,3),\cdots$ where each component is a pair of cluster address of image (obtained by SOM) and its observation time. Then we define the event e_i in the sequence as continuously occurring clusters, which is expressed by

$$e_i =< C_i, S_{if}, TS_i, TE_i > \ (i = 1, \cdots, n), \tag{6}$$

where C_i is the cluster address, S_{if} is the continuity, TS_i is the starting time, and TE_i is the ending time. The sequence S is then represented by

$$S =< e_1, e_2, \cdots, e_n >, \tag{7}$$

where n is the total number of the events in the sequence. Figure 4 shows a representation of event sequence in the case of $S_{if} \geq 2$.

We extract the event pairs that occur closely in the sequence by introducing a local time window. Assuming the local time window with the length of *neighbor*, the simple local variation pattern E is represented by

$$E = \langle [e_i, e_j], neighbor \rangle \ (i \in \{1, \cdots, n-1\}, j \in \{2, \cdots, n\}), \tag{8}$$

where $[e_i, e_j]$ is a combination of the two events e_i and e_j which satisfies $i < j$ and $TS_j - TE_i < neighbor$.

Although *neighbor* is an idea similar with a time window in episode rules [11] [12], we use this concept to extract only serial episodes such as $A \Rightarrow B$,

excluding parallel episode rules and combination of serial/parallel episode rules which are included in [11] [12].

Furthermore we use the method of co-occurring term-pair for document analysis [10] to evaluate strength of correlation of event pairs which occurs in the local time window and to extract the prominent pairs as rules. The cohesion of the event e_i and e_j in a local time window is defined by

$$cohesion(e_i, e_j) = \frac{E_f(e_i, e_j)}{\sqrt{[f(e_i) \times f(e_j)]}}, \tag{9}$$

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 in a local time window. 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 rules in each local time window with the length of *neighbor* is described in the following:

step 1 The frequency of each event $f(e)$ in a local time window is determined.
step 2 A combinational set of event pairs in a local time window are listed as rule candidates.
step 3 Candidate pairs are sorted lexicographically in regard to the first event and then the following event.
step 4 The same event pairs are bound, co-occurrence frequency of each candidate pair $E_f(e_i, e_j)$ is counted, and cohesion are calculated.
step 5 The event pairs that have larger cohesions than the threshold are extracted as rules.

It should be noted that extraction is performed for each local time window by sliding its position.

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

4.2 Result of Experiments Regarding Time-Dependent Association Rules

We applied the above-described time-dependent association rule extraction to the sequence of cluster address obtained in 3.4. Here we take the threshold of cohesion of 0.4 and *neighbor* ranging from 10 to 50. Since we sampled images every 8 hours, the virtual length of *neighbor* is between 3.3 days and 16.7 days.

Table 3 shows the relationship between the size of *neighbor* and the number of extracted rules. Although the assessment of the contents of the extracted rules and development of its user-interface are ongoing issues, 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 Integration of Extracted Rules and the Relational Database

We integrate image sequences and the extracted rules with the relational database and construct the system which supports analysis and discovery by domain-experts. Here we index the time sequences by using R-tree [6] to enable fast query operation.

5.1 Indexing by R-Tree

As shown in Fig. 5, there is a natural hierarchical enclosure relations between time sequences such as year, season, month, rule, and event. By using the method of R-tree [6], we can express these time sequences by the minimum bounding rectangles (defined by the staring time and the ending time of sequence) and store them into the hierarchical tree which reflects the enclosure relation. This enables fast query operation of weather patterns by using month or seasons as the search key.

5.2 Definition of Attributes

We stored extracted patterns in the following three tables: "series(l_term, r_term, cohesion, location, first, last)[1], "date_id (id, date), and "_series (term, first, last)[2] that represent contents of time-dependent rules, the relationship between image ID and the observation time, and the contents of time-dependent rule components (events), respectively.

5.3 Query by SQL

Storing extracted patterns in the database enables the secondary retrieval of the various complex patterns by using SQL statements. We show an example of complex queries and the corresponding SQL statement in the following:

[1] "l_term" and "r_term" are the cluster addresses of the first event and the second event of extracted rules, respectively, "location" is the reference to the R-tree rectangles, and "first" and "last" are the image IDs of the "l_term" starting point and the "r_term" ending point, respectively.

[2] "term" is the cluster number of the event, "first" and "last" are the image IDs of the starting point and the ending point of "term", respectively.

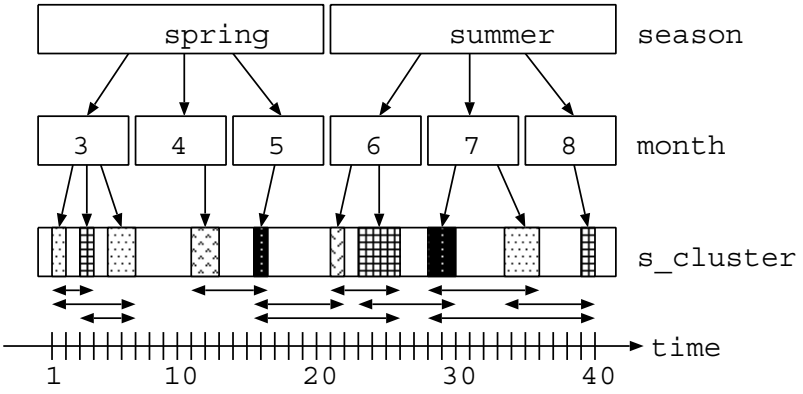

Fig. 5. Indexing by using R-tree, remarking at the continuing sequence. Arrows at the bottom represent minimum bounding boxes of rules.

"Search for a weather change in 1997 such that a typhoon occurred within 10 days after a front and a successive high pressure mass developed within the time interval of 5 days."

```
select t1.first, t2.last, t3.date, t4.date[3]
from series t1 , e_series t2 , date_id t3 , date_id t4
where ( t1.l_term = "A" or t1.l_term = "F" or t1.l_term = "I" or
t1.l_term = "J" or t1.l_term = "O") and ( t1.r_term = "B" or
t1.r_term = "C" or t1.r_term = "D" or t1.r_term = "H" or t1.r_term =
"I" or t1.r_term = "M" or t1.r_term = "N" ) and ( t2.term = "A" or
t2.term = "J" or t2.term = "M" or t2.term = "N") and (t1.last-t1.first
<15) and (t2.last- t1.first <30) and (t1.last <= t2.last) and t3.id
= t1.first and t4.id = t2.last
```

5.4 Result of Implementation and Issues in the Future Work

Figure 6 shows an example of the user interface of integrated KDD support system for weather information. Here we can retrieve the weather pattern by using the season, the first event and the second event as the search keys. Matched sequences are listed in the lower left frame, and by selecting one in the list, the corresponding weather variation is shown as an animation in the lower right frame. To deal with much more complex queries, we also prepare the user interface which accepts SQL query directly.

In this system, however, users are unable to operate the process of primary knowledge extraction by changing parameters such as the size of SOM, cohesion threshold, and the size of the local time window. There are two approaches to

[3] Note that time interval 1 in this SQL statement corresponds to 8 hours, and capital alphabets indicate the cluster addresses described in table 1.

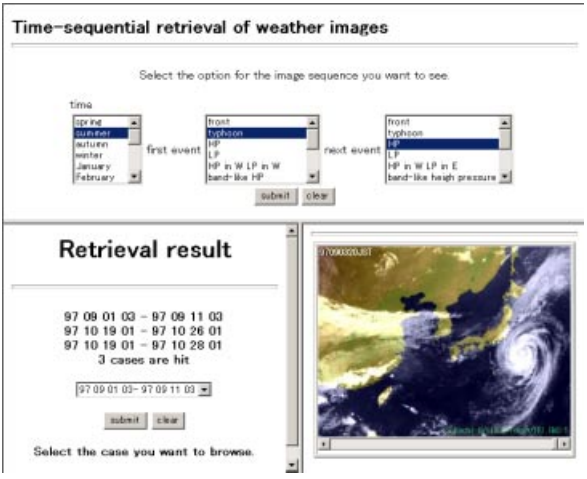

Fig. 6. Example of the result of retrieval from sequential image data.

solve this problem: one is incorporation of the optimization process of these parameters[4] and another is improvement of interactiveness of the user interface. Examination on both approaches will be one of most significant issues in the future work. Also we consider designing of the user interface to stimulate expert's natural discovery is also important. Furthermore, improvement of time sequential pattern analysis besides simple rule extraction will be significant to deal with temporal patterns more meaningful for domain-experts including prediction.

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 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. 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 also useful in managing enormous multimedia data sets which include

[4] For examples, the algorithm of growing hierarchical SOM [13], 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. We would like to examine this algorithm in the future work.

sequential patterns such as video and audio information or result of numerical simulation.

Acknowledgements. The authors are grateful to Prof. T. Kikuchi for offering us well-prepared GMS-5 images, We also thank to K. Katayama, and H. Takimoto for their past contribution. 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.

References

1. Agrawal, R., Imelinski, T., Swani, A.: Mining in association rules between sets of items in large database. Proc. ACM SIGMOD International Conference (1993) 207–216
2. Agrawal, R. and Srikant, R.: Fast Algorithms for mining association rules. Proceedings of 20th International Conference on VLDB (1994) 487–499.
3. Alex, A.F., Simon, H.L.: Mining very large databases with parallel processing. Kluwer Academic Publishers (1998)
4. Burl, M.C., Asker, L., Smyth, P., Fayyad, U.M., Perona, P., Crumpler, L., Aubele, J.: Learning to recognize volcanos on Venus. Machine Learning, Vol. 30, (2/3) (1998) 165–195
5. Fayyad, U.M., Djorgovski, S.G., Weir, N.: Automatic the analysis and cataloging of sky surveys. Advances in Knowledge Discovery and Data Mining, AAAI Press/MIT Press (1996) 471–493
6. Guttman, A.: R-trees: a dynamic index structure for spatial searching. Proc. ACM SIGMOD International Conference (1984) 47–57
7. Katayama, K., Konishi, O.: Construction satellite image databases for supporting knowledge discovery(in Japanese). Transaction of Information Processing Society of Japan, Vol. 40, SIG5(TOD2) (1999) 69–78
8. Katayama, K., Konishi, O.: Discovering co-occurencing patterns in event sequences (in Japanese). DEWS'99 (1999)
9. Kohonen, T.: Self-organizing maps. Springer (1995)
10. Konishi, O.: A statistically build knowledge based terminology construction system (in Japanese). Transaction of Information Processing Society of Japan, Vol. 30, 2 (1989) 179–189
11. Mannila, H., H. Tovinen and A. I. Verkano. Discovering frequent episodes in sequences. In First International Conference on Knowledge Discovery and Data Mining(KDD'95), AAAI Press (1995) 210–215
12. Mannila, H., Tovinen, H.: Discovering generalized episodes using minimal occurrences. In Proceeding of the Second International Conference on Knowledge Discovery and Data Mining(KDD'96), AAAI Press (1996) 146–151
13. Merkl, D. Rauber, A.: Uncovering the hierarchical structure of text archives by using unsupervised neural network with adaptive architecture. In PAKDD 2000 (2000) 384–395