Hierarchical Pattern Mining Based On Swarm Intelligence

Kazuaki Tsuboi
The University of Electro-Communications
1-5-1 Chofugaoka
Chofu, Tokyo, Japan 182-8585
tsuboi@ni.is.uec.ac.jp

Satoshi Suga
The University of Electro-Communications
1-5-1 Chofugaoka
Chofu, Tokyo, Japan 182-8585
ssuga@ni.is.uec.ac.jp

Satoshi Kurihara
The University of Electro-Communications
1-5-1 Chofugaoka
Chofu, Japan
skurihara@uec.ac.jp

ABSTRACT

The behavior patterns in everyday life such as home, office, and commuting, and buying behavior model by day of the week, season, location have hierarchies of various temporal granularity. Generally, in usual hierarchical data analysis, a basic hierarchical structure is given in advance. But it is difficult to estimate hierarchical structure beforehand for complex data. Therefore, in this study, we propose the algorithm to automatically extract both hierarchical structure and pattern from time series data using swarm intelligent method. We performed the initial operation test and confirmed that patterns can be extracted hierarchically.

CCS CONCEPTS

• **Computing methodologies** → *Intelligent agents*;

GENERAL TERMS

Algorithms

KEYWORDS

Ant Colony Optimization, sequential pattern mining, hierarchical structure

ACM Reference format:
Kazuaki Tsuboi, Satoshi Suga, and Satoshi Kurihara. 2017. Hierarchical Pattern Mining Based On Swarm Intelligence. In *Proceedings of GECCO '17 Companion, Berlin, Germany, July 15-19, 2017,* 2 pages.
DOI: http://dx.doi.org/10.1145/3067695.3082038

1 INTRODUCTION

Nowadays, sequential pattern mining for the time series data is one of the important techniques of big data mining. By extracting a pattern that takes time series into consideration, it became possible to extract patterns that this product should be bought next time after this product was bought. In these sequential pattern mining, the database to be analyzed is divided into units called transactions. , combinations of items that are considered patterns in the scope of this transaction are extracted. That is, the scale to be analyzed is clearly set from the database, and it is divided into transactions according to the scale. When analyzing the different size of the scale, analyze setting on each scale each time.

GECCO '17 Companion, Berlin, Germany

DOI: http://dx.doi.org/10.1145/3067695.3082038

However, when thinking about the real world, it is obvious that the behavior patterns in everyday life such as home, office, and commuting, and buying behavior model by day of the week, season, location have hierarchies of various temporal granularity. Therefore, in this study, we consider different time scales hierarchically and, propose the method to automatically extract both hierarchical structure and pattern from time series data. In this study, we focus on the Ant Colony Optimization(ACO) algorithm[2] in Swarm Intelligence. ACO algorithm is an optimization method which modeled the behavior of ant seeking for food in nature. ACO algorithm is famous as a solution of a traveling salesman problem. ACO algorithm consists of leaving a pheromone to the course which the ant passed and tending to choose the course in which the concentration of a pheromone is high. By this property, whenever ants act, the ants which pass the shortest path increase in number and the concentration of the pheromone which remains in a shortest path becomes greater. On the other hand, the concentration of pheromone will become low if time passes because of evaporation of the pheromone. As a result, the concentration of the pheromone which remained shows the answer to a shortest path problem. So the property of ACO algorithm is extremely robust and adaptable to dynamic change.

2 RELATED STUDIES

Apriori All algorithm[1] discovers the pattern in consideration of a time series. GSP algorithm[4] which aiming at the improvement of the mining speed by imposed time pressure constraints. SPADE algorithm[5] which also aiming at the improvement of the mining speed by dividing a candidate of a sequential pattern by the concept of "lattice". In [3], sequential pattern mining algorithm using rough set theory is proposed. But, in all of these conventional approaches, the goal is pattern extraction, that is, the point of view of time series is out of scope. So, the change of frequent pattern by the progress of time cannot be caught.

3 EXTRACTION METHOD OF PATTERNS WITH HIERARCHICAL STRUCTURE BASED ON ACO

In the proposed algorithm, frequent items that become patterns are extracted from time series data $D = \{d_1, d_2, \cdots, d_n\}$ $(d \in I)$ consisting of an item $I = \{i_1, i_2, \cdots, i_m\}$ taking into consideration the order relation. The length of the time series data$D$ is represented by n. The number of kind of items is represented by m.

The proposed algorithm has a hierarchical structure, and the processing in each layer is basically the same process. In the lower layer, patterns closely connected are extracted, while in the upper layer a pattern is extracted such that more elements are included

ALGORITHM 1: The pseudocode for the proposed algorithm

```
repeat
    Prepare the virtual map M_L;
    repeat
        Set the preference for each ant;
        Search and evaluate for each ant;
        Update the virtual map M_L;
    until for each data set S_L(t);
until for each Layer L ← 1 to MaxLayer;
```

from the lower layer while allowing some roughness based on the pattern extracted in the lower layer. That is, the pattern grows from the lower layer to the upper layer, including the ambiguity pattern. Patterns are constructed in order from the lower layer. Candidates for patterns are set as preferences for ant agents. Ant agents search according to preference, evaluate the result of the search and add the evaluation result as a pheromone to the mapM_L. We show the pseudocode of the proposed algorithm in the Algorithm1.

We prepared the virtual map $M_L = (K_L, v_L(K_L))$ ($K_L \subset I$). Subscription L means it is in the layer L. K_L is represented as the candidates for patterns in the layer L. $v_L(K_L)$ is represented as the amount of pheromone to K_L.

The candidate of the pattern to be searched by each ant agent is determined based on the virtual map M_{L-1} in which the candidate for pattern K_{L-1} extracted in one lower layer is recorded. It is selected according to the probability of the equation 1 from the arbitrary subset in K_{L-1} whose size is X_L or less. For example, in case of $X_L = 3$, it is select from $\{\{k_1\}, \cdots, \{k_{|K_{L-1}|}\}, \{k_1, k_1\}, \cdots, \{k_{|K_{L-1}|}, k_{|K_{L-1}|}\}, \{k_1, k_1, k_1\}, \cdots, \{k_{|K_{L-1}|}, k_{|K_{L-1}|}, k_{|K_{L-1}|}\}\}$.

$$p(k'_L) = \frac{\prod_{k \in k_{L-1}} v(k)}{\sum_{x=1}^{X_L} (\sum_K v_{L-1}(k_{L-1}))^x} \tag{1}$$

Furthermore, for the selected k'_L, the search is performed for the search item set k_L in which the dummy element j_L is inserted between elements in order to allow noise. That is, for example, if $k'_L = \{k_1, k_2, k_3\}$ is selected, the item set to be searched is $k_L = \{k_1, j_1, k_2, j_2, k_3\}$.

For each search data set$S_L(t)$, m_L number of ant agents perform a search and update the virtual map according to the equation2.

$$v(k) = v'(k)(1 - \rho_L) + C_k(t) \tag{2}$$

Evaporating the pheromone depend on the evaporation rate ρ_L. By the feature of evaporation pheromone, the new behavior is reflected in an analysis result in a constant ratio always. So, the evaporation rate ρ shows the speed of update.

4 EXPERIMENT

Currently, we are conducting experiments with the proposed algorithm. With 100 kinds of item sets $I = \{0, 1, 2, 3, \cdots, 99\}$, the patterns to be extracted are $\{0, 1, 2, 3\}$ and $\{4, 5, 6, 7\}$. We generate random numbers corresponding to the test data length with a uniform random number from item set I. In this test, we set the test data length to 1000. Then, for the time series data, $\{0, 1, 2, 3\}$ which is a pattern to be extracted is embedded in order of 20, and $\{4, 5, 6, 7\}$ is embedded of 10. To influence the noise on the pattern,

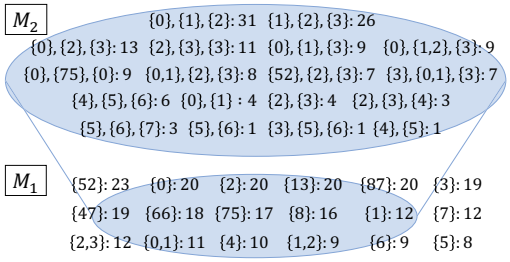

Figure 1: Extracted patterns of two layers

the width between the elements in the pattern follows the power-law distribution. In this test, the hierarchical structure adopts a two-layer structure, lower layer and upper layer. To make $S_L(t)$, time series data$D$ is divided by every 5 items in the lower layer, and by every 10 items in the upper layer. For each $S_L(t)$, 1000000 ant agents perform a search. The size of the candidate for the pattern as preferences is set to $X_L = 3$. Dummy element in the lower layer is $j_1 = \{\emptyset\}$ and in the upper layer is $j_2 = \{i | i \in I \wedge |i| < 1\}$. This is because we search for the lower layer in which elements to be searched continuously appear. On the other hand, in the upper layer it can be searched even when other items that become only one noise are included among the search item sets.

In the virtual map M_1 and M_2, top ten pattern candidates in which a large amount of pheromone remains and candidates related to assumed patterns to be extracted are shown in the Figure 1. In the Figure 1, for example, "{0}, {1}, {2} : 31 " shows that "{0}, {1}, {2} " is the extracted items and "31" shows the amount of pheromone. The higher the frequency of appearance is, the better it is extracted. And, patterns with lower frequency can also be extracted in a subset item, so it can be extracted well by increasing the hierarchical structure.

5 CONCLUSION

In this study, we have proposed a new pattern mining algorithm with the hierarchical structure based on swarm intelligence. In order to evaluate the performance of the proposed algorithm, we tested using test data embedded with patterns to be extracted. As a result, it is confirmed that the higher the frequency of appearance is, the better it is extracted. In the next step, we are going to increase hierarchical structure and we attempt to extract behavior patterns in the virtual living environment.

REFERENCES

[1] Rakesh Agrawal and Ramakrishnan Srikant. 1995. Mining sequential patterns. In *Data Engineering, 1995. Proceedings of the Eleventh International Conference on*. IEEE, 3–14.
[2] Marco Dorigo and Luca Maria Gambardella. 1997. Ant colony system: a cooperative learning approach to the traveling salesman problem. *IEEE Transactions on evolutionary computation* 1, 1 (1997), 53–66.
[3] Ken Kaneiwa and Yasuo Kudo. 2011. A sequential pattern mining algorithm using rough set theory. *International Journal of Approximate Reasoning* 52, 6 (2011), 881–893.
[4] Ramakrishnan Srikant and Rakesh Agrawal. 1996. Mining sequential patterns: Generalizations and performance improvements. *Advances in Database Technology—EDBT'96* (1996), 1–17.
[5] Mohammed J Zaki. 2001. SPADE: An efficient algorithm for mining frequent sequences. *Machine learning* 42, 1 (2001), 31–60.