확률적 다차원 연속패턴의 생성을 위한 효율적인 마이닝 알고리즘

(An Efficient Mining Algorithm for Generating Probabilistic Multidimensional Sequential Patterns)

이 창 환 †

(Chang-Hwan Lee)

요 약 연속패턴은 다양한 분야에서 사용되는 데이타 마이닝 기법의 한 종류이다. 하지만 현재의 연속패턴 방법은 한개의 속성내에서의 패턴만을 감지할 수 있으며 속성간의 패턴을 생성할 수 없다. 다차원의 연속패턴은 일차원에 비하여 훨씬 유용한 정보를 제공할 수 있다. 본 연구에서는 Hellinger 엔트로피 함수를 사용하여 다차원의 연속패턴을 생성하는 방법을 제시한다. 기존의 연속패턴방법과 달리 본 방법에서는 각 연속패턴의 중요도를 자동으로 계산할 수 있다. 또한 계산의 복잡도를 감소시키기 위한 다수의 법칙이 개발되었으며 다수의 실험 결과를 제시하였다.

키워드 : 데이타마이닝, 연속패턴, Hellinger 변량, 인공지능

Abstract Sequential pattern mining is an important data mining problem with broad applications. While the current methods are generating sequential patterns within a single attribute, the proposed method is able to detect them among different attributes. By incorporating these additional attributes, the sequential patterns found are richer and more informative to the user. This paper proposes a new method for generating multi-dimensional sequential patterns with the use of Hellinger entropy measure. Unlike the previously used methods, the proposed method can calculate the significance of each sequential pattern. Two theorems are proposed to reduce the computational complexity of the proposed system. The proposed method is tested on some synthesized purchase transaction databases.

Key words : Data Mining, Sequential Pattern, Hellinger measure, Artificial Intelligence

1. Introduction

Data mining has been widely used to detect important knowledge from a vast amount of data. It can provide useful, interesting, and high quality information to users. Among many techniques in data mining, sequential pattern mining is a technique which can discover more meaningful information by considering time attribute, together with other traditional attributes. For example, from a department store's transaction database, we can find the following sequential purchasing patterns: "*People who purchase a desktop PC later purchase a printer*" and "*Customers who buy a digital camera typically buy camera batteries later.*"

Sequential patterns can be widely used in many different applications, such as predicting certain kind of disease from history of symptoms, and predicting what product a customer will purchase based on his/her transaction history. Therefore, sequential pattern mining has been intensively studied during recent years, and there exist a number of algorithms for sequential pattern mining.

Almost all of the current methods for mining sequential patterns are based on the Apriori algorithm [1]. After that, a series of Apriori-like algorithms have been proposed: GSP [1], PSP [2], SPIRIT [3], FreeSpan [4], PrefixSpan [5], and SPADE [6].

However, one of the limitations of the current

· 본 연구는 2003년도 동국대학교 교내연구비 지원과제 임

† 종신회원 : 동국대학교 정보통신공학과 교수

chlee@dgu.ac.kr

논문접수 : 2004년 3월 8일

심사완료 : 2004년 11월 17일

sequential pattern algorithms is that they mine only one dimension. They only consider one attribute, and thus can not detect sequential patterns hiding across different attributes.

1.1 Multi-dimension vs. Single-dimension

In real applications, sequence patterns are usually associated with different circumstances, and such circumstances form a multiple dimensional space. Time-related databases might contain many other customer-related, product-related, and/or transaction-related attributes. For example when mining for purchase sequential patterns from a transaction database, the single dimensional mining methods only look at the items purchased.

On the other hand, multi-dimensional sequential pattern mining attempts to find sequential patterns across several dimensions of attribute. By incorporating the additional attributes, the sequential patterns found are richer and more informative to the user. For example, customer purchase sequences are associated with price, time, occupation, and others. Clearly, it is more interesting and useful to mine sequential patterns associated with multi-dimensional information. In multi-dimensional sequential pattern mining, many different attributes related to the transaction data were introduced and formed a multi-dimensional sequential datasets. The aim of this multi-dimensional sequential pattern mining is to get more interesting sequential patterns with different dimensional attributes.

In this paper, we propose the theme of multi-dimensional sequential pattern mining and thoroughly explore efficient methods for multi-dimensional sequential pattern mining with the use of Hellinger entropy measure. In addition, our method could calculate the significance of each sequential pattern as a numeric value, and these sequential patterns are given in a sorted order based on this numeric measure.

1.2 Related Work

The sequential pattern mining problem was first introduced by Agrawal and Srikant in [1]. Given a set of sequences, where each sequence consists of a list of elements and each element consists of a set of items, and given a user-specified minimum support threshold, sequential pattern mining is to find all of the frequent subsequences, i.e., the subsequences whose occurrence frequency in the set of sequences is no less than minimum support. Most of the basic and earlier algorithms for sequential pattern mining are based on the Apriori algorithm proposed in association rule mining.

Since then, many studies have contributed to the efficient mining of sequential patterns or other frequent patterns in time-related data. GSP [1] starts by finding the frequent 1-item patterns, and then these become the basis for generating all the potential candidate sets, and counts their support. The 2-item patterns with large support value are retained and become the seed set for the next pass. The algorithm continues to make multiple passes over the data until no more candidates can be generated or sequential patterns found.

PSP [2] was developed to improve the way in which GSP stored candidate patterns. PSP creates a prefix-tree, where any branch from its root to the leaf, stands for a candidate sequence, and the terminal node provides the support of the sequence.

SPIRIT [3] proposes the use of Regular Expressions (RE) as a flexible constraint specification tool that enables user-controlled focus to be incorporated into the sequential pattern mining process. While conventional mining systems provide users with only a very restricted mechanism for specifying patterns of interest, they develop a family of algorithms for mining frequent sequential patterns that also satisfy user-specified RE constraints.

However, the major limitation of the current sequential pattern algorithms is that they mine only one dimension. Usually, sequence patterns are associated with different attribute circumstances, and such circumstances form a multiple dimensional space.

In the data mining literature, not many study is known about multi-dimensional sequential pattern mining. One of them is UNISEQ method [7], and, in UNISEQ, the multi-dimensional attributes are embedded in the sequential database by adding a new element in the sequence database. With the newly formed sequential database, PrefixSpan [5] is employed in this method. However, this type of

multi-dimensional association pattern mining methods are fairly straightforward extensions from single-dimensional mining, and little research has been directed towards developing alternative methods.

In this paper, we propose a new paradigm for generating multi-dimensional sequential patterns. An efficient information theoretic approach for multi-dimensional sequential pattern mining, with the use of Hellinger entropy measure, is proposed. Our proposed method could calculate the significance of each sequential pattern as a numeric value, and these sequential patterns are given in a sorted order.

2. Information Content of Sequential Patterns

Sequential pattern generation can be viewed as a search for hypotheses to account for a set of time-related instances or examples which are often assumed to be restricted to some instance space. For the purpose of this paper, the hypothesis space of sequential pattern will be restricted to the conjunctive propositions in the discrete space defined by the Cartesian product of the sample spaces of the individual attributes-the extension to more general hypothesis spaces remains a topic for further investigation.

The format of sequential patterns which we will handle in this paper is as follows:

$$A=a \wedge B=b \wedge \ldots \rightarrow T=t$$

where A, B and T are attributes with a, b and t being values in their respective discrete alphabets. We restrict the right-hand expression to being a single value assignment expression while the left-hand side may be a conjunction of such expressions. The semantics of above format is that if a person does an action (e.g., purchase) based on the condition (left-hand side) of above pattern at a given time, then he will later do an action described in right-hand side with high possibility.

Since our sequential pattern method handles multi-dimensional databases, the format of database is different from the format used by traditional sequential pattern methods. Each transaction of database is associated with different circumstances for multi-dimensional sequential patterns mining, including the circumstances (attributes) of customer, item, and transaction, and such circumstances form a multiple dimensional space. The entire database is sorted based on customer-id and transaction-time.

The basic idea of sequential pattern generation starts with the assumption that the value assignments in the left hand side of each sequential pattern affects the probability distribution of the right-hand side (target attribute). The target attribute forms its a priori probabilities without presence of any left-hand conditions. It normally represents the class frequencies of the target attribute. However, its probability distribution changes when it is measured under certain conditions usually given as value assignments of other attributes. Intuitively speaking, if a certain value assignment has significantly changed the probability distribution of the target, it is clear that the given value assignment plays an important role in determining the class values of the target attribute. On the other hand, if the probability distribution of the target attribute remains the same regardless of a value assignment of other attribute, those two attributes are regarded as independent of each other. Therefore, in this paper, the significance of a sequential pattern is interpreted as the degree of dissimilarity between a priori probability distribution and a posteriori probability distribution of the target attribute.

In this paper, this dissimilarity is defined as instantaneous information, which is the information content of the sequential pattern given that the left-hand side happens. The critical part now is how to define or select a proper measure which can correctly measure the instantaneous information.

C4.5, which generates decision trees from data, has been widely used for classification in Quinlan [8]. C4.5 uses the following formula as a measure of information for attribute A.

$$H(T) - H(T|A=a) = \sum_t p(t) \log\left(\frac{1}{P(t)}\right) - \sum_t P(t|a) \log\left(\frac{1}{P(t|a)}\right) \quad (1)$$

where T and t represent the target attribute and its corresponding value, respectively. It calculates the difference between the entropy of a priori distribution and that of a posteriori distribution.

However, it is well-known that there is a fundamental problem with this measures. Consider the case of an n-valued variable where a particular value of $T=t$ is one, while all the other values in T's alphabet are zero. In this case, a conditional permutation of these probabilities would be significant, i.e., a rule which predicts the relatively rare event $T=t$. However, the formula (1), because it cannot distinguish between particular events, would yield zero information for such events.

In this paper a new information measure, called Hellinger measure, is used to define the information content of sequential pattern rules. The Hellinger divergence was originally introduced by Beran [9], and is defined as

$$\sqrt{\sum_i (\sqrt{p(t_i)} - \sqrt{p(t_i|a)})^2} \qquad (2)$$

where t_i denotes the value of attribute T. It becomes zero if and only if both a priori and a posteriori distributions are identical, and ranges from 0 to 1. Unlike other information measures, this measure is applicable to every possible case of probability distributions. In other words, the Hellinger measure is continuous on every possible combination of a priori and a posteriori values. It can be interpreted as a distance measure where distance corresponds to the amount of divergence between a priori and a posteriori distribution. Therefore, we employ Hellinger measure as a measure of divergence, which will be used as the information amount of sequential patterns.

3. Properties of *H* Measure

In terms of the sequential pattern rules, let us interpret the event $A=a$ as the target concept to be learned and the event(possibly conjunctive) $B=b$ as the hypothesis describing this concept. The information content of the sequential pattern rule is defined as

$$[\sqrt{P(a|b)} - \sqrt{P(a)}]^2 + [\sqrt{1-P(a|b)} - \sqrt{1-P(a)}]^2 \qquad (3)$$

where $P(a|b)$ means the conditional probability of $A=a$ under the condition $B=b$ has happened beforehand. Notice that Equation (3) has a different form of definition from that of Equation (2). In sequential pattern generation, one particular value of class attribute appears in the right hand side of the pattern, and thus the probabilities for all other values are included in *1-P(a)*. In addition, we squared the original form of Hellinger measure because (1) by squaring the original form of Hellinger measure, we could derive a boundary of the Hellinger measure, which allows us to reduce drastically the search space of possible sequential pattern rules, (2) the relative accuracy of each pattern is not affected by the modified Hellinger measure, and (3) the weights between two terms of Hellinger measure provides more reasonable trade-off in terms of their value range. This measure can be interpreted as the cross entropy of A with the variable "A conditioned on the event B=b." Cross entropy is well-known as an accuracy measure between two distributions and we address this measure as *accuracy* of patterns.

In addition, we need a preprocessing step when calculating the posterior probabilities($P(a|b)$) in Equation (3). Suppose a certain customer has $t_1, t_2, \ldots t_n$ transactions sorted on transaction time, and, among them, p transactions satisfy the left-hand side conditions of the given sequential pattern rule. For a certain transaction t_i, we define δ_i as *the number of transactions following* t_i *within the given customer*. Then, the posterior probability ($P(a|b)$) is defined as

$$P(a|b) = \frac{p + \sum_{i=1}^{p-1} \delta_i}{n + \sum_{i=1}^{p-1} \delta_i}. \qquad (4)$$

As an illustrative example, assume we have a sample database in Figure 1, and consider a sequential pattern we are going to consider is given as *Customer=C1* $\wedge$ *Item=P1* $\rightarrow$ *Item=P2* . In this database, Customer C1 has 7 transactions and three of them (transaction (1), (3), and (7)) satisfy the condition part of the above sequential pattern. The posterior probability($P(a|b)$) is calculated as

$$P(a|b) = \frac{3+(6+4+0)}{7+(6+4+0)} = \frac{13}{17}. \qquad (5)$$

Another criteria we have to consider is the *generality* of the sequential patterns. The basic idea behind generality is that the more often left-hand side occurs for a sequential pattern, the more useful the pattern becomes. The left-hand side

must occur relatively often for a pattern to be deemed useful. In this paper, we use $\sqrt{P(b)}$ to represent the probability that the sequential pattern will occur and, as such, can be interpreted as the measure of sequential pattern generality.

Table 1 Preprocessing Example

ID	Customer	Time	Item
(1)	C1	D1	P1
(2)	C1	D1	P4
(3)	C1	D2	P1
(4)	C1	D2	P2
(5)	C1	D2	P3
(6)	C1	D3	P2
(7)	C1	D3	P1
(8)	C2	D2	P2
...	...	...	...

Preprocessing step is also needed to calculate the generality of sequential patterns. Suppose the total number of transaction (of entire customers) is N, and there are p transactions satisfying the condition part of the given sequential pattern. We define, for a certain transaction t_i, Δ_i as *the number of transactions following t_i within the given customer*. Then, the generality($P(b)$) of the given sequential pattern is defined as

$$P(b) = \frac{p + \sum_{i=1}^{b-1} \Delta_i}{N + \sum_{i=1}^{b-1} \Delta_i}. \quad (6)$$

By multiplying the generality with the accuracy of the sequential pattern rules, we have the following term

$$\sqrt{P(b)}\,[(\sqrt{P(a|b)} - \sqrt{P(a)})^2 + (\sqrt{1-P(a|b)} - \sqrt{1-P(a)})^2] \quad (7)$$

which possesses a direct interpretation as a multiplicative measure of the generality and accuracy of a given sequential pattern rule. In this paper, we call the above multiplicative term H measure of sequential patterns.

3.1 Boundaries of H Measure

The algorithm starts with generating an initial set of sequential patterns, followed by specialization of these sequential patterns to optimize the pattern set. The characteristic of the specialization behavior is critical to the performance of the algorithm. Therefore, in this section, we are going to derive some quantitative bounds on the nature of specialization, which can be used to improve computational performance. Specialization is the process by which we try to increase a sequential pattern's accuracy by adding an extra condition to the pattern's left-hand side. The consequent necessary decrease in generality of the sequential pattern should be less than an increase in the accuracy to the extent that the overall H measure is increased.

We will examine specialization, using the H measure as the definition of sequential pattern goodness, with $\sqrt{p(a)}$ corresponding to generality and Equation (3) corresponding to accuracy. The question we pose is as follows: given a particular general sequential pattern, what quantitative statements can we make about specializing this sequential pattern ? In particular, if we define H_s and H_g as the H measures of the specialized and general sequential patterns, respectively, is it possible to find a bound of H_s in terms of H_g ?

Suppose we have a sequential pattern

$$B=b \rightarrow A=a. \quad (8)$$

We would like to specialize this sequential pattern by adding a condition $C=c$ so that we have a specialized sequential pattern

$$B=b \wedge C=c \rightarrow A=a. \quad (9)$$

For the sake of illustration, sequential patterns in formulas (8) and (9) are denoted as R_g and R_s, respectively. In this section, we deal with a sequential pattern which contains only one condition and try to specialize it. More general cases which have more than one condition in the left hand side can be easily understood. Suppose H_g and H_s are the H measures of the sequential patterns R_g and R_s, respectively. Our goal is to answer the question "Can we describe the bound of H_s in terms of H_g ?" In other words, is it possible to estimate the maximum value of H_s without knowing any information about attribute C? The motivation for bounding H_s in this manner is two-folds. Firstly, it produces some theoretical insight into specialization, while secondly, the bound

can be used by the sequential pattern algorithm to search the search space (hypothesis space) efficiently.

Consider that we are given a general sequential pattern whose H measure, H_g, is defined as

$$H_g = \sqrt{P(b)}([\sqrt{P(a|b)} - \sqrt{P(a)}]^2 + [\sqrt{1-P(a|b)} - \sqrt{1-P(a)}]^2)$$
$$= \sqrt{P(b)}[2 - 2\sqrt{P(a|b)P(a)} - 2\sqrt{(1-P(a|b))(1-P(a))}] \quad (10)$$

We try to calculate the bound of

$$H_s = \sqrt{P(bc)}[2 - 2\sqrt{P(a|bc)P(a)} - 2\sqrt{(1-P(a|bc))(1-P(a))}]$$
$$= \sqrt{P(c|b)}\sqrt{P(b)}[2 - 2\sqrt{P(a|bc)P(a)} - 2\sqrt{(1-P(a|bc))(1-P(a))}] \quad (11)$$

Given no information about C, we can state the following results.

Theorem 1: *If the H measure of a specialized pattern satisfies the following boundary:*

$$H_s \le \max\{\sqrt{P(a|b)}\sqrt{P(b)}[2\sqrt{m} - 2\sqrt{P(a)}],$$
$$2\sqrt{P(b)} - \sqrt{1-P(a|b)}\sqrt{P(b)}[2\sqrt{P(a)} + 2\sqrt{1-P(a)}]\}$$

where m represents the number of class in the target attribute, the general pattern discontinues specializing.

Proof: For given H_g and H_s, defined in Equation (10) and (11), respectively. We know that

$$P(ab) = P(abc) + P(ab\neg c) = P(a|bc)P(bc) + P(ab\neg c)P(b\neg c) \quad (12)$$

Dividing Equation (12) by $P(b)$, we have

$$P(a|b) = P(a|bc)P(c|b) + P(ab\neg c)P(\neg c|b)$$
$$= P(a|bc)P(c|b) + P(ab\neg c)(1 - P(c|b))$$

Therefore,

$$P(c|b) = \frac{P(a|b) - P(ab\neg c)}{P(a|bc) - P(ab\neg c)}. \quad (13)$$

Let ω denote $P(ab\neg c)$.

i) if $\omega \le P(a|b)$ and $\omega \le P(a|bc)$, both numerator and denominator of $P(c|b)$ are positive, and thus

$$\max_w P(c|b) = \frac{P(a|b)}{P(a|bc)} \text{ when } \omega = 0.$$

$$H_s \le \sqrt{\frac{P(a|b)}{P(a|bc)}}\sqrt{P(b)}[2 - 2\sqrt{P(a|bc)P(a)} - 2\sqrt{(1-P(a|bc))(1-P(a))}]$$
$$= \sqrt{P(a|b)}\sqrt{P(b)}[\frac{2}{\sqrt{P(a|bc)}} - 2\sqrt{P(a)} - 2\sqrt{\frac{1-P(a|bc)}{P(a|bc)}(1-P(a))}]$$
$$\le \sqrt{P(a|b)}\sqrt{P(b)}[\frac{2}{\sqrt{P(a|bc)}} - 2\sqrt{P(a)}].$$

Without loss of generality, we can assume that $1/m \le P(a|bc) \le 1$ since the highest frequency among the values of the target attribute is to be included in the right hand side of the pattern. Therefore,

$$H_s \le \sqrt{P(a|b)}\sqrt{P(b)}[2\sqrt{m} - 2\sqrt{P(a)}].$$

ii) if $\omega > P(a|b)$ and $\omega > P(a|bc)$, both numerator and denominator of $P(c|b)$ are negative, and thus

$$\max_w P(c|b) = \frac{1-P(a|b)}{1-P(a|bc)} \text{ when } \omega = 1.$$

$$H_s \le \sqrt{\frac{1-P(a|b)}{1-P(a|bc)}}\sqrt{P(b)}[2 - 2\sqrt{P(a|bc)P(a)} - 2\sqrt{(1-P(a|bc))(1-P(a))}]$$
$$= 2\sqrt{\frac{1-P(a|b)}{1-P(a|bc)}}\sqrt{P(b)} - \sqrt{1-P(a|b)}\sqrt{P(b)}[2\sqrt{\frac{P(a|bc)}{1-P(a|bc)}P(a)} + 2\sqrt{1-P(a)}]$$
$$\le 2\sqrt{P(b)} - \sqrt{1-P(a|b)}\sqrt{P(b)}[2\sqrt{P(a)} + 2\sqrt{1-P(a)}]$$

iii) otherwise, the probability of $P(c|b)$ in Equation (13), becomes less than zero, which is impossible to occur.

Therefore, from i), ii), and iii), we can prove Theorem 1. Q.E.D.

As a special case of Theorem 1, if the success rate (conditional probability) of general pattern becomes 1, the H measure of the specialized pattern is always less than or equal to that of general pattern.

Theorem 2 : *If the conditional probability ($P(a|b)$) of general pattern is 1, H measure of specialized pattern cannot be greater than that of general pattern. Therefore, the general pattern discontinues specializing.*

Proof: From $P(b) = P(ab) + P(\neg ab)$ and $P(a|b) = \frac{P(ab)}{P(b)} = 1$,

$$P(\neg ab) = P(b) - P(ab) = 0.$$

Therefore,

$$P(a|bc) = \frac{P(abc)}{P(bc)} = \frac{P(abc)}{P(abc) + P(\neg abc)} = \frac{P(abc)}{P(abc) + P(c|\neg ab)P(\neg ab)} = 1 \quad (14)$$

From Equation (10) and $P(a|b) = 1$, $H_g = \sqrt{P(b)}(2 - 2\sqrt{P(a)})$.

From Equation (11) and (14), $H_s = \sqrt{P(bc)}(2 - 2\sqrt{P(a)})$.

Since $P(bc) \le P(b)$, $H_s \le H_g$. Q.E.D.

As a consequence of these theorems, we note that since the bound of specialized sequential pattern is achievable without further information about C, we can decide in advance that the specialized sequential pattern cannot be improved

with respect to H measure. The logical consequence of this statement is that it precludes using the bound to discontinue specializing based on the value of H_g alone. Conversely, if $p(a|b)$ is not equal to 1, then with no information at all available about the other variables, there may always exist a more specialized sequential pattern whose information content is strictly greater than that of the general sequential pattern. However, as we shall see, we could certainly compare the bound with any sequential patterns we might already have. In particular, if the bound is less than the information content of the worst sequential pattern (H_*, described in the following section), then specialization cannot possibly find any better sequential pattern. This principle will be the basis for restricting the search space of the system.

4. Sequential Pattern Generation

We will now define the algorithm and discuss its basic ideas. The algorithm takes time-related database in the form of discrete attribute vectors and generates a set of K sequential patterns, where K is a user-defined parameter. The set of generated sequential patterns are the K most informative sequential patterns from the database as defined by the H measure. The algorithm proceeds by first finding K sequential patterns, calculating their H measures, and then placing these K sequential patterns in an ordered list. The smallest H measure, that of the Kth element of the list, is then defined as the running minimum H_*.

The critical part of the algorithm is the specialization criterion since it determines how much of the hypothesis space actually needs to be explored by the algorithm. The algorithm employs branch-and-bound with depth-first search over possible left-hand sides, starting with the first-order conditions(single value assignment in left-hand side). From that point onwards, new patterns which are candidates for inclusion in the sequential pattern set have their H measure compared with H_*. If they are greater than H_*, they are inserted in the list and the Kth sequential pattern is deleted. And H_* is updated with the value of the H measure of whatever sequential pattern is now Kth on the list. The algorithm systematically tries to specialize all first-order sequential patterns and terminates when it has determined that no more sequential patterns exist which can be specialized to achieve a higher H measure than H_*. The decision whether to continue specializing or to back-up on the depth-first search is determined by the algorithm in Figure 2. The H measure of each sequential pattern can be considered as the significance of the patterns.

5. Experimental Results

In order to test the functionality of the algorithm proposed in this paper, we assumed an artificial time-related database described in Table 2, and synthesized two sets of artificial datasets. The database in Table 2 contains 14 attributes. The databases for traditional sequential pattern algorithms (Apriori-like algorithms) contain the following three (or more) attributes: customer-id, transaction-time, (multiple) items. However, since our method is generating multi-dimensional sequential patterns, the database may contain many attributes in various categories. The attributes in the database can be grouped into the following four categories: 1) basic, 2) customer-related, 3) item-

If success rate of $H_g = 1$
Then
  cease to specialize; /* by Theorem 2 */
Else
  Let

$$2\sqrt{P(b)} - \sqrt{1-P(a|b)}\sqrt{P(b)}\,[\,2\sqrt{P(a)} + 2\sqrt{1-P(a)}\,]$$

  If $H_s \le H_*$, **Then** cease to specialize; /* by Theorem 1 */
End-if

Figure 1 Algorithm for specialization

related, 4) transaction-related.

The proposed algorithm was tested on two synthetic datasets. Each database contains 20,000 records, and data values are generated using random numbers. For each data set, the entire data set is read and then 100 most informative sequential patterns were generated.

The topmost 15 sequential patterns from the first dataset is shown in Table 3. For each pattern in Table 3, its corresponding values for confidence, generality, and H measure are shown, and the resulting patterns are sorted based on their H measure values. The confidence means the number of transactions satisfying both left-hand side and right-hand side of the pattern divided by the number of transactions satisfying left-hand side only.

The topmost pattern in Table 3 means that customers who purchased items (whatever the items are) of which price are between 20-29 later purchase item P07. This type of patterns can not be acquired from traditional sequential pattern methods. The pattern 4 shows a sequential pattern equivalent to the one generated from Apriori-like method. It illustrates that the functionality of our method includes that of traditional sequential pattern methods. In pattern 2, it shows that if male customers purchase item P06 then they later

Table 2 Sample Database

Attribute Category	Attribute name	Corresponding Values
Basic attributes	(1) Customer-ID	C00, C02, ... C20
	(2) Time	continuous
	(3) Item	P00, P02, ... P10
Customer related attributes	(4) Gender	male, female
	(5) Age	less than 20, 20 29, 30-39, 40-49, over 50
	(6) Married	single, married, others
	(7) Region	city, non city
	(8) Occupation	sales, engineer, teacher, others
Item related attributes	(9) Unit price	less than $20, $20-$29, $30-$39, $40-$49, over $50
	(10) SaleOrNot	sale, nonsale
	(11) Color	blue, red, white, others
Transaction related attributes	(12) Qty	1, 2-3, over 4
	(13) DayofWeek	Su., Mo., Tu., We., Th., Fr., Sa.
	(14) Payment	cash, credit card, others

Table 3 Sequential patterns using dataset I

No.	Sequential Patterns	Confidence	Generality	H
1	Price=20-29 → Item=P07	0.13137	0.00380	0.00023
2	Gender=male & Item=P06 → Item=P02	0.11015	0.00536	0.00021
3	Qty=1 → Item=P09	0.11800	0.00203	0.00021
4	Item=P09 → Item=P01	0.11067	0.00353	0.00019
5	SaleorNot=sale → Item=P00	0.11207	0.00477	0.00017
6	Age=20-29 & Qty=over 5 → Item=P03	0.10592	0.00621	0.00015
7	Price=30-39 & Qty=1 → Item=P02	0.10559	0.00649	0.00015
8	Gender=male → Item=P07	0.12526	0.00585	0.00014
9	Week=Su. → Item=P07	0.10678	0.00709	0.00012
10	Item=6 → Item=P05	0.12689	0.00364	0.00012
11	Occupation=sales → Item=P02	0.10658	0.00654	0.00012
12	Region=city & Item=P09 → Item=P01	0.12378	0.00607	0.00011
13	Price=20 29 → Item=P08	0.12515	0.00455	0.00011
14	Week=We. → Item=P05	0.10757	0.00770	0.00011
15	Gender=male & Payment=cash → Item=P03	0.10699	0.00347	0.00011
...	...	...	...	...

purchase item P02. This pattern shows how our method could combine the item attributes with other customer-related attributes, thus forming multi-dimensional pattern. Using the first dataset, we could demonstrate how the proposed sequential algorithm could detect and represent multi-dimensional sequential patterns.

The second dataset also contains 20,000 records, and data values are generated using random numbers. However, in the second dataset, we assumed that there are a number of sequential patterns hidden in the real world, and the dataset is generated based on those sequential patterns. The sequential patterns we have assumed are as follows.

- Color=white & Qty=1 → Item=P05
- Region=city & Item=10 19 → Item=P08

The goal of this experiment is to verify whether the proposed algorithm is able to detect these sequential patterns hidden in the dataset. For the second experiment, the entire data set is read and then 100 most informative sequential patterns were generated. The topmost 15 sequential patterns from the second dataset is shown in Table 4.

The sequential patterns we have assumed are generated from the system and shown in Table 4 as pattern 2 and pattern 5, respectively. We could also see many other multi-dimensional sequential patterns in Table 4. This experiment illustrates that our proposed algorithm is able to effectively detect the sequential patterns hidden within the dataset.

6. Conclusion

In this paper we have introduced a new method for generating multi-dimensional sequential patterns from time series databases. We developed an information theoretic measure, called H measure, which becomes the criteria for selecting and sorting inductive sequential patterns generated. The boundary of the H measure is analyzed and two theorems are developed to reduce the computational complexity of the system. In addition, missing values can be handled by considering them as separate categories. The algorithm is applied to two synthetic databases. The resulting sequential patterns generated from the data sets show how the system detects the hidden multi-dimensional patterns of data sets effectively.

References

[1] R. Agrawal and R. Srikant, Mining Sequential Patterns, *Int. Conf. on Data Engineering* pp. 3-14, 1995.

[2] F. Masseglia, F. Cathala and P. Poncelet. The PSP Approach for Mining Sequential Patterns. *The 2nd European Symposium on Principles of Data Mining and Knowledge Discovery (PKDD '98)*, Vol. 1510, pages 176-184, Nantes, France, LNAI, September 1998.

Table 4 Sequential patterns using dataset II

No.	Sequential Patterns	Confidence	Generality	H
1	Item=1 → Item=P07	0.17834	0.00214	0.000181
2	Color=white & Qty=1 → Item=P05	0.15481	0.00282	0.000103
3	Price=10 19 → Item=P03	0.13250	0.00200	0.000065
4	Price=30 39 → Item=P09	0.14624	0.00216	0.000051
5	Region=city & Item=10 19 → Item=P08	0.11951	0.00351	0.000040
6	Color=white & Qty=1 → Item=P08	0.11440	0.00459	0.000040
7	Payment=cash → Item=P03	0.12714	0.00882	0.000038
8	Occupation=sales & Color=red → Item=P02	0.11445	0.00443	0.000035
9	SaleOrNot=sale → Item=P05	0.11211	0.00478	0.000032
10	Gender=female & Color=red → Item=P01	0.11086	0.00693	0.000028
11	Price=10 19 & Qty=1 → Item=P04	0.10889	0.00488	0.000027
12	Sale=nonsale → Item=P03	0.11645	0.00424	0.000026
13	Age=10 19 & Week=Sa. → Item=P05	0.11092	0.00341	0.000025
14	Qty=1 → Item=P02	0.11051	0.00586	0.000023
15	Gender=male → Item=P02	0.11117	0.00522	0.000022
...	...	...	...	...

[3] M. Garofalaskis, R. Rastogi, and K. Shim, Spirit:Sequential Pattern Mining with Regular Expression Constraints, 1999 *Inernational Conference on Very Large Databases*, pp. 223-234, 1999.
[4] J. Han, J. Pei, B. Mortazavi-Asl, Q. Chen, U. Dayal and M-C. Hsu. FreeSpan: Frequent Pattern-Projected Sequential Pattern Mining. In *Proc. 2000 Int. Conf. Knowledge Discovery and Data Mining (KDD'00)*, 355-359, Boston, MA, Aug. 2000.
[5] J. Pei, J. Han, B. Mortazavi-Asl, H. Pinto, Q. Chen, U. Dayal, and M.-C. Hsu, PrefixSpan: Mining Sequential Patterns Efficiently by Prefix-projected Pattern Growth, *2001 Int. Conf. on Data Engineering*, pp. 215-224, 2001.
[6] M. J. Zaki. SPADE: An Efficient Algorithm for Mining Frequent Sequences. *In Proc. of Machine Learning Journal, special issue on Unsupervised Learning(Doug Fisher, ed.)*, Vol. 42 No. 1/2, pages 31-60, Jan/Feb 2001.
[7] H. Pinto, J. Han, J. Pei, K. Wang, Q. Chen, and U. Dayal, Multi-dimensional Sequential Pattern Mining, *Int. Conf. on Information and Knowledge Management*, Atlanta, GA, 2001.
[8] J. R. Quinlan, *C4.5: Programs for Machine Learning*, Morgan Kaufmann Publisher, 1993.
[9] R. J. Beran. Minimum Hellinger Distances for Parametric Models, *Ann. Statistics*, Vol. 5, pp. 445-463, 1977.
[10] R. Srikant and R. Agrawal, Mining Sequential Patterns: Generalizations and Performance Improvements, *the 5th Inernational Conference on Extending Database Technology*, pp. 3-17, 1996.
[11] J. Han and M. Kamber, *Data Mining: Concepts and Techniques*, San Francisco: Morgan Kaufmann Publishers, 2001.
[12] S.-J. Yen and A. Chen, An Efficient Approach to Discovering Knowledge from large Databases, *The 4th Int'l Conf. on Parallel and Distributed Information Systems*, 1996.

이 창 환

1982년 2월 서울대학교 계산통계학과 졸업(학사). 1988년 8월 서울대학교 계산통계학과 졸업(석사). 1994년 8월 University of Connecticut, Dept. of Computer Science(박사). 1982년 3월~1987년 2월 한국기계연구소. 1994년 12월~1996년 2월 AT&T Bell Laboratories, Middletown, USA. 1996년 3월~현재 동국대학교 정보통신학과 부교수. 관심분야는 기계학습, 마이닝, 생물정보학 등