Structural Relation Sequential Patterns Mining

Weiru CHEN
Faculty of Computer Science and Technology
Shenyang Institute of Chemical Technology
Shenyang China
willc@china.com

Shanshan CHEN
Faculty of Computer Science and Technology
Shenyang Institute of Chemical Technology
Shenyang China
54c33@163.com

Yang ZHANG
Faculty of Computer Science and Technology
Shenyang Institute of Chemical Technology
Shenyang China
sujiayang@163.com

***Abstract*—Structural Relation Patterns (SRPs) mining is proposed for mining relations among sequences, these relations are generally hidden behind sequential patterns. Concurrent Sequential Pattern (CSP) and Exclusive Sequential Pattern (ESP) are two important parts of SRP, called Structural Relation Sequential Pattern (SRSP). Upon the previous researches, the concepts of SRSP are redefined; the properties of SRSP are discussed; the algorithms for mining SRSPs are studied. All of these form a theoretical foundation for further study of structural relation patterns and relative mining algorithms. SRSPs mining is significant in practical applications same as sequential patterns mining.**

Keywords-Sequential Patterns Mining; Structural Relation Pattern; Structural Relation Sequential Pattern; Concurrent Relation; Exclusive Relation

I. INTRODUCTION

Structural Relation Patterns (SRPs) mining [1] is a new kind of data mining task, which is proposed based on sequential patterns mining [2] for setting out to find some new structural relation patterns which are generally hidden behind sequential patterns. SRPs mining is valuable in practical applications same as sequential pattern mining.

Graph mining[4.5], Tree mining[6] and Partial Order mining[7-9] are all similar to SRPs mining. By the view of the mining object, Partial Order mining is more similar to SRPs mining, but the structural relations are limited and extended of partial orders.

Concurrent Sequential Pattern (CSP) and Exclusive Sequential Pattern (ESP) are two important parts of SRP[1]. In general, the previous studies have laid a good foundation for researching further into the SRP. But, for the definitions of concurrence and exclusion, the relative relations among sequences were ignored. Therefore, some concepts are redefined with relative relations so that the impact to the relations among sequences caused by huge customer sequences database can be avoided, and the relation patterns among sequences with less frequent but more correlative can be discovered. Then, the mining result could be more complete and significant.

II. DEFINITIONS OF STRUCTURAL RELATION PATTERN

A. Structural relations among sequences

Structural relations among sequences include concurrent relations, exclusive relations, ordered relations and iterate relations. Concurrent relation and exclusive relation are discussed bellow.

Definition 1: Concurrent Relation

Relative to sequence c, sequences $\alpha_1, \alpha_2, \ldots, \alpha_n$ form a Concurrent Relation if they can simultaneously occur in sequence c, denoted by $[\alpha_1+\alpha_2+\ldots+\alpha_n]_c$. In particular, sequences α and β can simultaneously occur in sequence c, denote by $[\alpha+\beta]_c$.

Definition 2: Exclusive Relation

Relative to sequence c, sequences $\alpha_1, \alpha_2, \ldots, \alpha_n$ form an Exclusive Relation if one and only one sequence of sequences $\alpha_1, \alpha_2, \ldots, \alpha_n$ occurs in sequence c, denoted by $[\alpha_1-\alpha_2-\ldots-\alpha_n]_c$. In particular, only one sequence of α and β occurs in sequence c, denote by $[\alpha-\beta]_c$.

Example 1.

For a given Customer Sequences DataBase (CSDB) CSDB={<a(a,b,c)(a,c)d(c,f)>,<(a,d)c(b,c)(a,c)>,<(e,f)(a,b)(d,f)cb>,<eg(a,f)cbc>},

a).Sequences <dcb> and <fbc> are contained in sequence <(e,f) (a,b) (d,f) cb> and <eg(a,f) cbc>, that is：

<cb>∠<(e,f)(a,b)(d,f)cb>and

<fbc>∠<(e,f)(a,b)(d,f)cb>，

then $[<cb>+<fbc>]_{<(e,f)(a,b)\ (d,f)cb>}$

<cb><eg(a,f)cbc>and

<fbc><eg(a,f)cbc>，

then $[<cb>+<fbc>]_{<eg(a,f)cbc>}$

Similarly，

b)$[<fbc>-<a(b,c)(a,c)>]_{<(a,d)c(b,c)(a,c)>}$，

$[<fbc>-<a(b,c)(a,c)>]_{<eg(a,f)cbc>}$

Each structural relation discussed above is based on a sequence, such as sequence c.

B. Structural Relation Sequential Patterns

Suppose Sequential Pattern Set (SP) is the result of sequential patterns mining in a given CSDB, consider the structural relations among the sequential patterns of SP, some sets of the sequential patterns called Structural Relation Sequential Patterns (SRSPs), which consist sequential patterns, concurrent sequential patterns, exclusive sequential patterns, will be built under given conditions.

In the following sections, CSDB is the given customer sequences database, SP is the sequential patterns set mined in CSDB, the expression $|\{\ldots\}|$ denotes the size of a collection.

Definition 3: Concurrence degree

The concurrence degree of sequential patterns α and β in SP is defined as the fraction of the number of customer sequences which let α and β satisfy concurrent relation to the number of customer sequences which contain α or β. The formula is:

978-0-7695-3927-0/09 $26.00 © 2009 IEEE
DOI 10.1109/ICRCCS.2009.75

$$\text{Concurrence}(\alpha,\beta)=\frac{|\{c|[\alpha+\beta]_c,\alpha,\beta\in SP,c\in CSDB\}|}{|\{c|\alpha\angle c\vee\beta\angle c,\alpha,\beta\in SP,c\in CSDB\}|} \quad (1)$$

Generally, the concurrence degree of sequential patterns $\alpha_1,\alpha_2,\ldots,\alpha_n$ in SP is defined as the fraction of the number of customer sequences which let $\alpha_1,\alpha_2,\ldots$and α_n satisfy concurrent relation to the number of customer sequences which contain $\alpha_1,\alpha_2,\ldots$or α_n. The formula is:

$$\text{Concurrence}(\alpha_1,\alpha_2,\ldots,\alpha_n)=\frac{|\{c|[\alpha_1+\alpha_2+\ldots+\alpha_n]_c,\ \alpha_i\in SP,\ c\in CSDB\}|}{|\{c|(\exists i,i=1,2,\ldots n):\alpha_i\angle c,\ \alpha_i\in SP,\ c\in CSDB\}|} \quad (2)$$

The upper definition is different from the one in paper [1]. The denominator of relative fraction in that paper is |CSDB|, while this definition pays more attention to the relations among related sequences.

Definition 4: Concurrent Sequential Pattern (CSP)

Sequential patterns α and β form a Concurrent Sequential Pattern if the condition Concurrence(α,β) ≥mincon is satisfied, denoted by [α+β], where mincon is user specified minimum concurrence threshold.

Generally, Sequential patterns α_1, $\alpha_2,\ldots,\alpha_n$ form a CSP if the condition Concurrence($\alpha_1,\alpha_2,\ldots,\alpha_n$) ≥ mincon is satisfied, denoted by $[\alpha_1+\alpha_2+\ldots+\alpha_n]$.

Definition 5: Exclusive degree

The exclusive degree of sequential patterns α and β in SP is defined as the fraction of the number of customer sequences which let α and β satisfy exclusive relation to the number of customer sequences which contain α or β. The formula is:

$$\text{Exclusive}(\alpha,\beta)=\frac{|\{c|[\alpha-\beta]_c,\alpha,\beta\in SP,c\in CSDB\}|}{|\{c|\alpha\angle c\vee\beta\angle c,\alpha,\beta\in SP,c\in CSDB\}|} \quad (3)$$

Generally, the exclusive degree of sequential patterns $\alpha_1,\alpha_2,\ldots,\alpha_n$ in SP is defined as the fraction of the number of customer sequences which let $\alpha_1,\alpha_2,\ldots$and α_n satisfy exclusive relation to the number of customer sequences which contain $\alpha_1,\alpha_2,\ldots$or α_n. The formula is:

$$\text{Exclusive}(\alpha_1,\alpha_2,\ldots,\alpha_n)=\frac{|\{c|[\alpha_1-\alpha_2-\ldots-\alpha_n]_c,\alpha_i\in SP,c\in CSDB\}|}{|\{c|(\exists i,i=1,2,\ldots,n):\alpha_i\angle c,\alpha_i\in SP,c\in CSDB\}|} \quad (4)$$

Definition 6: Exclusive Sequential Pattern (ESP)

Sequential patterns α and β form an Exclusive Sequential Pattern if the condition Exclusive(α,β) ≥ minxcl is satisfied, denoted by [α-β], where minxcl is user specified minimum exclusive threshold.

Generally, Sequential patterns $\alpha_1,\alpha_2,\ldots,\alpha_n$ form an ESP if the condition Exclusive($\alpha_1,\alpha_2,\ldots,\alpha_n$) ≥ minxcl is satisfied, denoted by $[\alpha_1-\alpha_2-\ldots-\alpha_n]$.

Different from the CSP, an ESP requires that all pairs of the sequential patterns in the pattern are exclusive.

III. PROPERTIES OF SRSPS

The purpose of studying the properties of CSPs and ESPs is to find effective mining algorithms and to provide basic proofs for them. Due to the limitation of pages, only conclusions of the properties are given as followed.

Property 1. Anti-monotone: Remove any branch from a multi-branches CSP or ESP, the left part is also a CSP or ESP, and which is called sub pattern of the original CSP or ESP;

This property may be used to mine CSPs and ESPs within bottom-up method.

Property 2. Exchange rules:

[x+y]=[y+x], [x-y]=[y-x].

Property 3. Association rules:

[x+y+z]=[[x+y]+z]=[x+[y+z]],

[x-y-z]=[[x-y]-z]=[x-[y-z]].

In the process of mining CSPs and ESPs, properties 2 and 3 ensure that any mining order can get same result.

Property 4. Synthesize:

Concurrence(α, β)+Exclusive(α, β) = 1

The sum of concurrence degree and exclusive degree of any two sequences is 1.

IV. SRSPS MINING ALGORITHM

According to the definitions 1 to 6 and the above properties, the following SRSPs mining algorithm is based on sequent patterns set. Let CSDB=$\{c_1,c_2,\ldots,c_m\}$ be the customer sequence database, SP=$\{sp_1,sp_2,\ldots,sp_n\}$ is the result of sequential patterns mining with user specified minimum support threshold minsup.

A. Sequential Pattern Support Vector

The support vector S_i of each sequential pattern sp_i ($1\leqslant i\leqslant n$) is:,

$$S_i=\begin{bmatrix}S_{1i}\\S_{2i}\\\vdots\\S_{mi}\end{bmatrix}$$

where s_{ki} = 1 or 0 ($1\leqslant k\leqslant m$) denotes that customer sequence c_k contains or does not contain sequential pattern sp_i.

For any two sequential patterns sp_i and sp_j, sum the relevant support vector:

$$\text{SUM}=S_i+S_j=\begin{bmatrix}S_{1i}+S_{1j}\\S_{2i}+S_{2j}\\\vdots\\S_{mi}+S_{mj}\end{bmatrix},$$

The SUM is 2-branthes sequence support vector. Let Count(SUM,V) express the number of elements with value V in vector SUM.

The conclusions are as followed:

a)The number of nonzero elements in vector SUM, Count(SUM, V≠0), is the denominator of formulas (1), (3);

b)The number of elements with value 2 in vector SUM, Count(SUM,2), is the numerator of formula (1);

c)The number of elements with value 1 in vector SUM, Count(SUM,1) , is the numerator of formula (3);

Similarly, let SUM = Σ sp_i be the summation of k sequential patterns support vectors, called k-branched sequential support vectors, there are:

d)Count(SUM, V≠0) is the denominator of formulas (2), (4);

e)Count(SUM, k) is the numerator of formula (2), Count(SUM, 1) is the numerator of formula (4).

B. *The algorithm of SRSPs mining based on support vectors*

- **Algorithm:** Support vectors based SRSPs mining algorithm, SupSRSP
- **Input:** Customer Sequences Database (CSDB), minimize support threshold (minnsup), minimize concurrence threshold (mincon) and minimize exclusive threshold (minxcl).
- **Output:** The set of CSPs and ESPs
- **Method:**

a)Mine sequential patterns set in customer sequences database CSDB within minsup. Let $SP=\{sp_1,sp_2,\ldots,sp_n\}$ be the sequential patterns set;

b)Setup all the support vectors for each element of SP, the vectors set is *S*

c)Based on support vector *S*, calculate the SUM of all pairs of sequential patterns support vectors in the set *S*. According to the conclusions of section Ⅳ.A and formulas (1), (3) get 2-branthes CSPs set within mincon and 2-branthes ESPs set within minxcl.

d)According to the properties of SRSPs, conclusions of section Ⅳ.A, and formulas (2) and (4), compose k-branches CSPs set or ESPs set by using (k-1) branches CSPs set or ESPs set. To do so, the support vectors of k-branches sequential patterns should be set by summing each pair of support vectors, one of the pair is from (k-1) branches CSP set or ESP set, and the other one is from *S*; then the support vectors of k-branches CSP or ESP can be gotten.

e)Refine the set of the finding patterns by cutting out contained relationships among the branches of each CSP or ESP.

f)Repeat step *d* and step *e*, until there is no new pattern.

C. *Experiments*

Hereby we gave the result of the experiment for CSPs mining and ESPs mining.

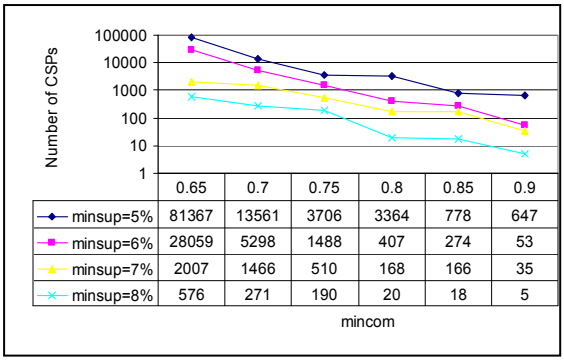
Firstly, we use the test data resource generator provided in paper [10].

The test parameter is below:

The number of customers |D|, the average number of transactions in sequences |C|, the average number of items in a transaction |T|, the potential maximize average length of sequential pattern |S| and the number of different items |N|. In this experiment, let |C|=10, |T|=2.5, |S|=8, |N|=100, |D|=100.

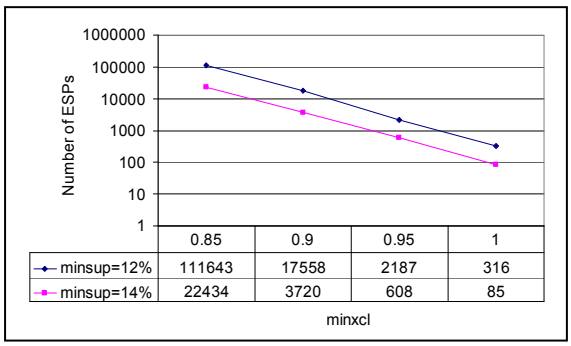
The experimental results are as followed.Set minsup = 5%,6%,7%,and8%,mincon=65%,70%,75%,80%,85%, and 90%, the number of CSPs are shown in Figure 1.

Set minsup=12% and 14%, minxcl=85%,90%,95%,and 100%, the number of ESPs are shown in Figure 2.



	0.65	0.7	0.75	0.8	0.85	0.9
minsup=5%	81367	13561	3706	3364	778	647
minsup=6%	28059	5298	1488	407	274	53
minsup=7%	2007	1466	510	168	166	35
minsup=8%	576	271	190	20	18	5

Figure 1. A part of result of CSPs mining



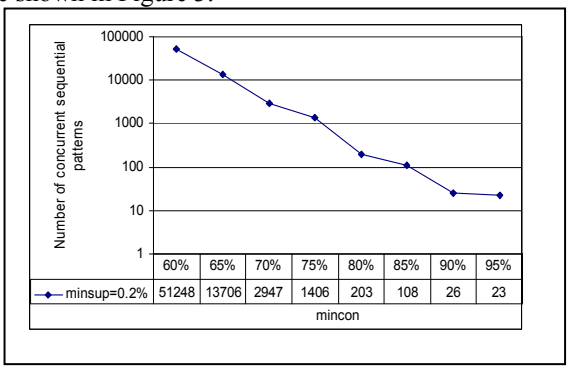
	0.85	0.9	0.95	1
minsup=12%	111643	17558	2187	316
minsup=14%	22434	3720	608	85

Figure 2. A part of result of ESPs mining

Figure 1 is a logarithmic curve diagram that shows the number of CSPs decreases exponentially with the increase in minimum concurrence threshold (mincon), and figure 2 shows the number of ESPs decreases exponentially with the increase in minimum exclusive threshold (minxcl). While we can get a conclusion that the number of ESPs is much more than the number of CSPs under same mining conditions.

Secondly, we mining with the practicality data. The data is coming from data mining web[11] which refers customer purchase sequential data, it contains 1831 customers id, 1927 exchanges and 999 items.

The experimental results are as followed. Set minsup=0.2%, mincon=60%-95%, the number of CSPs are shown in Figure 3.



	60%	65%	70%	75%	80%	85%	90%	95%
minsup=0.2%	51248	13706	2947	1406	203	108	26	23

Figure 3. A part of result of CSPs mining

Figure 3 is a logarithmic curve diagram that shows the number of CSPs decreases exponentially with the increase in minimum concurrence threshold (mincon).

The result of practicality data mining is according with the conclusion in synthesizes data mining. It has validated the correctness and validity of the algorithm.

V. CONCLUSIONS AND FURTHER WORKS

Structural relation patterns mining is a kind of data mining task for mining the structural relations among sequences based on sequential patterns mining. The structural relations among sequences patterns include concurrent, exclusive and etc. Structural relation patterns mining can be used to find some new inherent knowledge which can not be discovered by other methods.

An SRSPs mining algorithm has been researched based on the definitions of CSP, ESP and some relative concepts, and it has been applied in shopping analysis, web access analysis and bio-data analysis as samples. Study on algorithms for mining SRPs, efficient mining algorithms and practical applications are the further works, especially the significance of the application needs to be proved.

REFERENCES

[1] Jing Lu, Osei Adjei, Weiru Chen, Jun Liu. "Post Sequential Pattern Mining: A new method for discovering Structural Patterns". In Proceedings of 2nd International Conference on Intelligent Information Processing, Beijing, China, October 2004 and for Springer Publications

[2] Agrawal R., and Srikant, R. "Mining sequential patterns". Proceedings of the 11th International Conference on Data Engineering, Taipei, Taiwan, 1995, IEEE Computer Society Press, 3-14.

[3] ZHANG Yang, CHEN Weiru, JI Yuan. "Study on algorithm for mining exclusive relation patterns"(in chinese), Computer Engineering and Design.Vol. 29 No. 22, pp.5776-5779, 2008

[4] Kuramochi, M., Karypis, G., "Discover Frequent Geometric Subgraphs", Proceedings of the Second IEEE International Conference on Data Mining (ICDM'02), pp.258-264. 2002

[5] Zaki, M.J., "Efficiently Mining Frequent Trees in a Forest", Proceedings of the SIGKDD, pp.71-80. 2002

[6] Ruckert, U., Kramer, S., "Frequent Free Tree Discovery in Graph Data", Proceedings of 2004 ACM Symposium on Applied Computing, Nicosia, Cyprus, pp.564-570.2004

[7] Jian Pei, Jian Li, Haixun Wang, Ke Wang, Yu, P.S., Jianyong Wang, "Efficiently mining frequent closed partial orders",Data Mining, Fifth IEEE International Conference on,Volume , Issue , 27-30 Nov. 2005 Page(s): 4 pp.

[8] G. Casas-Garriga. Summarizing sequential data with closed partial orders. In SDM, pp. 380-391, 2005.

[9] Guozhu Dong,Jian Pei, "Mining Partial Orders from Sequences", Advances in Database Systems Volume 33, Sequence Data Mining, Springer US , pp.89-112,2007

[10] JI Yuan, CHEN Weiru, ZHANG Xue. "Synthetic method of data resource for concurrent relation patterns", Journal of Shandong Universi(Natural Science),Vol. 42,No. 9,PP.84-87,2007

[11] David Heckerman. MSNBC.com Anonymous Web Data Data Set[DB/OL]. (2001-09-09)[2008-11-14]. http://archive.ics.uci.edu/ml/datasets/MSNBC.com+Anonymous+Web+Data.