Towards Unexpected Sequential Patterns

Haoyuan Li, Anne Laurent, Pascal Poncelet

Towards Unexpected Sequential Patterns

Extended Abstract

Dong (Haoyuan) Li* — **Anne Laurent**** — **Pascal Poncelet*****

* *Laboratoire LGI2P - École des Mines d'Alès*
Parc Scientifique Georges Besse, 30035 Nîmes, France
Haoyuan.Li@ema.fr

** *LIRMM - UMR CNRS 5506*
161 rue Ada, 34392 Montpellier, France
laurent@lirmm.fr

*** *Laboratoire LGI2P - École des Mines d'Alès*
Parc Scientifique Georges Besse, 30035 Nîmes, France
Pascal.Poncelet@ema.fr

RÉSUMÉ. *Dans cet article, nous nous intéressons à la recherche de motifs séquentiels inattendus. Ces derniers représentent des motifs qui apparaissent dans la base de données mais ne respectent pas la croyance que nous avons de ces données. Nous proposons un nouvel algorithme USP basé sur des arbres préfixes pour extraire de telles séquences.*

ABSTRACT. *In this article we are interested in searching unexpected sequential patterns in database that do not respect the beliefs we have. We also propose a new algorithm called USP, based on user beliefs represented as a prefix tree, for mining such sequential patterns.*

MOTS-CLÉS : *Extraction de séquences, motifs séquentiels, base de croyances, inattendu*

KEYWORDS: *Sequence mining, sequential patterns, belief base, unexpectedness*

1. Introduction

The general *support/confidence* based pattern mining approaches use the statistical frequency as the primary interestingness measure in finding potentially interesting patterns in large database. In (McGarry, 2005; Silberschatz *et al.*, 1995; Silberschatz *et al.*, 1996) the *interestingness measures* for data mining are classified as *objective measures* and *subjective measures*. Frequency based criteria are considered as objective and are useful in early phases of data mining where domain knowledge is not yet available. According to these approaches, belief based criteria like unexpectedness are considered as subjective.

Usually unexpectedness measures were mainly considered for association rule algorithms. For instance, (Padmanabhan *et al.*, 1998; Padmanabhan *et al.*, 2006) proposed and improved an belief based unexpected association rules mining approach. In this approach with respect to the belief $X \rightarrow Y$ on dataset $\mathcal{D}$, a rule $A \rightarrow B$ is unexpected if : (i) $B \sim \neg Y$, means that B and Y logically contradict each other ; (ii) The pattern $A \cup X$ holds ; (iii) The rule $A \cup X \rightarrow B$ holds ; (iv) The rule $A \cup X \rightarrow Y$ does not hold. Algorithms derived from the *a priori* algorithm (Agrawal *et al.*, 1994) were also proposed in this approach.

In this article we present a new approach to sequence mining that uses unexpectedness, which is defined by user knowledge (or called *beliefs*), as constraints on finding unexpected behaviors in sequences.

Example 1. *Let us consider the sequential pattern mining in a customer transaction database of supermarket. Assuming we have already know that most of the customers who buy breads and butters like to buy Coca in short future, and who buy always Coca would not like to buy beers. These facts constitute a basic belief base that corresponds to a sequence* $\langle (bread, butter)(Coca) \rangle$ *and a negation relation* $beer \sim \neg Coca$. *According to such user beliefs, the sequential pattern* $\langle (bread, butter)(Coca) \rangle$ *with* $support > 80\%$ *generated by the mining process is valueless since it is already presented in user beliefs, but the sequential pattern* $\langle (bread, butter)(beer) \rangle$ *with* $support = 15\%$ *may be much more valuable because it is unexpected to the belief base and may result in finding undiscovered shopping behaviors.* □

2. Beliefs and Unexpected Sequences

We consider a belief b as a pair of sequence rule p and a set of constraints C, denoted by $b : (p, C)$. A sequence rule p is a relation $s_\alpha \models s_\beta$ where s_α, s_β are two sequences occurring temporally ordered and $t_{end}(s_\alpha) < t_{begin}(s_\beta)$. The constraints set C consists of a contradiction relation $\eta : s_\beta = \neg s_\gamma$ and an expression $\tau : n \ \{<, \leq, =, \neq, \geq, >\} \ N \ (N \in \mathbb{N}), n = 0, n = *$ of temporal order between s_α and s_β. This belief represents that if s_α occurs then s_β should occurs in an order with respect to τ, however if s_β does not occur with respect to τ or s_γ occurs instead of s_α then the sequence s where $s_\alpha, s_\beta \sqsubseteq s$ and $t_{end}(s_\alpha) < t_{begin}(s_\beta)$ is an unexpected sequence.

Example 2. *Given belief* $b : (p, C)$ *where* $p : (A)(C) \models (B)(D)$ *and* $\eta : (B)(D) \sim \neg(E)(F)(G), \tau : n = 0/*$. *Sequence* $s_1 = \langle(A)(B)(C)(B)(C)(D)\rangle$ *is expected to* $(A)(C) \xrightarrow{0} (B)(D)$ *; sequence* $s_2 = \langle(A)(B)(C)(D)(B)(C)(D)\rangle$ *is expected to* $(A)(C) \xrightarrow{*} (B)(D)$ *; sequence* $s_3 = \langle(A)(B)(C)(B)(E)(B)(F)(G)\rangle$ *is unexpected to* $(A)(C) \xrightarrow{0} (B)(D)$ *; sequence* $s_4 = \langle(C)(A)(B)(C)(D)(B)(E)(C)(F)(B)(G)\rangle$ *is unexpected to* $(A)(C) \xrightarrow{*} (B)(D)$. □

The belief base must be consistent. Let b_i and b_j denote two beliefs, l_i denotes the rule part of b_i and l_j of b_j, h_i denotes the left-hand part of l_i and t_i denotes the right-hand part of l_i (and so on for l_j), η_i denotes the contradiction of t_i, we have (1) for a consistent belief base.

$$\forall h_i \sqsubseteq h_j \Longrightarrow \eta_i \not\sqsubseteq t_j \quad [1]$$

A belief constrained by contradiction can be extended to two types of rules. For instance, the sequence rule $p : (A)(B) \models (C)(D)$ and contradiction $\eta : (C)(D) \sim \neg(E)(F)$ imply two rules $(A)(B) \rightarrow (C)(D)$ and $(A)(B) \not\rightarrow (E)(F)$ which can be formally described as follows.

$$(A)(B) \rightarrow (C)(D) \Longleftrightarrow \forall s = (I_1)(I_2)\ldots(I_n) \text{ and } (A)(B)s, \text{ we have}$$
$$\exists i, j \text{ that } i < j, I_i = C \text{ and } I_j = D \quad [2]$$
$$(A)(B) \not\rightarrow (E)(F) \Longleftrightarrow \forall s = (I_1)(I_2)\ldots(I_n) \text{ and } (A)(B)s, \text{ we have}$$
$$\forall i, j \text{ that } i < j, I_j \neq E \text{ or } I_i = E, I_j \neq F \quad [3]$$

Any sequence corresponding to (2) is an expected sequence but any sequence contradicting (3) is an unexpected one.

3. The USP Algorithm

In order to extract unexpected sequences we propose the USP (Mining **U**nexpected **S**equential **P**atterns) algorithm. In this algorithm the belief base $\mathcal{B}$ is represented as a prefix tree and all unexpected sequential patterns with respect to $\mathcal{B}$ and frequent sequential patterns predicated by minimal support σ are stored in another prefix tree $\mathcal{T}$. The USP algorithm uses the PSP (Masseglia *et al.*, 1998) algorithm for appending $\mathcal{T}$ by level.

At each level, the $AppendBeliefs$ routine first appends each $b \in \mathcal{B}$ to each $\mathbf{i} \in \mathcal{T}$ that does not correspond to any belief, then finds unexpected sequential patterns to each $\mathbf{i}_b \in \mathcal{T}$ that corresponds to the first item of any belief, at last removes all non frequent nodes in current path if $\mathbf{i}_b$ does not occurred in any sequence ; the $CountSequence$ routine counts the frequency of each $\mathbf{i} \in \mathcal{T}$ in each path of current level, and finds frequent sequences for next level by the PSP approach. When no more nodes from the belief base can be returned and no more frequent items can be found

in the data set $\mathcal{S}$, the algorithm stops and returns the prefix tree $\mathcal{T}$ that contains item counting information on each node.

Input : Belief base $\mathcal{B}$ represented by a prefix tree, a data set $\mathcal{S}$ of sequences
Output: $\mathcal{T}$

$\mathcal{T} := \emptyset$;
$k := 1$;
$\mathcal{C}_B := AppendBeliefs(\mathcal{B}, \mathcal{T}, k)$;
$\mathcal{C}_k := CountSequence(\mathcal{S}, \mathcal{T}, \mathcal{C}_B, k)$;
while $\mathcal{C}_k \neq \emptyset$ **do**
    $\mathcal{T} := \mathcal{T} \cup \mathcal{C}_k$;
    $\mathcal{C}_B := AppendBeliefs(\mathcal{B}, \mathcal{T}, k+1)$;
    $\mathcal{C}_{k+1} := CountSequence(\mathcal{S}, \mathcal{T}, \mathcal{C}_B, k+1)$;
    $k := k + 1$;
end
return $\mathcal{T}$;

Algorithm 1: Main routine of the algorithm USP.

4. Conclusion

In this article we introduce a new approach to unexpected sequence mining. A belief can be extended to a set of expected and unexpected sequences, with which the computational task to find potentially interesting sequences with multi criteria. Our approach ensures a targeted discovery of unexpected behaviors in sequence mining.

5. Bibliographie

Agrawal R., Srikant R., « Fast Algorithms for Mining Association Rules in Large Databases », *VLDB*, p. 487-499, 1994.

Masseglia F., Cathala F., Poncelet P., « The PSP approach for mining sequential patterns », *PKDD*, p. 176-184, 1998.

McGarry K., « A survey of interestingness measures for knowledge discovery », *Knowl. Eng. Rev.*, vol. 20, n° 1, p. 39-61, 2005.

Padmanabhan B., Tuzhilin A., « A belief-driven method for discovering unexpected patterns », *KDD*, p. 94-100, 1998.

Padmanabhan B., Tuzhilin A., « On characterization and discovery of minimal unexpected patterns in rule discovery », *IEEE Trans. Knowl. Data Eng.*, vol. 18, n° 2, p. 202-216, 2006.

Silberschatz A., Tuzhilin A., « On subjective measures of interestingness in knowledge discovery », *KDD*, p. 275-281, 1995.

Silberschatz A., Tuzhilin A., « What makes patterns interesting in knowledge discovery systems », *IEEE Trans. Knowl. Data Eng.*, 1996.