Mining Negative Sequential Patterns

NANCY P. LIN, HUNG-JEN CHEN, WEI-HUA HAO
Department of Computer Science and Information Engineering
Tamkang University,
151 Ying-Chuan Road, Tamsui, Taipei,
TAIWAN

Abstract: Sequential pattern mining is to discover all frequent sequences from a sequence database and has been an important issue in data mining. A lot of methods have been proposed for mining sequential patterns. However, conventional methods consider only sequences in a sequence database, and the sequential patterns are referred to as positive sequential patterns. In practice, the absence of a frequent itemset in a sequence may imply significant information. We introduce negative sequential pattern concept in which the absence of an itemset in a sequence is also considered. The major difficulties of negative sequential pattern mining are that there may be huge amounts of the candidates of negative sequences and most of them are meaningless. We propose an algorithm for mining negative sequential patterns (NSPM). Using NSPM, we prune a number of redundant candidates by applying apriori-principle, and extract meaningful negative sequences from a large number of frequent sequences using the interestingness measure.

Key-Words: Data mining, Negative sequential pattern, Large sequence

1 Introduction
Sequential pattern mining is to discover all frequent subsequences from a given sequence database and has become an important data mining problem with many divers applications, such as basket analysis, web access patterns and quality control in manufactory engineering, etc. For example, users’ web pages access sequential patterns can be used to improve a company’s website structure to provide more convenient access to the most popular links. The sequential pattern can be divided into Sequential Procurement [1], [2] and Cyclic Procurement [3], [4], [5], [6], [7], [8] by the sequence and the section of time.

A number of methods have been proposed to discover sequential patterns. All the conventional methods for sequential pattern mining were developed to discover positive sequential patterns from database up to now. [1], [8], [9], [10], [11], [12]. Such positive sequential patterns consider only the occurrences of itemsets in a sequence. However, in practice, the absence of an itemset in a sequence may imply valuable information. For example, a web page access sequence can be denoted as <A, B, C, D>, and called a negative sequence. This sequence could give us some valuable information. For example, the link between C and D is need for improving web pages service. However, it is a difficult problem to find such negative patterns because there may be a huge number of candidates generated and most of them are meaningless.

In this paper, we propose a method for mining negative sequential patterns, which can avoid a number of redundant candidates and extract meaningful negative sequences from a large number of frequent negative sequences.

2 Problem Statement
A sequence is an ordered list of itemsets. A positive sequence is denoted by <s1, s2, ..., sn> and a negative sequence is denoted by <s1, s2, ..., ¬sn>, where ¬sn represents the absence of itemset sn. The length of a sequence is the number of itemsets in the sequence. A sequence with length l is called an l-sequence. We may note that a sequence <s1, s2, ..., sn> (or a negative sequence <s1, s2, ..., ¬sn>) can also be written as <s1, s2, ..., sn>, <sn> (or <s1, s2, ..., sn>, ¬sn>). That is a sequence can be regarded as an (n-1)-sequence <s1, s2, ..., sn-1>, denoted by sprev, and called a preceding subsequence, followed by a l-sequence <sn>, denoted by slast, and called a target subsequence. A sequence database D is a set of tuples (cid, s) with primary key cid that is a customer-id, and s that is a customer transaction sequence.

A positive sequence <a1, a2, ..., an> is contained in a sequence <s1, s2, ..., sn> if there exist integers l
Negative Sequential Patterns

Two major difficulties of mining negative sequential pattern problem are: there may be huge amounts of candidates of negative sequence and most of these candidates are meaningless. To overcome the first problem, we design two generation function, \( p_gen() \) and \( n_gen() \), which can generate a few numbers of the candidates of negative sequence. And we adopt the measure of interestingness to solve the second problem. Two generation function, \( p_gen() \) and \( n_gen() \), and the measure of interestingness are introduced as in the following subsections.

### 3.1 Candidates Generation

The generation function of the candidates of positive sequences, \( p_gen() \), includes two phases: the first for generating new candidates and the second for pruning redundant candidates [1]. In the first phase, the candidates of \( k \)-sequences are generated from the set of large positive \((k-1)\)-sequences join with itself. For example, we can generate two candidates \(< s_1, s_2, ... , s_n, b_{n-1} > \) and \(< s_1, s_2, ... , s_n, b_{n-1}, a_{n-1} > \) from \(< s_1, s_2, ... , s_n, b_{n-1} > \) and \(< s_1, s_2, ... , s_n, b_{n-1}, a_{n-1} > \) using \( p_gen() \) function. In the second phase, a candidate of positive \( k \)-sequence will be deleted if any \((k-1)\)-subsequence of it is not a large positive sequence. This is because the apriori-principle states the fact that any super-pattern of an infrequent pattern cannot be frequent.

The generation function of the candidates of negative sequences, \( n_gen() \), includes two phases: the first for generating new candidates and the second for pruning redundant candidates. In the first phase, the candidates of \( k \)-sequences are generated from the set of large positive \((k-1)\)-sequences join with the set of large negative \((k-1)\)-sequences. Note that the way to combine two sequences is slightly different from \( p_gen() \). For example, we combine \(< a_1, s_2, ... , s_n, b_{n-1} > \) and \(< s_1, s_2, ... , s_n, b_{n-1} > \) to generate \(< a_1, s_2, ... , s_n, b_{n-1} > \). In the second phase, a candidate of negative \( k \)-sequence will be deleted if any \((k-1)\)-subsequence of it is not a large negative sequence.

Function: \( n_gen(LP_{k-1}, LN_{k-1}) \)

Parameters:

- \( LP_{k-1} \): Large positive sequences of length \( k - 1 \)
- \( LN_{k-1} \): Large negative sequences of length \( k - 1 \)

Output:

\( CN_k \) : // Negative sequence Candidates

Method:

1. Generating new candidates
2. For each sequence \( p = < p_1, p_2, ..., p_{k-2}, p_{k-1} > \) in \( LP_{k-1} \) do
3. For each sequence \( q = < q_1, q_2, ..., q_{k-2}, q_{k-1} > \) in \( LN_{k-1} \) do
4. If \( ((p_{j+1} = q_j)) \), for all \( j = 1 ... k - 2 \) then
5. Begin
6. new = \( < p_1, p_2, ..., p_{k-1}, q_{k-1} > \)
7. \( CN_k = CN_k \cup \{ new \} \)
8. End
9. Pruning redundant candidates
10. \( CN_k = CN_k - \{ c | c \in CN_k \text{ and any } (k-1)- \text{subsequence of } c \notin LN_{k-1} \} \)
11. Return \( CN_k \)

### 3.2 Measure of Interestingness

There may be a huge number of sequences generated during mining process, and most of them are not interesting. Therefore, defining a function to measure the degree of interestingness of a sequence is needed. Suppose that \( s = < s_1 ... s_n > \) (or \( < s_1 ... s_n > \)), then...
**Algorithm: NSPM**

**Input:**
- \( TD \): Transaction database
- \( \lambda_{ps} \): Threshold of positive sequences
- \( \lambda_{ns} \): Threshold of negative sequences
- \( \lambda_{ni} \): Threshold of interestingness of negative sequences

**Output:**
- \( N \): Negative sequential patterns

**Method:**
1. \( LP_1 = \{ \langle i \rangle | i \in I, \text{Supp}(i) \geq \lambda_{ps} \} \)
2. \( LN_1 = \{ \langle -i \rangle | i \in LP_1 \} \)
3. \( N = \emptyset \)
4. for \( (k = 2; P_{k-1} \neq \emptyset; k++) \) do
5. begin
6. \( // \) Mining positive sequential patterns
7. \( CP_k = p_{-gen}(LP_{k-1}) \)
8. \( LP_k = \{ \langle i \rangle | i \in I, \text{Supp}(i) \geq \lambda_{ns} \} \)
9. \( // \) Mining negative sequential patterns
10. \( CN_k = n_{-gen}(LP_{k-1}, LN_{k-1}) \)
11. \( LN_k = \{ \langle c \rangle | c \in CN_k, \text{Supp}(c) \geq \lambda_{ni} \} \)
12. \( IN_k = \{ \langle l \rangle | l \in LN_k, \text{Im}(l) \geq \lambda_{ni} \} \)
13. \( N = N \cup IN_k \)
14. end
15. return \( N \);

we have the preceding subsequence \( <s_1...s_{n-1}> \), \( s_{pre} \), and the target subsequence \( <s_{n-1}> \) (or \( <-s_{n-1}> \)), \( s_{tar} \).

We define a measure of interestingness as in the following equation:

\[
\text{Im}(s) = \text{Supp}(s) / \text{Supp}(s_{pre}) - \text{Supp}(s_{tar})
\]

If the value of \( \text{im}(s) \) is greater than or equal to a user-predefine threshold, we can predict that \( s_{tar} \) follows \( s_{pre} \) with a relatively high probability. In our method, we use \( \text{Im}(s) \) to measure the degree of interestingness of a sequence and extract meaningful sequences.

### 3.3 Algorithm NSPM

In this algorithm, each iteration \( k \) consists of two phases: the positive sequential patterns mining phase and the negative sequential patterns mining phase. In the positive sequential patterns mining phase (line 6 - 7), the positive candidates of length \( k \), \( CP_k \), are generated from \( LP_{k-1} \) join with \( LP_{k-1} \) by \( p_{-gen} \) function described in 3.1. Then, support of these candidates is counted by scanning the database \( D \) to select large \( k \)-sequences, \( LP_k \). In the negative sequential patterns mining phase (line 10 - 13), the negative candidate sequences of length \( k \), \( CN_k \), are generated from \( LP_{k-1} \) join with \( LN_{k-1} \) by \( n_{-gen} \) function described in 3.1. Next, support of these candidates is counted to determine large \( k \)-sequences \( LN_k \). Then, the value of the interestingness measure function \( \text{im} \) of these large sequences is computed for finding negative sequential patterns \( IN_k \) that we are interested in. Finally, \( IN_k \) is added into \( N \) which contains all negative patterns that have already been mined so far.

### 3.4 Example

Suppose we are given a customer sequence database shown in as Table 1. The threshold of the support of a positive sequence, \( \lambda_{ps} \), the threshold of the support of a negative sequence, \( \lambda_{ns} \), and the threshold of interestingness of a negative, \( \lambda_{ni} \), are set to 0.4, 0.6 and 0.8, respectively. The processes of the algorithm are shown as in table 2 to table 7. The discovered negative sequential patterns are shown as in table 8.

<table>
<thead>
<tr>
<th>CID</th>
<th>Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>C01</td>
<td>( &lt;(1),(2,3,6),(4)&gt; )</td>
</tr>
<tr>
<td>C02</td>
<td>( &lt;(2,3,6)&gt; )</td>
</tr>
<tr>
<td>C03</td>
<td>( &lt;(1),(3,4,7)&gt; )</td>
</tr>
<tr>
<td>C04</td>
<td>( &lt;(2)&gt; )</td>
</tr>
<tr>
<td>C05</td>
<td>( &lt;(1),(2,3,6),(4,5,8)&gt; )</td>
</tr>
</tbody>
</table>

Table 1. Sequence database

In Table 2, all candidates of positive 1-sequences (\( CP_1 \)), their support (\( \text{Supp} \)), large positive 1-sequences (\( LP_1 \)) obtained from \( CP_1 \), and large negative 1-sequences (\( LN_1 \)) obtained from \( LP_1 \) are listed.

<table>
<thead>
<tr>
<th>( CP_1 )</th>
<th>( \text{Supp} )</th>
<th>( LP_1 )</th>
<th>( LN_1 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( &lt;1&gt; )</td>
<td>0.6</td>
<td>( &lt;1&gt; )</td>
<td>( &lt;1&gt; )</td>
</tr>
<tr>
<td>( &lt;2&gt; )</td>
<td>0.8</td>
<td>( &lt;2&gt; )</td>
<td>( &lt;2&gt; )</td>
</tr>
<tr>
<td>( &lt;3&gt; )</td>
<td>0.8</td>
<td>( &lt;3&gt; )</td>
<td>( &lt;3&gt; )</td>
</tr>
<tr>
<td>( &lt;4&gt; )</td>
<td>0.6</td>
<td>( &lt;4&gt; )</td>
<td>( &lt;4&gt; )</td>
</tr>
<tr>
<td>( &lt;5&gt; )</td>
<td>0.2</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>( &lt;6&gt; )</td>
<td>0.6</td>
<td>( &lt;6&gt; )</td>
<td>( &lt;6&gt; )</td>
</tr>
<tr>
<td>( &lt;7&gt; )</td>
<td>0.2</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>( &lt;8&gt; )</td>
<td>0.2</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

Table 2. Positive and negative 1-sequences
In table 3, all candidates of positive 2-sequences (CP$_3$) and large positive 2-sequences (LP$_3$) obtained from CP$_3$ are listed.

<table>
<thead>
<tr>
<th>CP$_3$</th>
<th>Supp</th>
<th>LP$_3$</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;1,2&gt;</td>
<td>0.4</td>
<td>&lt;1,2&gt;</td>
</tr>
<tr>
<td>&lt;1,3&gt;</td>
<td>0.6</td>
<td>&lt;1,3&gt;</td>
</tr>
<tr>
<td>&lt;1,4&gt;</td>
<td>0.6</td>
<td>&lt;1,4&gt;</td>
</tr>
<tr>
<td>&lt;1,6&gt;</td>
<td>0.4</td>
<td>&lt;1,6&gt;</td>
</tr>
<tr>
<td>&lt;2,1&gt;</td>
<td>0</td>
<td>&lt;4,3&gt;</td>
</tr>
<tr>
<td>&lt;2,3&gt;</td>
<td>0</td>
<td>&lt;4,6&gt;</td>
</tr>
<tr>
<td>&lt;2,4&gt;</td>
<td>0.4</td>
<td>&lt;4,6&gt;</td>
</tr>
<tr>
<td>&lt;2,6&gt;</td>
<td>0</td>
<td>&lt;6,2&gt;</td>
</tr>
<tr>
<td>&lt;3,1&gt;</td>
<td>0</td>
<td>&lt;6,3&gt;</td>
</tr>
<tr>
<td>&lt;3,2&gt;</td>
<td>0</td>
<td>&lt;6,4&gt;</td>
</tr>
</tbody>
</table>

Table 3. Positive 2-sequences

Now, we consider negative sequences, in table 4, all candidates of negative 2-sequences (CN$_n$) are generated from the joint of LP$_3$ and LN$_n$. After the comparisons of support (Supp) and measure of interestingness (Im) with $\lambda_{si}$ and $\lambda_{ni}$, large negative 2-sequences (LN$_n$) obtained from CN$_n$, and interested negative 2-sequences (IN$_n$) are obtained and listed.

<table>
<thead>
<tr>
<th>CN$_n$</th>
<th>Supp</th>
<th>Im</th>
<th>LN$_n$</th>
<th>IN$_n$</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;1,2&gt;</td>
<td>0.2</td>
<td>0.13</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>&lt;1,3&gt;</td>
<td>0</td>
<td>-0.2</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>&lt;1,4&gt;</td>
<td>0</td>
<td>-0.4</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>&lt;1,6&gt;</td>
<td>0.2</td>
<td>-0.07</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>&lt;2,1&gt;</td>
<td>0.8</td>
<td>0.6</td>
<td>&lt;2,1&gt;</td>
<td>-</td>
</tr>
<tr>
<td>&lt;2,3&gt;</td>
<td>0.8</td>
<td>0.6</td>
<td>&lt;2,3&gt;</td>
<td>-</td>
</tr>
<tr>
<td>&lt;2,4&gt;</td>
<td>0.4</td>
<td>0.1</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>&lt;2,6&gt;</td>
<td>0.4</td>
<td>0.6</td>
<td>&lt;2,6&gt;</td>
<td>-</td>
</tr>
<tr>
<td>&lt;3,1&gt;</td>
<td>0.8</td>
<td>0.6</td>
<td>&lt;3,1&gt;</td>
<td>-</td>
</tr>
<tr>
<td>&lt;3,2&gt;</td>
<td>0.8</td>
<td>0.6</td>
<td>&lt;3,2&gt;</td>
<td>-</td>
</tr>
<tr>
<td>&lt;3,4&gt;</td>
<td>0.1</td>
<td>0.4</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>&lt;3,6&gt;</td>
<td>0.8</td>
<td>0.6</td>
<td>&lt;3,6&gt;</td>
<td>-</td>
</tr>
<tr>
<td>&lt;4,1&gt;</td>
<td>0.6</td>
<td>0.6</td>
<td>&lt;4,1&gt;</td>
<td>-</td>
</tr>
<tr>
<td>&lt;4,2&gt;</td>
<td>0.6</td>
<td>0.8</td>
<td>&lt;4,2&gt;</td>
<td>-</td>
</tr>
<tr>
<td>&lt;4,3&gt;</td>
<td>0.6</td>
<td>0.8</td>
<td>&lt;4,3&gt;</td>
<td>-</td>
</tr>
<tr>
<td>&lt;4,6&gt;</td>
<td>0.6</td>
<td>0.6</td>
<td>&lt;4,6&gt;</td>
<td>-</td>
</tr>
<tr>
<td>&lt;6,1&gt;</td>
<td>0.6</td>
<td>0.6</td>
<td>&lt;6,1&gt;</td>
<td>-</td>
</tr>
<tr>
<td>&lt;6,2&gt;</td>
<td>0.6</td>
<td>0.8</td>
<td>&lt;6,2&gt;</td>
<td>-</td>
</tr>
<tr>
<td>&lt;6,3&gt;</td>
<td>0.6</td>
<td>0.8</td>
<td>&lt;6,3&gt;</td>
<td>-</td>
</tr>
<tr>
<td>&lt;6,4&gt;</td>
<td>0.2</td>
<td>-0.07</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

Table 4. Negative 2-sequences

In table 5, all candidates of positive 3-sequences (CP$_3$) and large positive 3-sequences (LP$_3$) obtained from CP$_3$ are listed.

<table>
<thead>
<tr>
<th>CP$_3$</th>
<th>Supp</th>
<th>LP$_3$</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;1,2,4&gt;</td>
<td>0.4</td>
<td>&lt;1,2,4&gt;</td>
</tr>
<tr>
<td>&lt;1,3,4&gt;</td>
<td>0.4</td>
<td>&lt;1,3,4&gt;</td>
</tr>
<tr>
<td>&lt;1,6,4&gt;</td>
<td>0.4</td>
<td>&lt;1,6,4&gt;</td>
</tr>
</tbody>
</table>

Table 5. Positive 3-sequences

In table 6, all candidates of negative 3-sequences (CN$_3$) generated from the joint of LP$_3$ and LN$_3$, support (Supp), measure of interestingness (Im), large negative 3-sequences (LN$_3$) obtained from CN$_3$, and interested negative 3-sequences (IN$_3$) are listed.

<table>
<thead>
<tr>
<th>CN$_3$</th>
<th>Supp</th>
<th>Im</th>
<th>LN$_3$</th>
<th>IN$_3$</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;1,2,3&gt;</td>
<td>0.4</td>
<td>0.8</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>&lt;1,2,4&gt;</td>
<td>0</td>
<td>-0.4</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>&lt;1,3,4&gt;</td>
<td>0.4</td>
<td>0.6</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>&lt;1,3,6&gt;</td>
<td>0.2</td>
<td>-0.07</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>&lt;1,4,3&gt;</td>
<td>0.6</td>
<td>0.6</td>
<td>&lt;1,4,3&gt;</td>
<td>-</td>
</tr>
<tr>
<td>&lt;1,5,4&gt;</td>
<td>0.6</td>
<td>0.8</td>
<td>&lt;1,5,4&gt;</td>
<td>-</td>
</tr>
<tr>
<td>&lt;1,4,5&gt;</td>
<td>0.6</td>
<td>0.8</td>
<td>&lt;1,4,5&gt;</td>
<td>-</td>
</tr>
<tr>
<td>&lt;1,5,6&gt;</td>
<td>0.6</td>
<td>0.6</td>
<td>&lt;1,5,6&gt;</td>
<td>-</td>
</tr>
<tr>
<td>&lt;1,6,4&gt;</td>
<td>0.6</td>
<td>0.8</td>
<td>&lt;1,6,4&gt;</td>
<td>-</td>
</tr>
<tr>
<td>&lt;1,5,6&gt;</td>
<td>0.6</td>
<td>0.8</td>
<td>&lt;1,5,6&gt;</td>
<td>-</td>
</tr>
<tr>
<td>&lt;1,6,5&gt;</td>
<td>0.6</td>
<td>0.6</td>
<td>&lt;1,6,5&gt;</td>
<td>-</td>
</tr>
</tbody>
</table>

Table 6. Negative 3-sequences

In table 7, all candidates of negative 4-sequences (CN$_4$) generated from the joint of LP$_3$ and LN$_4$, After the comparisons of support (Supp) and measure of interestingness (Im) with $\lambda_{si}$ and $\lambda_{ni}$, no more sequences are satisfied, therefore we stop.
Table 7. Negative 4-sequences

<table>
<thead>
<tr>
<th>CN4</th>
<th>Supp</th>
<th>Im</th>
<th>LN4</th>
<th>IN4</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;1,2,4,¬3&gt;</td>
<td>0.4</td>
<td>0.8</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>&lt;1,2,4,¬6&gt;</td>
<td>0.4</td>
<td>0.6</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>&lt;1,3,4,¬2&gt;</td>
<td>0.4</td>
<td>0.8</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>&lt;1,3,4,¬6&gt;</td>
<td>0.4</td>
<td>0.6</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>&lt;1,6,4,¬2&gt;</td>
<td>0.4</td>
<td>0.8</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>&lt;1,6,4,¬3&gt;</td>
<td>0.4</td>
<td>0.8</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

Table 8. The discovered negative sequential patterns

Finally, in table 8, all negative sequential patterns discovered are listed.

2-sequences | 3-sequences
---|---
<2,¬3> | <1,3,¬2>
<3,¬2> | <1,4,¬2>
<4,¬2> | <1,4,¬3>
<4,¬3> | 
<6,¬2> | 
<6,¬3> |

Table 8. The discovered negative sequential patterns

4 Conclusion
We introduced negative sequential pattern mining concept in which the absence of itemsets in a sequence are also considered. The major difficulties of negative sequential pattern mining are that there may be huge amounts of negative sequence candidates and most of them are meaningless. In the proposed algorithm NSPM, we reduce a number of redundant candidates by applying the apriori-principle and therefore the computational time is reduced. Additionally, we extract meaningful sequential patterns that we are interested in by using the interestingness measure.

References: