A NEW EFFICIENT METHOD FOR MINING FREQUENT ITEMSETS IN MARKET BASKET DATA ANALYSIS

S.M. Fakhrahmad  
mfakhrahmad@cse.shirazu.ac.ir

M.H. Sadredini  
sadedin@shirazu.ac.ir

M. Zolghadri Jahromi  
zjahromi@shirazu.ac.ir

Abstract: Discovery of hidden and valuable knowledge from large data warehouses is an important research area and has attracted the attention of many researchers in recent years. Most of Association Rule Mining (ARM) algorithms start by searching for frequent itemsets by scanning the whole database repeatedly and enumerating the occurrences of each candidate itemset. In data mining problems, the size of data is often too large to fit in main memory. However, in some cases such as records of sales of a large supermarket, the probability of a particular item to be present in a transaction is too large to fit in main memory. This is due to the fact that a large number of items are usually available for purchase and also the fact that a small set of items is purchased by a customer in a shopping. In this paper, we make use of these facts to propose an efficient method for mining frequent itemsets. In our approach, the database is scanned just once, and data is encoded into a compressed form and held in a proper data structure in main memory. In each iteration, the time required to measure the frequency of itemsets, is reduced further (i.e., enumerating n-dimensional candidate itemsets is much faster than (n-1)-dimensional itemsets). We evaluate the efficiency of our technique using both synthetic and real-life datasets and compare it with other ARM methods proposed in past research.

روشی کارا برای کاوش اقلام پرتنگر در تحلیل داده‌های 
سید خرید

سیدمحمد فخراحمد، محمدهادی صدرالدینی و منصور ذوالفقاری چهرمی

چکیده: کشف الگوهای یکپارچه و ارزشمند از درون حجم وسیعی از داده‌های خام، اخیراً توجه بسیاری از محققان را به خود جلب کرده است. اغلب الگوهای کاوش قوی‌ترین ندایی در مرحله اول کار کوئی اقلام پرتنگر (ساده و ترکیبی) را از بین تمام اقلام موجود در داده‌ها جستجو می‌کنند که این اثر نیازمند به خواندن مکرک گل داده‌ها از دوپسک است. در مسائل داده‌کاپی، حجم پایگاه داده‌های تراکنش معمولاً افزایش زیادی است که قبل بر رشد در حافظه اصلی نیز باید. در این مقاله در مورد الگوی داده‌ای که در حافظه به صورت متغیر به سیستم خریدی یک فروشگاه با توجه به تعداد نسبی زیاد اقلام ممکن (کل اجناس فروشگاه) و نیز محدودیت نسبی اندازه تراکنش‌ها (اقلای خریداری شده در هر بسته)، احتمال رخ داده یک قلم داده (خریداری شده یک کلیه نخی) بیشتر است. در این مقاله با بهره‌گیری از این ویژگی، روشی کارا برای کاوش اقلام پرتنگر در مجموعه داده‌هایی از این قبیل ارائه می‌دهیم. در روایی پیشنهادی، داده‌ها تاکسید یک پار در دوپسک خوده‌ای می‌شوند و بعد از آن به یک ساختار رمز شده و خلاصه‌شکلی‌برداری می‌گردد. بطوریکه اولاً قابل نگهداری در حافظه می‌باشد و ثانیاً با توجه به ساختار خاصی که

تاريخ وصول: 1495/10/20
تاريخ تصویب: 1496/1/13

mfakhrahmad@cse.shirazu.ac.ir
sadedin@shirazu.ac.ir
zjahromi@shirazu.ac.ir
در این روش، عملیات شمارش به سه مرحله نیاز می‌پردازد: 

1. **درخواست نیازمندی دائمی** 
   - **الگو قرار دادن** 
     - برای یک مجموعه از گروه‌های مشابه، اگر گروهی در مجموعه دنباله‌ای عمده شود، در نتیجه آن را به عنوان الگوی دنباله‌ای تعریف می‌کنیم.
   - **تفاوت نیازمندی دائمی** 
     - تفاضل مجموعه‌های می‌شود که برای یک مجموعه از گروه‌های مشابه، مجموعه دنباله‌ای‌هایی که دارای نیازمندی دائمی ندارند، را می‌توانند.

2. **درخواست نیازمندی باربر** 
   - **الگو قرار دادن** 
     - برای یک مجموعه از گروه‌های مشابه، اگر گروهی در مجموعه دنباله‌ای عمده شود، در نتیجه آن را به عنوان الگوی دنباله‌ای تعریف می‌کنیم.
   - **تفاوت نیازمندی باربر** 
     - تفاضل مجموعه‌های می‌شود که برای یک مجموعه از گروه‌های مشابه، مجموعه دنباله‌ای‌هایی که دارای نیازمندی باربر ندارند، را می‌توانند.

3. **درخواست نیازمندی اقدامی** 
   - **الگو قرار دادن** 
     - برای یک مجموعه از گروه‌های مشابه، اگر گروهی در مجموعه دنباله‌ای عمده شود، در نتیجه آن را به عنوان الگوی دنباله‌ای تعریف می‌کنیم.
   - **تفاوت نیازمندی اقدامی** 
     - تفاضل مجموعه‌های می‌شود که برای یک مجموعه از گروه‌های مشابه، مجموعه دنباله‌ای‌هایی که دارای نیازمندی اقدامی ندارند، را می‌توانند.

**کلیدهای وابسته:** داده‌های قوی، الگوی دنباله، اقاقیا پرکرکر، ترکش، تحلیل ضریب خرد

1. **مقادمه** 
   - کاوش قوی‌های قوی به وسیله تحقیقاتی در صحنه با داده‌های قوی است. در راستای کشف ارتباطات جالب و با اهمیت بین اقاقیا اطلاعاتی در باکتری‌های اطلاعاتی پرکرکر و باکتری‌های داده‌های ترکش، می‌باشد که از ارتباطات تحقیقاتی قوی‌های را به بهبود اصلاح داده‌اند. از مسئولیت‌های قوی‌های آن‌ها نیز به سبب مسائل معرفی تحلیل سید

2. **Fast Frequent Pattern Miner**
   - به‌طور کلی، فناوری تحقیقاتی قوی‌های داده‌های قوی برای میانگین‌های قرار داده‌شده داده‌های قوی می‌باشد که در نتیجه آن نوعی از داده‌های قوی شده، مقایسه می‌کنیم.

3. **فناوری بازی‌های یکم‌العملی**
   - برای یک مجموعه از گروه‌های مشابه، اگر گروهی در مجموعه دنباله‌ای عمده شود، در نتیجه آن را به عنوان الگوی دنباله‌ای تعریف می‌کنیم.

4. **Fast FPM**
   - تعریف عمليه مقایسه الگوی پیشنهادی و تعدادی از الگوهای موجود، در بخش ۲ و بخش ۴ ارائه و بخش می‌شود.

5. **متغیر X**
   - تعریف عمليه مقایسه الگوی پیشنهادی و تعدادی از الگوهای موجود، در بخش ۲ و بخش ۴ ارائه و بخش می‌شود.

6. **Consequence**
   - به عنوان مثال، می‌توان نتیجه‌گیری کرد که X به عنوان نتیجه‌گیری می‌شود.
روش گارا برای کاوش مجموعه اقلام پرترکر در تحلیل داده‌های سید خرید

با حذف یک قلم داده از پرترک، درجه پشتیبانی ترکیب حاصل را محاسبه کرده تا نهایت بزرگترین ترکیب یک پرترکر پیدا شود. [11-12]  از آنجا که در مسائل واقعی معمولاً نتوان اقلام داده زاید به وجود آورد و اطمینان خودرویی از طرفی داده‌ها ایجاد نمی‌گردد، شور و دوم (جستجوی بالا به پایین) برای کمک به کنترل در مقایسه با روش اول (جستجوی پایین به بالا) کاربری خودی ندارد. در مرجع [13] یک مقایسه کلی از کاربری الگوریتم‌های مختلف بر روی داده‌ها واقعی ارائه شد.

در مجموع مشکلات عدم روش‌های ذکر شده، نیاز به دسترسی به‌طور معمول اقلام داده جهت کشف اقلام پرترکر است. هدف اصلی الگوریتم‌های k-اریکی‌کاربردی آن‌ها در پایین است.

2. الگوریتم پیشنهادی

برای قبل فهم‌تر شدن الگوریتم پیشنهادی (FastFPM)، از این به بعد پایگاه داده تراکنش‌ها را که در این هیچ ساختار خاصی نیست، به‌صورت یک بانک اطلاعاتیقابل انتقال درون‌دایری قرار می‌دهد. پس کم‌ترین مقدار تنها نشان دهنده کلمی هر مورد به‌کلمه خود یک مجموعه اقلام داده‌ای ممکن و هر کوکردار منتوکارکشیده از پایگاه داده تراکنش‌ها است. هر مقدار یک به‌کلمه در یک تراکنش برترین نشان دهنده و جوابی از عدم وجود کلم در تراکنش است.

Milk, Butter, Bread, Beer

1 1 1 0
1 0 1 0
1 1 0 1
0 1 0 1

شکل 1: دو قلم نمایش یک پایگاه داده تحلیلی مربوط به مسئله سید خرید. (الف) نمایش بدون ساختار تراکنش ها. (ب) نمایش ساختار یافته تراکنش ها.

به عنوان مثال، رابطه نشان دهنده شده در شکل 1(b) نمایش ساختار یافته پایگاه تراکنشی شناسه داده در شکل (الف) است که متشکل بر چهار تراکنش است. در مجموعه داده‌های مربوط به مسئله سید خرید و ساختار مشابه آن تعداد موردی‌یک یک به‌کلمه می‌باشد. این می‌تواند در تعداد اقلام مختلف یک فرودایه و همچنین محصول بودن محصولات آن پیشتر باشد. احتمال بیت 1 در هر ستون از متقابلی دارند. تلاقی بسیاری از این اقلام بر این که نمایش داده شده است.

Milk, Butter

1 1

Bread, Milk, Beer

1 1 0

Bread, Beer

1 1

Milk, Bread, Butter, Beer

1 0 1 0

FastFPM

1 1 0 1

0 1 0 1

فهرست کتاب‌ها و مقالاتی که در راه‌اندازی این الگوریتم مورد استفاده قرار گرفته که آن را به‌طور کامل در مطالعات انتخاب نمایند.

روش‌های ساختاردار

روش‌های ساختاردار نیز جوی دارند که هر یک از کلمات یک پارامتر می‌باشد. [13-17]  به‌جای کلمه اقلام پرترکر از مجموعه داده می‌باشد. همین‌طور ممکن است این واقعیت، اقلام را به‌طور مشابه با کلمات به‌طور ثابت مشاهده نماید. نشان دهنده کلم نشان دهنده و جوابی از عدم وجود کلم در تراکنش است.

روش‌های تعریف دیگری نیز جوی دارند که یکی از این الگوریتم‌های اقلام فشرده می‌باشد. [13-17]  به جای کلمه اقلام پرترکر از مجموعه داده می‌باشد. همین‌طور ممکن است این واقعیت، اقلام را به‌طور مشابه با کلمات به‌طور ثابت مشاهده نماید. نشان دهنده کلم نشان دهنده و جوابی از عدم وجود کلم در تراکنش است.

روش‌های تعریف دیگری نیز جوی دارند که یکی از این الگوریتم‌های اقلام فشرده می‌باشد. [13-17]  به جای کلمه اقلام پرترکر از مجموعه داده می‌باشد. همین‌طور ممکن است این واقعیت، اقلام را به‌طور مشابه با کلمات به‌طور ثابت مشاهده نماید. نشان دهنده کلم نشان دهنده و جوابی از عدم وجود کلم در تراکنش است.

روش‌های تعریف دیگری نیز جوی دارند که یکی از این الگوریتم‌های اقلام فشرده می‌باشد. [13-17]  به جای کلمه اقلام پرترکر از مجموعه داده می‌باشد. همین‌طور ممکن است این واقعیت، اقلام را به‌طور مشابه با کلمات به‌طور ثابت مشاهده نماید. نشان دهنده کلم نشان دهنده و جوابی از عدم وجود کلم در تراکنش است.

روش‌های تعریف دیگری نیز جوی دارند که یکی از این الگوریتم‌های اقلام فشرده می‌باشد. [13-17]  به جای کلمه اقلام پرترکر از مجموعه داده می‌باشد. همین‌طور ممکن است این واقعیت، اقلام را به‌طور مشابه با کلمات به‌طور ثابت مشاهده نماید. نشان دهنده کلم نشان دهنده و جوابی از عدم وجود کلم در تراکنش است.

روش‌های تعریف دیگری نیز جوی دارند که یکی از این الگوریتم‌های اقلام فشرده می‌باشد. [13-17]  به جای کلمه اقلام پرترکر از مجموعه داده می‌باشد. همین‌طور ممکن است این واقعیت، اقلام را به‌طور مشابه با کلمات به‌طور ثابت مشاهده نماید. نشان دهنده کلم نشان دهنده و جوابی از عدم وجود کلم در تراکنش است.

روش‌های تعریف دیگری نیز جوی دارند که یکی از این الگوریتم‌های اقلام فشرده می‌باشد. [13-17]  به جای کلمه اقلام پرترکر از مجموعه داده می‌باشد. همین‌طور ممکن است این واقعیت، اقلام را به‌طور مشابه با کلمات به‌طور ثابت مشاهده نماید. نشان دهنده کلم نشان دهنده و جوابی از عدم وجود کلم در تراکنش است.

روش‌های تعریف دیگری نیز جوی دارند که یکی از این الگوریتم‌های اقلام فشرده می‌باشد. [13-17]  به جای کلمه اقلام پرترکر از مجموعه داده می‌باشد. همین‌طور ممکن است این واقعیت، اقلام را به‌طور مشابه با کلمات به‌طور ثابت مشاهده نماید. نشان دهنده کلم نشان دهنده و جوابی از عدم وجود کلم در تراکنش است.
Hash Table

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

A: يماثل C
B: يماثل C
C: يماثل A

Hash Table

<table>
<thead>
<tr>
<th>كيلبيها</th>
<th>مقدار</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>7</td>
</tr>
<tr>
<td>II</td>
<td>14</td>
</tr>
<tr>
<td>III</td>
<td>11</td>
</tr>
</tbody>
</table>

Hash Table

<table>
<thead>
<tr>
<th>كيلبيها</th>
<th>مقدار</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>7</td>
</tr>
<tr>
<td>IV</td>
<td>13</td>
</tr>
<tr>
<td>VI</td>
<td>15</td>
</tr>
</tbody>
</table>

Hash Table

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

Hash Table

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

Hash Table

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>
لا يمكنني قراءة النص العربي من الصورة في هذه المرحلة. إذا كنت بحاجة إلى المساعدة في شيء آخر، فأنا هنا لمساعدتك!
روشی کارا برای کاوش مجموعه اطلاعات بر تحلیل داده‌های سبد خرید

4. نتایج عملکرد مقایسه الگوریتم‌ها

برای ارزیابی الگوریتم‌های پیشنهادی و مقایسه آن با الگوریتم‌های موجود دو مدل آزمایش تریپل داده‌بندی در حالت اول، آزمایش بر روی داده‌های ساختگی و در حالت دوم عوامل انجام شد. همانطور که می‌دانیم الگوریتم‌های کاوش قوانین تداعی عموماً روال منشی را یافته‌می‌کنند و در شرایط مشابه خروجی‌های کسبی دارند.

بنابراین، معیار مهم در مقایسه روش‌های مختلف کارایی آن‌ها از نظر زمان پاسخ و حافظه مورد نظر می‌باشد. لذا، در این آزمایش‌ها، کارایی الگوریتم پیشنهادی (با در نظر گرفتن $k = 32$) را با الگوریتم‌های معرفت‌برداری و کاراکتر الگوریتم‌های موجود هستند، مقایسه کردیم. الگوریتم معروف دیگری به نام $FP-growth$ نشان داد که الگوریتم معرفت‌برداری به نام $Apriori$ و $ARMOR$ نسبت به الگوریتم $Viper$ و $FastFPM$ کارایی نسبت به کارایی $Apriori$ و الگوریتم $FastFPM$ در مقایسه با $Viper$ و $Apriori$ نسبت به $FastFPM$ وجود دارد. $ARMOR$ با $Viper$ و $Apriori$ کارایی اصلی ۱ گیگابایت و سرعت پردازنده ۳ گیگاهرتز اجرا شده‌اند.

شکل ۵. کارایی روش‌های مختلف بر روی داده‌های تصادفی با اختلال‌های مختلف برای بیت ۱ در داده‌ها

الف) ۵۰/۰۵ (ب) ۰/۰۵ (ت) ۰/۱
جدول 1. مقدار حافظه مصرفی در مراحل مختلف آزمایش 1

<table>
<thead>
<tr>
<th>حافظه مصرفی (مگابایت)</th>
<th>احتمال بیت 1 در داده‌ها</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.5</td>
<td>0.5</td>
</tr>
<tr>
<td>0.7</td>
<td>0.7</td>
</tr>
<tr>
<td>0.9</td>
<td>0.9</td>
</tr>
<tr>
<td>1.1</td>
<td>1.1</td>
</tr>
</tbody>
</table>

آزمایش 2: مجموعه داده‌های واقعی

آزمایش دوم مربوط به وضعیت جمع‌آوری داده‌های واقعی به نام‌های BMS-WebView-2 و BMS-WebView-1 بوده است. مجموعه داده‌ها از داده‌های موجود در بسیاری از اپلیکیشن‌های وب ساخته شده است. در این آزمایش، مقدار حافظه مصرفی توسط کاربران در حال تصادف در این اپلیکیشن‌ها ارائه شده است.

دلاله عبارت برتری آزمایش 2 در مقایسه با آزمایش 1 است. در آزمایش 2، مقدار حافظه مصرفی بین 0.5 تا 1.1 مگابایت بوده است.

در این آزمایش، میانگین حافظه مصرفی بین 0.7 تا 0.9 مگابایت بوده است. در این آزمایش، میانگین حافظه مصرفی بین 0.7 تا 0.9 مگابایت بوده است. در این آزمایش، میانگین حافظه مصرفی بین 0.7 تا 0.9 مگابایت بوده است. در این آزمایش، میانگین حافظه مصرفی بین 0.7 تا 0.9 مگابایت بوده است.
مراجع


معمولاً برقرار است. در این شرایط، با توجه به اختلال بسیار کم حضور یک قلم داده در یک تراکش، محدوده درجه پیشنهادی اقلام مختلف به سیار پایین می‌شود. بنابراین، نمونه‌گیری که در آزمایش‌های مربوط به این موضوع انجام شد، حد آسانه مورد قبول برای درجه پیشنهادی در هنگام کار با این نوع داده‌ها معمولاً عدد کوچکی انتخاب می‌شود.

[14] BMS-WebView-1

(الف)


(ب)

با توجه به تصویربندیقلیدی، ممکن است برای ایجاد کاراگاهی بیشتر در داده‌های مختلف بر روی موضوع بسیار اجرایی کنیم‌هاي مختلف بر روی مجموعه

BMS-WebView-1 (ب) BMS-POS (الف)

BMS-WebView-2 (ب)

شکل 5 زمان اجرایی کلونی‌هاي مختلف بر روی مجموعه


