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

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

M.H. Sadredini
sadredin@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 often very low. 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 مجموعه اقلام پرترکار در تحلیل داده‌های
سید خرید

سیدمحمد فخراحمد، محمدهدایت صدرالدینی و منصور ذوالفقاری جهرمی

چکیده: کشف الگوها و پهنای و ارزشمند از درون حجم ویژه از داده‌های خام، اجرای توجه بسیاری از محققان را به خود جلب کرده است. اغلب روش‌های کاوش قوانین تداوم در مرحله اول کار خود کلیه اقلام پرترکار (ساده و ترکیبی) را از بین تمام اقلام موجود در داده‌ها جستجو می‌کنند که این امر نیازمند به نگهداری کل داده‌ها از دیسک است. در مسائل داده‌کاوی، حجم پایگاه داده‌های تراکنش معمولاً بیشتر است که قابل بردن در حافظه اصلی نیست. اما در برخی موارد مانند پایگاه داده‌های تجربی مربوط به سیدهای خرید یا فروشگاه، با توجه به تعداد نسبتاً زیاد اقلام ممکن (اکثر اقسام فروشگاه)، نیز محدودیت نسبی اندیزه تراکنش‌ها (اقلام خریداری شده در هر سبد)، احتمال رخ دادن یک قلم داده (خریداری شدن یک کالا خاص)، باید است. در این مقاله با پژوهشگری از این ویژگی روشی کارا برای چک اقلام پرترکار در مجموعه داده‌هایی از این قبیل را ارائه می‌دهیم. در روش پیشنهادی، داده‌ها نسبتا یک بار از دیسک خوده‌های مورد و بعد از آن به یک ساختار رسیده و خلاصه نتایج می‌گردد. پتیره‌کلا اولاً قابل نگه‌داری در حافظه‌ای پادشاهی و داده‌ها نسبتا یک بار از دیسک خوده‌های محدود و بعد از آن به یک ساختار رسیده و خلاصه نتایج می‌گردد.
دوازدهم کلیدی: داده‌کاوی، قوانین تندی، اقامه برترکر، تراکنش، تحلیل سیب خرید

1. مقدمه

کاوش قوانین تندی یک زمینه تحقیقاتی در مبحث داده‌کاوی و در راستای کشف ارتباطات جالب و اهمیت بین اقلام اطلاعاتی در بانک‌های اطلاعاتی برگ و پایگاه‌های داده‌ای تراکنشی می‌باشد. این مسئله تلاش‌های تحقیقاتی فراوانی را به خود اختصاص داده است. از مهم‌ترین کاربردهای این نوع تحقیقات به سیستم‌های معروف تحلیل سیب (Apriori) و چندین الگوی دیگر چون [1], [2] و [3] است. با استناد به تحقیق [4], کاربردی یک قانون تندی به شکل $X \rightarrow Y$ می‌باشد که شامل داده‌های مربوط به اپلیکیشن و بهترین قانون مربوط به اکثر آنها به اکثر آنها بازدهی و حداکثر این قانون می‌باشد.

2. تحقیقات مرتب

1. Fast Frequent Pattern Miner


class Data Mining

market basket data analysis

clustering

classification

predecessor

consequence

frequency items
روش گرا برای کاوش مجموعه اقلام پرترک در تحلیل داده‌های سید خرد

واحدی که قلم داده از ترکیب، درجه شبیه‌سازی ترکیب حاصل را مطابقتی درد. تلاش سیسی از بُرُوا بر کاهش تعداد دفعات مراجعه به دیسک جهت خواندن داده‌های

برای این منظور، با تغییر روشهای دیگر راههای مستقیم جهت به دست آوردن بسته‌های خودروی دیسک خودداری می‌کنند و برخی با خطا پذیری ساختار داده‌های خاصی در

Hash Tree

FP-Growth

ARMOR

Apriori

VIPER

که از خارجی داده‌های بیمیشی است. هدف از کلرگی که محققانی که از دسترسی AR گرایش که مد را به دیسک FP-Growth و VEPER روشهای

روشهای متعدد دیگری نیز وجود دارند که هریک باید گام بعد یک مدل مورد حاصل خاصی نیست.

که از دسترسی به مدل FP-Growth و VEPER است. هدف از کلرگی که محققانی که از دسترسی AR گرایش که مد را به دیسکFP-Growth و VEPER روشهای

روشهای متعدد دیگری نیز وجود دارند که هریک باید گام بعد یک مدل مورد حاصل خاصی نیست.

که از دسترسی به مدل FP-Growth و VEPER است. هدف از کلرگی که محققانی که از دسترسی AR گرایش که مد را به دیسکFP-Growth و VEPER روشهای

روشهای متعدد دیگری نیز وجود دارند که هریک باید گام بعد یک مدل مورد حاصل خاصی نیست.

که از دسترسی به مدل FP-Growth و VEPER است. هدف از کلرگی که محققانی که از دسترسی AR گرایش که مد را به دیسکFP-Growth و VEPER روشهای

روشهای متعدد دیگری نیز وجود دارند که هریک باید گام بعد یک مدل مورد حاصل خاصی نیست.

که از دسترسی به مدل FP-Growth و VEPER است. هدف از کلرگی که محققانی که از دسترسی AR گرایش که مد را به دیسکFP-Growth و VEPER روشهای
رابطه شکل 1 (الف) کمتر می‌شود. شرط اصلی برای کارایی الگوریتم پیشنهادی، وجود این خاصیت در مجموعه داده است.

در گام نخست از الگوریتم، داده‌ها را بصورت افتی به مجموعه قطعه با طول مساوی (k) کرده و در هر قطعه، بدون اشتراک تجسم می‌کنیم. در مورد مقدار مناسب برای k، ودعا به وسیله از این مجموعه این اشتقاق با حالت اصلی شده است. با اشتراک این داده‌ها در هر قطعه حداقل دو یک مورد از الگوریتم پیشنهادی، کل الگوریتم

برترکنید برروی کت جستجو می‌شود. در گام اول بایستی دفعات ثروه کمیک از الگوریتم تکی شده‌اند و برای هر کدام که پیشنهادی اش حداقل با حد استانه پایین‌تر است، و دیگری که می‌شود متقابل با الگوریتم A در شکل 2 نشان داده شده است.

| کلیه‌ها | متقابل
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>II</td>
</tr>
<tr>
<td>9</td>
<td>14</td>
</tr>
<tr>
<td>III</td>
<td>V</td>
</tr>
</tbody>
</table>

شکل 2. جدول‌های در هر ساز مربوط به (الف) A، (ب) B

جهت محاسبه پیشنهادی تک ترکیب از الگوی مانند AB دیگر بایستی به پیشنهاد کل داده‌ها نیست و این کار با استفاده از جدول‌های در هر ساز عنصر ان (یعنی A و B) انجام می‌شود. برای این منظور از جدول در هر ساز گنجینش (شامل عنصر کمتر) شروع می‌کنیم. پس از ترکیب کلی که در جدول دوم می‌توان آورده شود، یک مقدار صحیح مerto با آن AND از دو جدول در هر ساز را با هم می‌گیریم. تنها عمل AND می‌تواند کلیک در حالت ارگی که مکنای در میان 2 نشان دهنده محل و تعداد رخ‌دادن همزمان عنصر A و B در قطعه مربوطه است.

| کلیه‌ها | متقابل
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>IV</td>
</tr>
<tr>
<td>7</td>
<td>13</td>
</tr>
<tr>
<td>V</td>
<td>VI</td>
</tr>
</tbody>
</table>

رابطه فقط‌ننی‌شده فقط یک بار پیش‌شده می‌شود و پیشنهاد
کلیه الگوریتم تکی (1 بر 9) محاسبه می‌گردد تا الگوریتم ترکیب
کشف شود. در ضمن این عمل مقدار به‌طور غیر از (اعداد
مربوط به هر یک از قطعه‌ها) از تمام فتقها استخراج می‌شود.

باز از شروع در ساخته که قطعه مربوط به آن پرترکر
پاژی، یک جدول است دستیسی مستقیم (جدول در هر ساز) در هر
سازیم. مقدار دیگری مشابه در این جدول، همان اعداد صحیح

در صورت که میرک از قطعه‌ها انتخاب دو الگوریتم
کلید

دستیسی مستقیم به هریک از مقدار می‌شماره قطعه مربوط به آن
می‌شود (جدول 1 و حاکم برای تعداد قطعه‌ها)، از آنجا که ما

می‌کنیم. این الگوریتم مربوط به صفر (تعداد قطعه‌ها) را در جدول‌های
در هر ساز درج می‌کنیم. الگوریتم دستیسی مستقیم در هر جدول
در هر ساز به همکاری با پیش‌شده‌که قطعه مربوط به آن

را باید کنید از میان 1 تا 24 این آمار در مراحل
بعدی الگوریتم، یعنی جستجوی الگوریتم با ابعاد بالاتر به مراتب

1. Hash Table
روش گارا برای کاوی مجموعه اقلام پرترکر در تحلیل داده‌های سبد خرید

1-3 تحلیل داده‌های سبد خرید

گروهی که ممکن است اغلب از جمله شما که مدت زمانی طولانی در بازار معاملاتی واقع شده باشید، با احتمال بالا در حال خرید اقلام پرترکر توانسته باشند. این گروه برای کمک به کسانی که در صورت انتخاب اقلام پرترکر در سبد خرید شتاب آورده می‌شود. برای اینکه بتوانید از این نوع کمک بهره ببرید، می‌توانید بررسی کنید که چگونه از اقلام پرترکر در سبد خرید خود استفاده می‌کنید.

**شماره 1 جدول در هر سبد مربوط به AB**

<table>
<thead>
<tr>
<th>شماره 1</th>
<th>مقدار</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
</tr>
</tbody>
</table>

**شماره 2 جدول در هر سبد مربوط به AB**

<table>
<thead>
<tr>
<th>شماره 2</th>
<th>مقدار</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>4</td>
</tr>
</tbody>
</table>

**شماره 3 جدول در هر سبد مربوط به AB**

<table>
<thead>
<tr>
<th>شماره 3</th>
<th>مقدار</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>6</td>
</tr>
</tbody>
</table>

**شماره 4 جدول در هر سبد مربوط به AB**

<table>
<thead>
<tr>
<th>شماره 4</th>
<th>مقدار</th>
</tr>
</thead>
<tbody>
<tr>
<td>7</td>
<td>8</td>
</tr>
</tbody>
</table>

**شماره 5 جدول در هر سبد مربوط به AB**

<table>
<thead>
<tr>
<th>شماره 5</th>
<th>مقدار</th>
</tr>
</thead>
<tbody>
<tr>
<td>9</td>
<td>10</td>
</tr>
</tbody>
</table>
دیوید رایت، مقدمه‌ای از انسان خردسالی، ۱۹۹۳

در این مطالعه، به توصیفی که در انسان‌های خردسالی انجام می‌شود، اشاره می‌گردد. این توصیف شامل چندین گروه از عوامل مختلفی می‌باشد که در فکری اندام شاید به خردسالی اشاره کنند. این عوامل شامل امراض و افزایش درجه نارضایتی در افراد خردسالی می‌باشند.

2.2 مقدار مناسب

یک مثال مهم از اینگونه اعتقاد می‌تواند در صورتی باشد که در انتخاب محدوده مناسبی از فاکتورها به ترتیب اینگونه انتخاب می‌شود. این مقدار مناسب به‌طور کلی به شرح زیر است:

- انتخاب یک اعداد مناسب
- محاسبه طول دوره مطالعه
- تعریف اجمالی
- محدوده‌بندی
- اندازه‌گیری

2.3 تعیین اساسهای پیشنهادی

نتیجه‌گیری‌های این مقاله به‌صورت اصلی بر اساس این وابستگی نتایج دقیق در جهت پیشنهادی است. مطالعات زمانی که در انتخاب‌هایی به‌طور کلی به‌صورت واحدی می‌باشند، این مقدار مناسب را در محدوده انتخاب می‌سازند. این به‌رونمایی از این است که در انتخاب‌هایی مانند این، این مقدار مناسب را در محدوده انتخاب می‌سازند.
روشی کارا برای کاوش مجموعه اقلام پر بردگان در تحلیل داده‌های سبد خرید

4- نتایج عملی مقایسه الگوریتم‌ها

برای ارزیابی الگوریتم‌های مختلف و مقایسه آن با الگوریتم‌های موجود در مدل آزمایش ترتیب داده شده، در حالت اول، آزمایش بر روی داده‌های ساختگی و در حالت دوم روی داده‌های واقعی انجام شد. همانطور که می‌دانیم الگوریتم‌های کاوش قوانین تداعی عموماً روال مشخصی را تی می‌کند و در شرایط مشابه خروجی‌های پیشینی دارند.

بنابراین، معیار مهم در مقایسه روش‌های مختلف کارایی آنها از لحاظ زمان پاسخ و حافظه مورد نظر باید دو می‌باشد. بنابراین نتایج آزمایش در مدل آزمایش اول با اعمال الگوریتم‌های معرفی گردیده بر روی داده‌های مورد نظر بوده و محدود نشان داده می‌شود. میزان اجرایه الگوریتم در مدل آزمایش اول به عنوان یک مجموعه داده مختلف است. جدول 1 حاصل حافظه کمیابی الگوریتم (جهت تحلیل جدول‌های مشابه در هر سر) را در هر مرحله از آزمایش نشان می‌دهد. از مقایسه آن بحث می‌باشد. انتخاب مجموعه داده اولیه (تقییاً 100 مگابایت) به‌منظور به‌بیناندن بزرگی الگوریتم پی. برداشته با توجه به شکل 5 بیشینه که یک برتری قابل توجه برای کارایی این الگوریتم و FastFPM در مقایسه با یک برتری یکی در مقایسه با یکی وجود دارد.

این نتایج با FastFPM و ARMOR بوده است. صورت بسته که یک برتری قابل توجه برای کارایی FastFPM و ARMOR بوده است. صورت بستهی زمان حافظه بسیار زیادی دارد، این تجویز آن را برای انتخاب مجموعه داده‌های واقعی که کوچک‌ترین حافظه دارد. الگوریتم‌های C++ و بر روی یک سیستم با حافظه است. اصل یک گیگابایت و سرعت پردازندگی گیگاهرتز اجرا شده است.

شکل 5 - کارایی روشهای مختلف بر روی داده‌های تصادفی با احتمال های مختلف برای بیت 1 در داده‌ها

(الف) $0/0/5$ (ب) $0/0/5$ (ت) $0/0/1$
دسته‌بندی و معیار
را در ارزیابی کارایی الگوریتم‌ها اعمال می‌کرد.

زمان اجزای الگوریتم‌ها مختلف برای کشف مجموعه قاطع‌ترین برترکار
و تولد قانون بر روی سهمیه‌گزاری مشخص اندازه گیری و در شکل‌های ۱-۲ (د) تا ۵ (ب) نمایش داده شده است. همانطور که در شکل‌ها دیده می‌شود، زمان اجزای الگوریتم بسیاری شده‌است (FastFPM)
که از آنجا روی دامنه حرف‌کاری ARMOR
است. می‌باشد.

نتیجه گیری
به دلیل اهمیت و کاربرد زیاد قوانین داخلی روش‌های زیادی جهت کشف الگوها در برابر دقیق‌ترین قاطع‌ترین برترکار است. به‌طور کلی، این کشف مجموعه قاطع‌ترین برترکار به‌طور کامل شده است. مسئله با توجه به اینکه مجموعه اقامت غیرین کشف شده با گم شدن حضور است. پشتیبانی صورت نماینده از مایه‌ای، قبل از درخواست مورد نظر الگوریتم
با وجود اینکه تعداد روش‌ها کاملاً داخلی در این راستا پشتیبانی می‌شود. اما در اغلب موارد راه بی‌پایان چند فصل تمرکز آنها هماهنگ خواهد در فضای گروهی به‌طور کامل زیر دانسته و از مراحل
به‌طور کلی، این کشف مجموعه قاطع‌ترین برترکار به‌طور کامل است. در این راستا پشتیبانی می‌شود. برای ارزیابی کارایی این روش مقارن است با روی دامنه حرف‌کاری ARMOR.

جدول ۱. مقادیر حافظه مصرفی در مراحل مختلف آزمایش ۱

<table>
<thead>
<tr>
<th>حافظه مصرفی (مگابایت)</th>
<th>احتمال بیت ۱ در داده‌ها</th>
</tr>
</thead>
<tbody>
<tr>
<td>۶/۴</td>
<td>۰/۰۵</td>
</tr>
<tr>
<td>۱۱/۳</td>
<td>۰/۱۰</td>
</tr>
<tr>
<td>۲۴/۶</td>
<td>۰/۵۰</td>
</tr>
<tr>
<td>۸/۸۷۳۷۱</td>
<td>۰/۱۱</td>
</tr>
</tbody>
</table>

۲-۱ آزمایش ۱: مجموعه داده‌ای واقعی
آزمایش دوم ما بر روی مجموعه داده واقعی به نام‌های BMS-View-۲ و BMS-POS می‌باشد. مجموعه‌ای از داده‌های مربوط به اقلام فروخته شده BMS-POS یک فروشگاه از اتمسفریک بزرگ است که ساکورا، که در آن اقامت گزارش‌های جدید این اقلام داده شده است. مجموعه همان‌اکنون این اقلام در کتاب‌خانه یک راهکرد را تشکیل می‌دهد. هدف از جمع آوری این داده بیان انتباهی بین اقلام خریداری شده بود. این مجموعه داده مستقل بر

۲-۲ آزمایش ۲: مجموعه داده‌ای واقعی
هم شمار داده‌های BMS-View-۲ و BMS-POS مربوط به مراحل کاربران طی چند ماه به سایریهار و فروشگاه
اینترنتی است که بر اساس مجموعه‌های که کاربر در هر پار مراحل دارد، تبدیل به یک پار مراحل محسوب می‌شود. در نتیجه، این کاربر در کتاب‌خانه
با هدف از جمع آوری این داده هم باعث بیان انتباهی بین اقلام بازیابی‌های همزمان است. این داده مشابه داده داده‌های BMS-View-۲ و BMS-POS مربوط به مراحل کاربران طی چند ماه به سایریهار و فروشگاه
اینترنتی است که بر اساس مجموعه‌های که کاربر در هر پار مراحل دارد، تبدیل به یک پار مراحل محسوب می‌شود. در نتیجه، این کاربر در کتاب‌خانه
با هدف از جمع آوری این داده هم باعث بیان انتباهی بین اقلام بازیابی‌های همزمان است. این داده مشابه DAS انتباهی بین اقلام بازیابی‌های همزمان است. می‌باشد. در این آزمایش‌ها حد آستانه مورد

MinConf


