به نام خدا

دانشگاه تهران

دانشکدگان فنی

دانشکده مهندسی برق و کامپیوتر

درس داده کاوی

تمرین کامپیوتری دوم(CA2)

نام و نام خانوادگی : حسین سیفی

شماره دانشجویی : ۸۱۰۱۰۰۳۸۶

اردی‌بهشت ۱۴۰۱

فهرست

تمرین‌های تشریحی

سوال ۱

الف

ابتدا جدول Itemهای مکرر کاندید و ساپورت آن‌ها را ایجاد می‌کنیم. جدول ایجاد شده با عنوان C1 قابل مشاهده است:

Table 1: C1

Item	Support
A	4/4=1
B	4/4=1
C	2/4=0.5
D	3/4=0.75
E	2/4=0.5
K	1/4=0.25

همانطور که در جدول ۱ قابل مشخص است، Itemهای C، E و K مقدار ساپورتی کمتر از حداقل ساپورت دارند و در نتیجه از مجموعه Itemهای مکرر حذف می‌شوند. Itemهای مکرر در جدول ۲ قابل مشاهده هستند:

Table 2: L2

Item	Support
A	1
B	1
D	0.75

سپس به ایجاد Itemsetهای کاندید با اندازه ۲ می‌پردازیم که از Itemهای مکرر موجود در جدول L1 ایجاد شده‌اند. این Itemsetها در جدول C2 قابل مشاهده هستند:

Table 3: C2, L2

Itemset	Support
A, B	1
A, D	0.75
B, D	0.75

با توجه به جدول فوق، هیچکدام از Itemsetها شرط حداقل ساپورت را نقض نمی‌کنند و در نتیجه جدول L2 به شکل جدول C2 خواهد بود. با استفاده از محتویات سطرهای جدول ۳ می‌توان Itemsetهای کاندید با طول ۳ را ایجاد کرد. این Itemsetها در جدول ۴ قابل مشاهده هستند:

Table 4: C3, L3

Itemset	Support
A, B, D	0.75

با بررسی جدول ۳ متوجه می‌شویم که تنها یک Itemset با اندازه ۳ می‌توان ایجاد کرد و Itemset ایجاد شده نیز شرط حداقل ساپورت را نقض نمی‌کند و در نتیجه جدول ۴ همان جدول L3 نیز خواهد بود. به دلیل اینکه تنها یک Itemset با اندازه ۳ وجود دارد امکان ایجاد جدول C4 وجود ندارد و الگوریتم در همین نقطه متوقف می‌شود. تمامی Itemsetهای مکرر ایجاد شده به شرح زیر می‌باشند:

Frequent itemsets: A, B, D, {A, B}, {A, D}, {B, D}, {A, B, D}

ب

در جدول زیر Association Ruleهایی که شروط Min_support و Min_confidence را برآورده می‌کنند و Metarule مورد نظر مطابقت دارند به همراه ساپورت و Confidence آورده شده‌اند:

Table 5: Strong Association Rules

Item 1	Item 2	Item 3	Support	Confidence
A	D	B	0.75	1
B	D	A	0.75	1

با توجه به Itemsetهای مکرر معرفی شده در بخش الف سوال، می‌توانستیم حالت {A,B}→D را نیز به جدول فوق اضافه کنیم اما این حالت شرط Min_confidence را نقض می‌کرد بنابراین در این جدول جایی ندارد. همچنین در هر دو Association Rule فوق در حالتی که Item1 و Item2 جابجا شوند نیز صادقند.

سوال ۲

الف

معیار لیفت برای همبرگر و هات‌داگ به شکل زیر محاسبه می‌شود:

$$Lift(Hamburgers, Hotdogs) = \frac{\frac{2000}{5000}}{\frac{3000}{5000} * \frac{2500}{5000}} = \frac{4}{3}$$

مقدار محاسبه شده لیفت برای همبرگر و هات‌داگ از ۱ بزرگتر است بنابراین معیار لیفت وابستگی مثبت(Positive correlation) را بین این دو کالا نشان می‌دهد.

ب

معیار All-Confidence و Cosine برای محصولات همبرگر و هات‌داگ به شکل زیر محاسبه می‌شود:

$$All\_Confidence(Hamburgers, Hotdogs) = \frac{2000}{\max\{2500,3000\}} = \frac{2}{3}$$

$$Cosine(Hamburgers, Hotdogs) = \frac{2000}{\sqrt{2500 * 3000}} \approx 0.73$$

هر دو معیار All-Confidence و Cosine مقداری بیش از ۰.۵ را نشان می‌دهند و در نتیجه نشانگر وابستگی مثبت(Positive correlation) بین دو محصول مورد نظر هستند.

معیار لیفت ویژگی Null_invariant را ندارد و دو معیار دیگر دارای این ویژگی هستند اما با توجه به اینکه تعداد تراکنش‌های Null(تراکنش-هایی که شامل هیچکدام از محصولات همبرگر و هات‌داگ نباشند) زیاد نیست(کمتر از ۳۰ درصد کل تراکنش‌ها را شامل می‌شود)، هر سه معیار وابستگی یکسانی را بین محصولات مورد نظر نشان می‌دهند و در صورتیکه تعداد تراکنش‌های Null زیاد باشد این معیارها می‌توانند نتایج متفاوتی را نشان دهند. در چنین شرایطی معیارهای Cosine و All-Confidence نسبت به معیار لیفت نتایج معتبرتری دارند.

سوال ۳

برای محاسبه Itemsetهای مکرر با استفاده از الگوریتم FPgrowth ابتدا باید ساپورت هر یک از Itemها را محاسبه کنیم.

Table 6: Items' support

Item	Support	Value
Milk	3	3000
Butter	2	2500
Peanut	5	2300
Chips	4	2000
Cake	2	1500
Cheese	3	1200
Water	1	1000
Tea	1	3000

F-list = {Milk, Butter, Peanut, Chips, Cake, Cheese}

دو محصول Water و Tea مقدار ساپورتی کمتر از حداقل ساپورت مورد نظر دارند و در نتیجه از جدول فوق حذف می‌شوند و در ادامه محاسبات مورد استفاده قرار نمی‌گیرند. همچنین F_list نیز به ترتیب Value محصولات به شکل فوق درمی‌آیند.

در جدول ۷ تراکنش‌ها به ترتیب Value محصولات و با حذف محصولات Water و Tea بازنویسی شده‌اند.

Table 7: Transactions

TID	Items
100	Milk, Butter, Peanut, Cake
200	Peanut, Chips, Cake
300	Peanut, Chips, Cheese
400	Milk, Butter, Peanut, Chips, Cheese
500	Milk
600	Peanut, Chips, Cheese

در پس از محاسبات اولیه FP-tree به شکل زیر درمی‌آید:

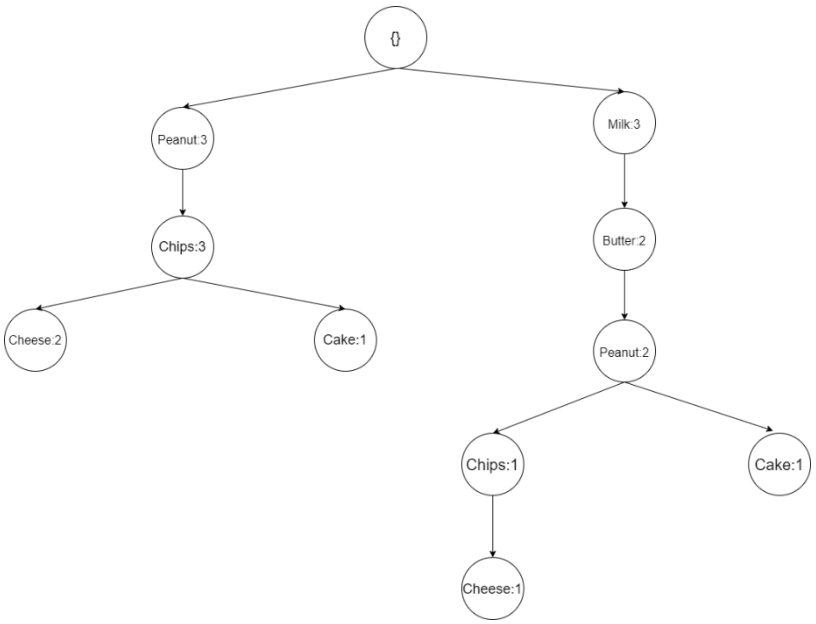

Figure 1: FP-Tree

سپس به ایجاد جدول Conditional pattern baseها می‌پردازیم. جدول به شکل زیر خواهد بود:

Table 8: Conditional Pattern Base for Each Item

Item	Conditional Pattern Base	FP-Tree
Cheese	{Peanut,Chips}:2, {Milk,Butter,Peanut,Chips}:1	<Peanut:3,Chips:3>
Cake	{Peanut,Chips}:1, {Milk,Butter,Peanut}:1	<Peanut:2>
Chips	{Milk,Butter,Peanut}:1, {Peanut}:3	<Peanut:2>

برای سه متغیر دیگر Conditional Pattern Base ایجاد نمی‌شود چرا که در هر شرایط و با هر Itemset ممکن، شرط حداقل Value نقض می‌شود. در Conditional Pattern Baseهای محصولات Cheese، Cake و Chips به ترتیب محصولات {Milk, Butter}، {Milk,Chips,Butter} و {Milk,Butter} به دلیل اینکه دارای ساپورتی کمتر از حداقل ساپورت هستند حذف می‌شوند و با حذف هر کدام از این مجموعه محصولات از Condtional Pattern Base جدول فوق، در درخت هرکدام از این محصولات تنها یک شاخه باقی می‌ماند و در این شرایط به راحتی می‌توان Itemsetهای مکرر هر کدام از درخت‌ها را محاسبه کرد(درخت‌ها رسم نشده‌اند اما شیوه نوشتاری آن‌ها در جدول فوق موجود است). جدول زیر Itemsetهای مکرر شامل هر کدام از محصولات جدول فوق را نشان می‌دهد:

Table 9: Frequent Patterns

Item	Frquent Patterns
Cheese	{Peanut,Chips}:3, {Peanut,Cheese}:3, {Peanut,Chips,Cheese}:3, {Chips,Cheese}:3
Cake	{Peanut,Cake}:2
Chips	{Peanut,Chips}:2

در جدول ۹ Itemset مشخص شده با رنگ قرمز تکراری است و در نتیجه نیاز به ذکر کردن دوباره آن نیست. در کنار Itemsetهای موجود در جدول فوق Itemهای Cheese، Chips و Cake نیز جزو Itemsetهای مکرر هستند که دارای شرط Succinct مورد نظر می‌باشد در نتیجه لیست نهایی مورد نظر به شکل زیر می‌باشد:

Frequent Itemsets: [{Peanut,Chips}, {Peanut,Cheese}, {Peanut,Chips,Cheese}, {Chips,Cheese}, {Peanut,Cake}, {Cake}, {Chips}, {Cheese}]

تمرین‌های عملی

سوال ۱

مجموعه داده موجود در این سوال که در قالب CSV قرار دارد را می‌توان به عنوان یک دیتافریم به کمک کتابخانه Pandas در محیط پایتون بارگزاری کرد. ابتدا می‌خواهیم لیستی از تمامی Itemهای موجود در مجموعه تراکنش‌ها به دست بیاوریم ولی در کنار این کار می‌توانیم تعداد حضور هر Item در تراکنش‌ها را محاسبه کنیم. بدین منظور یک دیکشنری ایجاد کردیم که کلیدهای موجود در آن نام Itemها هستند و مقادیر متناظر با هر کلید تعداد حضور آن Item در مجموعه تراکنش‌ها است و نام آن را Items می‌گذاریم. سپس با توجه به استفاده از کتابخانه MLxtend برای استخراج Itemsetهای مکرر و Association ruleهای قوی نیاز داریم تا مجموعه داده سبدهای خرید را به فرمت ورودی مورد نظر این کتابخانه تبدیل کنیم. فرمت ورودی مورد نیاز یک دیتافریم است که به ازای هر تراکنش یک سطر و به ازای هر Item یک ستون دارد و مقادیر موجود در این دیتافریم از نوع باینری هستند و هر خانه نشان دهنده حضور یا عدم حضور یک Item مشخص در یک تراکنش مشخص است.

نمودار زیر میزان فروش هر یک از Itemها را نشان می‌دهد:

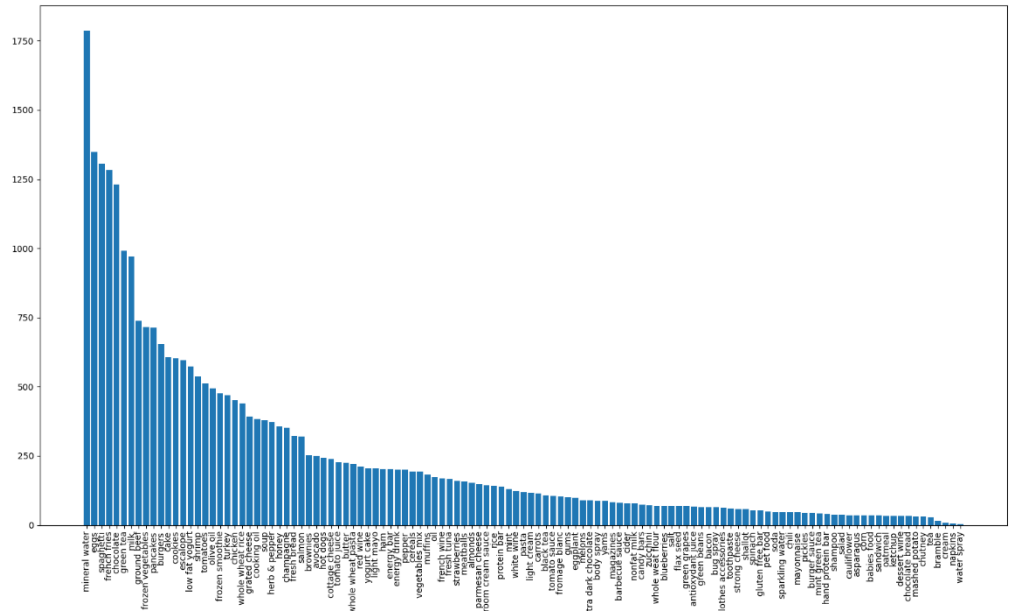

Figure 2

نمودار ۱ نشان می‌دهد بین خرید Itemهای متفاوت توازن وجود ندارد. تعداد کمی از Itemها در ۱۴ درصد تراکنش‌ها حضور دارند که مقدار زیادی است و از طرف دیگر تعداد زیادی از Itemها وجود دارند که در کمتر از ۰.۰۴ تراکنش‌ها حضور دارند. با توجه به نمودار فوق برای پیدا کردن Itemsetهای مکرر باید مقدار حداقل ساپورت را به دقت انتخاب کنیم چرا که حتی تغییر ۱ درصدی این مقدار می‌تواند موجب حذف تعدادی زیادی از روابط معنادار بین Itemها می‌شود. همچنین با استفاده از این نمودار Itemهایی با بیشترین و کمترین فرکانس حضور در سبدهای خرید را می‌توان به راحتی تشخیص داد. سه Item با بیشترین فرکانس آب معدنی، تخم مرغ و اسپاگتی و سه Item با کمترین فرکانس اسپری آب، دستمال و bramble هستند.

سوال ۲

الف

تعداد تراکنش‌های موجود در مجموعه داده برابر با تعداد سطرهای آن و به تعداد ۷۵۰۱ عدد می‌باشد.

ب

تعداد اقلام متمایز موجود در سبدهای خرید که با تعداد سطرهای ایجاد شده در دیکشنری Items برابر است به تعداد ۱۱۹ می‌باشد.

ج

برای پاسخ به این سوال با استفاده از دیکشنری Items ایجاد شده می‌توان ۵ کلیدی را یافت که بیشترین مقدار را دارند. جدول زیر ۵ جنس پرفروش به همراه تعداد فروش آن‌ها را نشان می‌دهد:

Table 10: 5 items with most sales

Item	Sales
Mineral water	1788
Eggs	1348

Spaghetti	1306
French fries	1282
Chocolate	1230

د

تعداد تراکنش‌های شامل محصول Black Tea برابر مقدار متناظر با کلید Black Tea در دیکشنری Items ایجاد شده و برابر ۱۰۷ است.

سوال ۳

برای سوالات عملی که نیاز به اعمال الگوریتم‌های Apriori و FPGrowth دارند می‌توان از کتابخانه‌های Apyori و MLxtend استفاده کرد. به دلیل سهولت نسبی استفاده از MLxtend ما از این کتابخانه برای پاسخ دادن به ادامه سوالات استفاده کردیم.

الف

تعداد Itemsetهای مکرر شناسایی شده با استفاده از الگوریتم Apriori برای هر کدام از حداقل ساپورت‌های ۰.۳، ۰.۰۳ و ۰.۰۰۳ و حداقل طول ۲ در جدول زیر قابل مشاهده است.

Table 11: #Frequent Itemsets

Min_support	#Frequent Itemsets
0.003	1328
0.03	18
0.3	0

همانطور که در جدول فوق مشخص است با افزایش مقدار حداقل ساپورت تعداد Itemsetهای مکرر که از شرط مورد نظر تبعیت کنند کاهش می‌یابد. با مقدار حداقل ساپورت ۰.۳ هیچ Itemset مکرری با حداقل طول ۲ وجود ندارد . در چنین شرایطی دو نتیجه متفاوت می‌توان گرفت، این امکان وجود دارد که Itemset مکرر در این مجموعه داده وجود ندارد یا اینکه مقدار حداقل ساپورت انتخاب شده برای این مجموعه نامناسب است و اطلاعات اشتباه به ما نشان می‌دهد. با بررسی مقادیر دیگر برای حداقل ساپورت می‌توان فهمید که کدامیک از دو حالت مطرح شده صحیح است. با بررسی Itemsetهایی که با حداقل ساپورت ۰.۰۳ متوجه می‌شویم که در این مجموعه داده Itemsetهای مکرری وجود دارد که از نظر معنایی با هم رابطه دارند. برای مثال Itemsetهای {Spaghetti, Ground beef}، {Eggs, Milk} و {Chocolate, Milk} بین ۱۸ Itemset با حداقل ساپورت ۰.۰۳ هستند که دارای رابطه معناداری هستند. با حداقل ساپورت ۰.۰۰۳ نیز تعداد ۱۳۲۸ Itemset به عنوان Itemset مکرر معرفی شده است که تعداد زیادی از آن‌ها دارای روابط معنایی خاصی نیستند. برای مثال روابط معناداری همانند { Eggs, Cereal}، {Spaghetti, Burgers} و {Eggs, Pancakes} علاوه بر موارد ذکر شده برای حداقل ساپورت ۰.۰۳ وجود دارند اما برعکس Itemsetهای بی معنی بیشتری نیز مانند {Mineral water, Bacon} و {Spaghetti, Chocolate, Soup} به عنوان Itemset مکرر معرفی شده‌اند.

ب

به نظر می‌رسد مقدار حداقل ساپورت ۰.۰۳ بهترین مقدار برای پیدا کردن Itemsetهای مکرر این مجموعه داده باشد چرا که حضور محصولاتی که با هم رابطه مثبت دارند در ۲۲۵ تراکنش از ۷۵۰۱ تراکنش عددی منطقی برای نشان دادن رابطه بین محصولات می‌باشد. مقدار حداقل ساپورت ۰.۳ برای این تعداد تراکنش و چنین تعداد Item به نظر می‌رسد که مقدار بالایی باشد چرا که هیچ Itemset وجود ندارد که این شرط برایش برقرار باشد و حتی در صورتی که چند محصول با هم ارتباط مثبت داشته باشند نیز بعید به نظر می‌رسد در ۰.۳ تراکنش‌ها(۲۲۵۰ تراکنش از مجموعه داده سبدهای خرید!) به طور همزمان وجود داشته باشند. از طرفی دیگر حداقل ساپورت ۰.۰۰۳ نیز به شدت پایین می‌باشد(حضور چند محصول در ۲۲ تراکنش از ۷۵۰۱ تراکنش سبدهای خرید!) و حتی امکان حضور چند محصول به صورت تصادفی در این تعداد تراکنش

وجود دارد. همچنین با توجه به بخش الف این سوال، Itemsetهای مکرر معنادارتر بیشتری با مقدار حداقل ساپورت ۰.۰۳ معرفی شده‌اند و از Itemsetهای با مقدار حداقل ساپورت ۰.۰۰۳ بیشتر هستند و با مقدار حداقل ساپورت ۰.۳ نیز هیچ Itemset مکرری پیدا نشده است.

ج

Itemsetهای مکرر با حداقل ساپورت ۰.۰۵ با استفاده از الگوریتم FPGrowth کاوش شدند و نتایج زیر به همراه مقدار ساپورت آن‌ها به دست آمد:

Table 12: Found frequent itemsets by FPgrowth

Itemset	Support	Itemset	Support
Mineral water	0.238	Low fat yoghurt	0.076
Eggs	0.179	Shrimp	0.071
Spaghetti	0.174	Tomatos	0.068
French fries	0.170	Olive oil	0.065
Chocolate	0.163	Frozen smoothie	0.063
Green tea	0.132	Turkey	0.062
Milk	0.129	Chicken	0.059
Ground beef	0.098	Mineral water, Spaghetti	0.059
Frozen vegetables	0.095	Whole wheat rice	0.058
Pancakes	0.095	Mineral Water, Chocolate	0.052
Burgers	0.087	Grated Cheese	0.052
Cake	0.081	Cooking oil	0.051
Cookies	0.080	Mineral water, Eggs	0.050
Escalope	0.079	Soup	0.050

سوال ۴

الف

Association Ruleهای خواسته شده به کمک کتابخانه Mlxtend بدین صورت جمع آوری شدند که ابتدا Itemsetهای مکرر با حداقل ساپورت ۰.۰۳ محاسبه شدند و سپس Association Ruleهایی با Min-confidence برابر ۰.۲ به دست آمدند. تعداد قوانین به دست آمده برابر ۲۷ می‌باشد و سه قانون با بالاترین مقدار معیار لیفت به شرح زیر می‌باشند.

Table 13: Top 3 association rules

Antecedents	Consequents	Lift	Support	Confidence
Spaghetti	Ground beef	2.291	0.039	0.225
Ground beef	Spaghetti	2.291	0.039	0.398
Ground beef	Mineral water	1.747	0.040	0.416

ب

برای به دست آوردن Ruleهای خواسته شده در این بخش سوال از Itemsetهای مکرر بخش قبل استفاده می‌کنیم و Ruleهایی با Min-confidence برابر ۰.۳۵ را به دست می‌آوریم. تعداد قوانین به دست آمده برابر ۵ می‌باشد و سه قانون با بالاترین مقدار معیار لیفت در جدول زیر گزارش شده‌اند.

Table 14: Top 3 association rules

Antecedents	Consequents	Lift	Support	Confidence
Ground beef	Spaghetti	2.291	0.039	0.398
Ground beef	Mineral water	1.747	0.040	0.416

Frozen Vegetables	Mineral water	1.572	0.035	0.374

تعداد قوانین حالت به نسبت به حالت الف ۲۲ کمتر است. علت این کاهش تعداد قوانین افزایش مقدار Min_confidence از ۰.۲ به ۰.۳۵ می‌باش و بدین معناست که در حالت اول هر دو محصول در ۰.۲ تراکنش‌های محصول اول حضور دارند و در حالت دوم هر دو محصول در ۰.۳۵ تراکنش‌های محصول اول حضور دارد. این تغییر باعث حذف محصولاتی می‌شود که وابستگی کمتری نسبت به یکدیگر دارند می‌شود و در نتیجه تعداد قوانین کاهش می‌یابد. تصمیم‌گیری روابط بین محصولات با استفاده از Min-confidence می‌تواند منطقی‌تر باشد چرا که در حالتی که بخواهیم فقط بر اساس حداقل ساپورت تصمیم گیری کنیم امکان گمراه شدن به وسیله محصولاتی که به تنهایی نیز جزو Itemsetهای مکرر هستند وجود دارد. برای مثال در یک مجموعه با ۱۰۰۰ تراکنش که محصول A و B هر کدام در ۲۰۰ تراکنش وجود دارد و در ۳۰ تراکنش هر دو محصول وجود دارد و حداقل ساپورت برابر ۰.۰۳ و Min-confidence برابر ۰.۲ باشد، Itemset {A, B} دارای ساپورت ۰.۰۳ است و شرط حداقل ساپورت را نقض نمی‌کند اما محصول B فقط در ۰.۱۵ تراکنش‌های محصول A حضور دارد(Confidence) و شرط Min-Confidence را نقض می‌کند در نتیجه حداقل ساپورت رابطه بین این محصولات را نشان می‌دهد در صورتی که Min-Confidence نشانگر هیچ رابطه خاصی بین آن‌ها نیست.

تمرین تشریحی امتیازی

الف

با در نظر گرفتن اینکه ترتیب Itemها در زیر مجموعه‌های سه تایی درخت مهم است، برای محاسبه Itemsetهایی با طول ۳ به کمک درخت هش موجود در صورت سوال بدین صورت عمل می‌کنیم:

۱. Itemsetهایی که با ۱ شروع می‌شوند: در برگ L1 به دنبال {۱،۴،۵}، در برگ L5 به دنبال {۱،۳،۴}، {۱،۳،۵} و {۱،۳،۸} و در برگ L3 به دنبال {۱،۵،۸} می‌گردیم.
۲. Itemsetهایی که با ۳ شروع می‌شوند: در برگ L9 به دنبال {۳،۴،۵}و{۳،۴،۸} و در برگ L11 به دنبال {۳،۵،۸} می‌گردیم.
۳. Itemsetهایی که با ۴ شروع می‌شوند: در برگ L3 به دنبال {۴،۵،۸} می‌گردیم.

هیچ Itemsetی وجود ندارد که با Itemهای ۵ و ۸ آغاز شود بنابراین فقط برگ‌های {L1، L3، L5، L9، L11} ملاقات می‌شوند.

ب

با توجه به برگ‌های ملاقات شده Itemsetهای کاندیدهای سه تایی {۱،۴،۵}، {۱،۵،۸}، {۴،۵،۸} وجود دارند. دیگر Itemsetهای سه تایی که در بخش الف و در برگ‌های مربوطه دنبال آن‌ها گشتیم در درخت هش موجود نبودند و بنابراین در Itemsetهای کاندید قرار نمی‌گیرند.