مقدمة

تعني قواعد الارتباط التنقيبية البحث عن الارتباطات ما بين البيانات في قواعد البيانات فمع عمليات التجميع والتخزين المستمرة لكميات هائلة من البيانات بدأت العديد من القطاعات الصناعية تهتم بشكل أكبر بالتنقيب عن قواعد الارتباط في قواعد بياناتها لأن اكتشاف العلاقات الارتباطية ما بين الكميات الكبيرة من سجلات الصفقات التجارية يمكن أن يساعد العديد من القطاعات المصرفية أو الصناعية وغيرها في عملية دعم القرار مثل التسويق والتصميم وغيرها.

ويعتبر تحليل سلة التسوق (market basket analysis) المثال العملي عن قواعد الارتباط التنقيبية وتقوم هذه العملية بتحليل العادات الشرائية لإيجاد الارتباطات بين مختلف المواد التي يشتريها الزبون ويضعها في سلة التسوق الخاصة به وبالتالي يساعد اكتشاف تلك الارتباطات بائعي التجزئة من تطوير استراتيجياتهم التسويقية من خلال تحديد المواد التي تشترى عادة معاً من قبل الزبائن.

٥-١-قواعد الارتباط التنقيبية (Association Rule Mining)

تبحث قواعد الارتباط التنقيبية عن الارتباطات المهمة ما بين العناصر (المواد) في مجموعة البيانات. يزود هذا القسم بمقدمة عن قواعد الارتباط التنقيبية فنبدأ بتوضيح مثال عن تحليل سلة التسوق التي تعتبر الشكل الأبسط لقواعد الارتباط التنقيبية ومن ثم نعطي بعض المفاهيم الأساسية والأنواع المختلفة لقواعد الارتباط الممكن التنقيب عنها.

٥-١-١-تحليل سلة التسوق: مثال عن قواعد الارتباط التنقيبية

بفرض أنك مدير لإحدى الفروع الكبرى الخاصة بمؤسستك وترغب بالتعلم أكثر عن العادات الشرائية لزبائنك من خلال الإجابة عن السؤال التالي: ما هي المجموعات أو المواد المحتمل أن يشتريها الزبائن خلال جولتهم في هذا الفرع؟ للإجابة عن هذا السؤال يتم تحليل سلة التسوق على البيانات الخاصة بصفقات الزبائن وتستخدم نتائج هذا التحليل لتخطيط التسوق أو استراتيجيات الإعلانات التجارية وتصميم قائمة المنتجات. ربما يساعد تحليل سلة التسوق مثلاً المدراء في ترتيب المتجر من خلال توضع المواد التي تشترى عادة معاً في أماكن قريبة من بعضها البعض وذلك لتشجيع الزبائن على شرائها معاً ففي حالة كان الزبائن الذين يشترون المعدات الحاسوبية يميلون أيضاً لشراء

برمجيات الإدارة المالية بنفس الوقت عندئذ يعرض العتاد الصلب (hardware) بجانب البرمجيات مما يساعد في زيادة مبيعاتهما معاً أو كإستراتيجية بديلة يتم وضع العتاد الصلب والبرمجيات مقابل بعضهما البعض في المتجر فيجذب الزبائن على شرائهما معاً. إذا اعتبرنا وجود مجموعة من المواد المتوفرة في المتجر عندئذ فإن كل مادة تمتلك متحول منطقي يمثل وجود أو غياب تلك المادة وعندئذ يمكن تمثيل كل سلة بواسطة شعاع منطقي من القيم المسندة لتلك المتحولات وبالتالي يمكن تحليل الأشعة المنطقية للعينات المشتراة والتي تعكس المواد المشتراة معاً وتمثل تلك العينات بشكل قواعد ارتباط (association rules). كمثال المعلومات التي تبين بأن الزبائن الذين يشترون الحواسيب يميلون إلى شراء برمجيات الإدارة المالية بنفس الوقت يمكننا تمثيلها بقاعدة ارتباط من الشكل:

Computer => financial_management_software [support=2%, confidence= 60%]

قاعدة الدعم (support) والموثوقية (confidence) قياسان لدرجة أهمية قاعدة الارتباط والتي تحسب بالعلاقات التالية:

$$\text{Confidence (A => B)} = \frac{number\_tuples\_containing\_both\_A\_and\_B}{number\_tuples\_containing\_A}$$

$$\text{Support (A => B)} = \frac{number\_tuples\_containing\_both\_A\_and\_B}{total\_number\_of\_tuples}$$

وتعكسان درجة فائدة وتأكيد القواعد المكتشفة وتعني 2% لقاعدة الدعم بأن 2% من كل الصفقات قيد التحليل تبين بأن الحواسيب وبرمجيات الإدارة المالية يتم شراؤهما معاً وتعني قاعدة الموثوقية 60% بأن 60% من الزبائن الذين يشترون الحواسيب يشترون معها البرمجيات الحاسوبية.

٥-١-٢-مفاهيم أساسية

بفرض I={i1,i2,…..,im} هي مجموعة المواد و D هي قاعدة بيانات الصفقات الخاصة بتلك المواد وبفرض أن كل صفقة T منها تحتوي مجموعة من المواد وهذا يعني أن T محتواه أو تساوي I. تمتلك كل صفقة معرّف خاص بها هو T_{ID} وبفرض أن A هي مجموعة مواد فنقول عن الصفقة T أنها تحتوي A بحالة وبحالة فقط كانت A محتواه أو تساوي T. تصاغ قاعدة الارتباط بالشكل A => B حيث أن A محتواه

في I وB محتواه في I و$A \cap B = \emptyset$ حيث أن **الدعم (s)** يمثل النسبة المئوية من الصفقات في D المحتوية A وB معاً (AUB) ويعبر عنه بالاحتمال P(AUB) **والموثوقية (c)** في مجموعة الصفقات D هي النسبة المئوية من الصفقات في D المحتوية A والتي تحتوي B أيضاً ويعبر عنها بالاحتمال الشرطي P(B|A) وهذا يعني:

Support (A => B) = P(AUB) Confidence (A => B) = P(B|A)

تدعى القواعد التي تحقق كل من عتبة الدعم الأصغرية (min_sup) وعتبة الموثوقية الأصغرية (min_conf) بأنها قواعد قوية (strong) حيث يتم وضع قيم الدعم والموثوقية لتقع بين 0% و100% بدلاً من 0 و1. تتألف مجموعة البنود (المواد) (itemset) من K مادة ويعبر عنها بـ k-itemset فمثلاً المجموعة {computer, financial_management_software} هي 2-itemset ويعبر عن عدد الصفقات التي تحتوي مجموعة المواد بتكرار حدوث مجموعة المواد ويدعى بالتكرار (frequency) أو مقدار الدعم (support count) للمواد وتحقق مجموعة المواد الدعم الأصغري بحالة كان تكرار حدوثها أكبر من أو يساوي لحاصل جداء min_sup والعدد الأعظمي للصفقات في D ويشير إليه بمقدار الدعم الأصغري (minimum support count) وبحالة كانت مجموعة المواد تحقق الدعم الأصغري فعندئذ هي مجموعة مواد متكررة.

تتألف عماية التنقيب عن قواعد الارتباط التنقيبية من مرحلتين:

١. إيجاد كل مجموعات المواد المتكررة حيث أن كل مجموعة من مجموعات المواد سوف تتكرر على الأقل بمقدار الدعم الأصغري المحدد مسبقاً.
٢. توليد قواعد الارتباط القوية من مجموعة المواد المتكررة بمعنى أنها تحقق شرط الدعم الأصغري والموثوقية الأصغرية.

٥-٢-قواعد الارتباط ذات البعد الواحد

سوف نتعلم في هذا الفصل طرق (مناهج) للتنقيب عن أبسط شكل من قواعد الارتباط ذات البعد الواحد وأهمها خوارزمية Apriori وتعتبر الخوارزمية الأساسية لإيجاد مجموعة المواد المتكررة وسوف نتطرق إلى طرق استخراج القواعد القوية منها.

٥-٢-١-خوارزمية Apriori

تستخدم هذه الخوارزمية لإيجاد مجموعة المواد المتكررة باستخدام توليد المرشح (candidate generation) وتعتمد هذه الخوارزمية على المعرفة المسبقة بخواص المواد المتكررة وتقوم بتوظيف طريقة تكرارية تعرف بالبحث level-wise حيث تستخدم K مجموعة من المواد لتوليد K+1 مجموعة من المواد. يتم في البداية إيجاد مجموعة مواد متكررة ذات طول يساوي 1 ويرمز لها بـ L1 وتستخدم لإيجاد L2 والتي تعبر عن مجموعة المواد المتكررة ذات طول يساوي 2 والتي تستخدم بدورها لتوليد L3 وهكذا حتى لا يمكننا تشكيل أي k-itemset متكررة ويتطلب الإيجاد لكل Lk مرور واحد على كامل قاعدة البيانات.

لتحسين فعالية توليد level-wise للمواد المتكررة فإنه تستخدم خاصية مهمة تدعى بخاصية الـ Apriori وتستخدم لتقليل فضاء البحث وسنوضح هذه الخاصية في البداية ومن ثم سنعطي مثالاً توضيحياً عنها.

خاصية Apriori: يجب أن تكون كل المجموعات الجزئية غير الفارغة من مجموعة البنود (المواد) المتكررة متكررة أيضاً وتعتمد هذه الخاصية على الملاحظة التالية: إذا كانت مجموعة مواد I لا تحقق عتبة الدعم الأصغرية عندئذ فإن I ليست متكررة هذا يعني P(I)<min_sup وبحالة كانت المادة A قد أضيفت إلى مجموعة المواد I عندئذ فإن مجموعة المواد الناتجة (أي AUI) لا يمكن أن تتكرر أكثر من تكرار I بالتالي AUI ليست مكررة أيضاً وهذا يعني أن P(AUI)<min_sup.

لفهم هذه الخاصية دعنا ننظر إلى كيفية استخدام L(k-1) لتوليد L(k) وتتألف هذه العملية ذات المرحلتين من خطوتي الربط (join) والتشذيب (prune):

١. *خطوة الربط:* يتم إيجاد L(k) والتي هي عبارة عن K مجموعة مواد مرشحة بربط L(k-1) مع بعضها البعض ويرمز لها بـ C(k). لتكن l1 و l2 مجموعتي بنود في L(k-1) ويشير التدوين Li[j] إلى البند ذو الرقم j في li (مثال تشير l1[k-2] إلى العنصر الثاني لآخر بند في l1) بحسب التعريف تفترض Apriori بأن المواد ضمن الصفقة أو مجموعة البنود تخزن بترتيب أبجدي

عندئذ يطبق الربط L(k-1)L(k-1) حيث أن عناصر L(k-1) قابلة للربط مع بعضها البعض بحالة

(l1[1]=l2[1]^l1[2]=l2[2]^……..^l1[k-1]<l2[k-1]) ويشير التدوين l1[k-1]<l2[k-1] ببساطة إلى عدم توليد تكرارات.

٢. *خطوة التشذيب:* C(k) هي المجموعة العليا لـ L(k) وهذا يعني ربما تكون أعضائها متكررة أو لا ولكن كل K مجموعة من المواد المتكررة متضمنة في C(k). يتم فحص قاعدة البيانات لتحديد عدد كل مرشح في C(k) وبالتالي تحديد L(k) (هذا يعني كل العناصر المرشحة التي تمتلك عدد ليس أقل من الدعم الأصغري). يمكن أن تكون C(k) كبيرة جداً ولتقليل حجمها تستخدم خاصية Apriori بالشكل التالي: أي (k-1)-itemset ليست متكررة لا يمكن أن تكون مجموعة جزئية من k-itemset المتكررة بالنتيجة إذا كانت أية (k-1) مجموعة جزئية من k-itemset المرشحة ليست من L(k-1) عندئذ فإن المرشح لا يمكن أن يكون متكرر أيضاً وبالتالي تزال من C(k).

مثال ٥-١:

لنأخذ مثال عن خوارزمية الـ Apriori بالاعتماد على الصفقات D المبينة في الجدول التالي. توجد تسعة صفقات في قاعدة البيانات، هذا يعني |D|=9.

T_{ID}	List of item_ID_s
T_{100}	I_1, I_2, I_5
T_{200}	I_2, I_4
T_{300}	I_2, I_3
T_{400}	I_1, I_2, I_4
T_{500}	I_1, I_3
T_{600}	I_2, I_3
T_{700}	I_1, I_3
T_{800}	I_1, I_2, I_3, I_5
T_{900}	I_1, I_2, I_3

توضيح عمل الخوارزمية:

١. في أول تكرار للخوارزميـة فإن كل بند هو عنصر في مجموعة من 1-itemset المرشحة C_1. تفحص الخوارزمية ببساطة كل الصفقات لتعد عدد مرات حدوث كل بند.

٢. بفرض أن عتبة الدعم الصغرى هي 2 (هذا يعني $sup = {}^{2}/_{9} = 22\%$).

بالتالي يمكن تحديد مجموعة من 1-itemset المتكررة والتي يرمز لها بـ L_1 وتتألف من 1-itemset المرشحة والتي تحقق شرط الدعم الأصغري.

٣. لاكتشاف مجموعة من 2-itemset المتكررة، L_2، تستخدم الخوارزمية $L_1 \bowtie L_1$ لتوليد المجموعة 2-itemset المرشحة، C_2.

٤. ثم يتم فحص الصفقات في D لتحديد عدد الدعم لكل من العناصر المرشحة في C_2.

٥. يتم تحديد مجموعة من 2-itemset المتكررة L_2 وتتألف من 2-itemset المرشحة في C_2 والتي تحقق شرط العتبة الأصغري.

٦. يتم توليد مجموعة من 3-itemset المرشحة C_3 حيث أن:

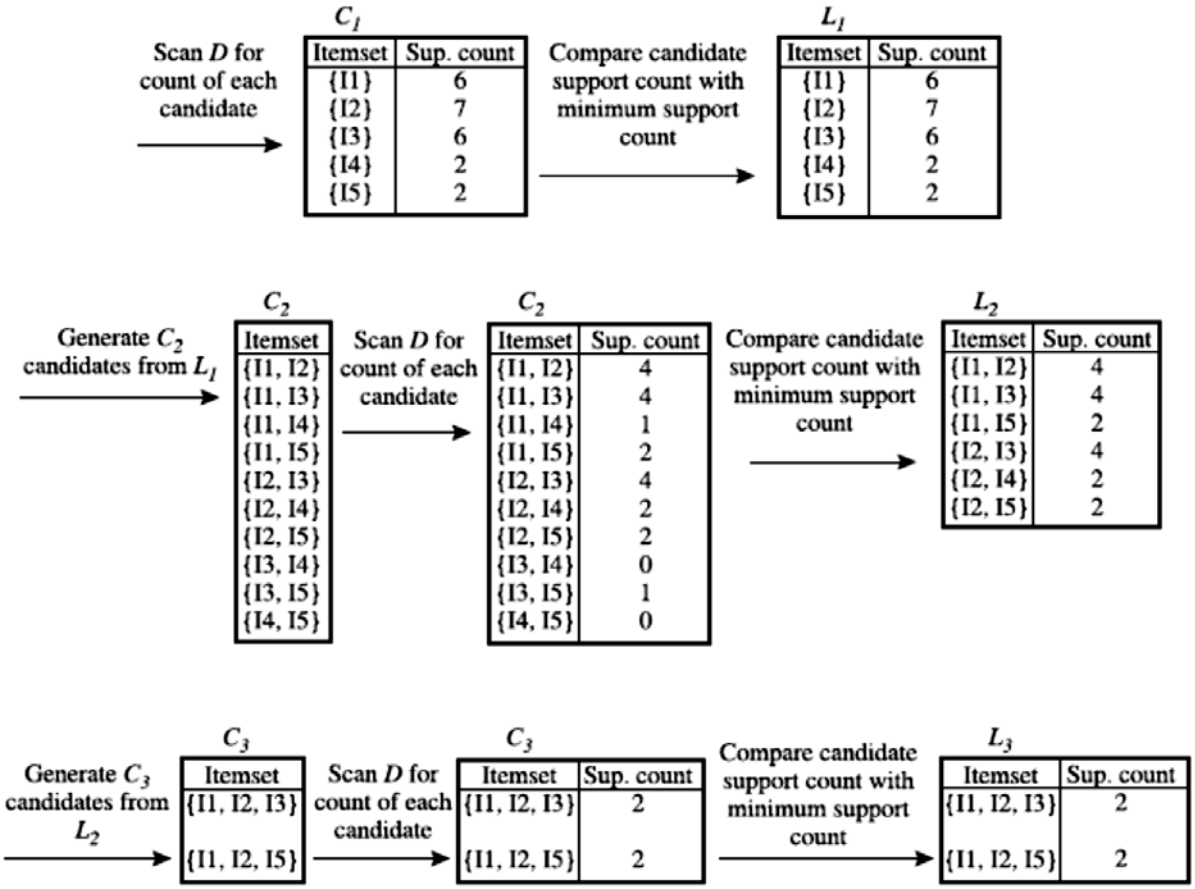
$$C_3 = L_2 \bowtie L_2$$
$$= \{\{I_1, I_2, I_3\}, \{I_1, I_2, I_5\}, \{I_1, I_3, I_5\}, \{I_2, I_3, I_4\}, \{I_2, I_3, I_5\}, \{I_2, I_4, I_5\}\}$$

وبالاعتماد على خاصية Apriori فإن كل المجموعات الجزئية من مجموعة البنود المتكررة يجب أن تكون متكررة أيضاً.

٧. تُفحص الصفقات في D لتحديد L_3 والتي تحتوي 3-itemset المرشحة في C_3 والتي تمتلك الدعم الأصغري.

٨. تستخدم الخوارزمية $L_3 \bowtie L_3$ لتوليد مجموعة 4-itemset المرشحة C_4 وبالرغم من أن نتائج الربط هي $\{\{I_1, I_2, I_3, I_5\}\}$ فإنه يتم إزالته بسبب أن المجموعة الجزئية $\{I_1, I_3, I_5\}$ ليست مكررة وبالتالي $C_4 = \emptyset$ وتنتهي الخوارزمية (تتوقف) بإيجاد كل مجموعات العناصر المكررة.

يوضح الشكل (1-5) توليد مجموعة العناصر المرشحة والمكررة حيث أن شرط عتبة الدعم الأصغري هي 2.



الشكل 1-5: توليد مجموعة العناصر المرشحة والمتكررة بطريقة Apriori

يبين الشكل (5-2) توليد 3-itemset من L_2 باستخدام خاصية Apriori.

(a) Join: $C_3 = L_2 \bowtie L_2 = \{\{I1, I2\}, \{I1, I3\}, \{I1, I5\}, \{I2, I3\}, \{I2, I4\}, \{I2, I5\}\} \bowtie$
$\{\{I1, I2\}, \{I1, I3\}, \{I1, I5\}, \{I2, I3\}, \{I2, I4\}, \{I2, I5\}\}$
$= \{\{I1, I2, I3\}, \{I1, I2, I5\}, \{I1, I3, I5\}, \{I2, I3, I4\}, \{I2, I3, I5\}, \{I2, I4, I5\}\}.$

(b) Prune using the Apriori property: All nonempty subsets of a frequent itemset must also be frequent. Do any of the candidates have a subset that is not frequent?

- The 2-item subsets of {I1, I2, I3} are {I1, I2}, {I1, I3}, and {I2, I3}. All 2-item subsets of {I1, I2, I3} are members of L_2. Therefore, keep {I1, I2, I3} in C_3.
- The 2-item subsets of {I1, I2, I5} are {I1, I2}, {I1, I5}, and {I2, I5}. All 2-item subsets of {I1, I2, I5} are members of L_2. Therefore, keep {I1, I2, I5} in C_3.
- The 2-item subsets of {I1, I3, I5} are {I1, I3}, {I1, I5}, and {I3, I5}. {I3, I5} is not a member of L_2, and so it is not frequent. Therefore, remove {I1, I3, I5} from C_3.
- The 2-item subsets of {I2, I3, I4} are {I2, I3}, {I2, I4}, and {I3, I4}. {I3, I4} is not a member of L_2, and so it is not frequent. Therefore, remove {I2, I3, I4} from C_3.
- The 2-item subsets of {I2, I3, I5} are {I2, I3}, {I2, I5}, and {I3, I5}. {I3, I5} is not a member of L_2, and so it is not frequent. Therefore, remove {I2, I3, I5} from C_3.
- The 2-item subsets of {I2, I4, I5} are {I2, I4}, {I2, I5}, and {I4, I5}. {I4, I5} is not a member of L_2, and so it is not frequent. Therefore, remove {I2, I4, I5} from C_3.

(c) Therefore, $C_3 = \{\{I1, I2, I3\}, \{I1, I2, I5\}\}$ after pruning.

الشكل 2-5: طريق توليد 3-itemset

٥-٢-٢-توليد قواعد الارتباط من مجموعة العناصر المتكررة

بعد إيجاد مجموعة العناصر المتكررة في قاعدة البيانات D يتم مباشرة توليد قواعد الارتباط القوية منها (قواعد الارتباط القوية هي القواعد التي تحقق شرطي عتبة الدعم الأصغري والموثوقية الأصغرية). هذا يمكن تطبيقه باستخدام المعادلات التالية للموثوقية والتي تعبر عن الاحتمال الشرطي مفسراً بعدد الدعم لمجموعة العناصر:

$$confidence(A => B) = P(B|A) = \frac{support\_count(A \cup B)}{support\_count(A)}$$

حيث أن $support\_count(A \cup B)$ هو عدد الصفقات المحتوية A وB و$support_count(A)$ هو عدد الصفقات المحتوية A وبالاعتماد على هذه المعادلة يمكن توليد قواعد الارتباط بالشكل التالي:

- من أجل كل مجموعة عناصر متكررة ℓ، يتم توليد كل المجموعات الجزئية غير الفارغة من ℓ.

- من أجل كل مجموعة جزئية غير فارغة من ℓ نقوم بفحص كل قاعدة محتملة من الشكل $s => (\ell - s)$ و في حالة كانت نتيجة الفحص $\frac{support\_count(\ell)}{support\_count(s)} \geq \min\_conf$ حيث أن $\min\_conf$ هو عتبة الموثوقية الأصغرية تكون تلك القاعدة قاعدة ارتباط تنقيبية مكتشفة.

مثال 2-5: من أجل المثال 1-5 حصلنا على مجموعة العناصر المتكررة التالية: $\ell = \{I_1, I_2, I_5\}$ فيمكن توليد قواعد الارتباط منها بالشكل التالي: المجموعة الجزئية غير الفارغة من ℓ هي: $\{I_1, I_2\}, \{I_1, I_5\}, \{I_2, I_5\}, \{I_1\}, \{I_2\}, \{I_5\}$ لذلك فإن قواعد الارتباط هي:

$I1 \wedge I2 \Rightarrow I5$, $confidence = 2/4 = 50\%$
$I1 \wedge I5 \Rightarrow I2$, $confidence = 2/2 = 100\%$
$I2 \wedge I5 \Rightarrow I1$, $confidence = 2/2 = 100\%$
$I1 \Rightarrow I2 \wedge I5$, $confidence = 2/6 = 33\%$
$I2 \Rightarrow I1 \wedge I5$, $confidence = 2/7 = 29\%$
$I5 \Rightarrow I1 \wedge I2$, $confidence = 2/2 = 100\%$

وبفرض أن عتبة الموثوقية الأصغرية هي 70% عندئذ فقط القاعدة الثانية والثالثة والأخيرة هي القواعد القوية.

٥-٢-٣-التنقيب عن مجموعات العناصر المتكررة بدون توليد المرشح

لقد رأينا بأن الطريقة السابقة تقلل من حجم مجموعة العناصر المرشحة بشكل ملحوظ فتقود إلى أداء جيد في النهاية ولكنها تعاني من عدة سيئات:

- أنها تحتاج إلى توليد عدد كبير من مجموعات العناصر المرشحة فمثلاً بحالة وجود مجموعة مؤلفة من 10^4 من 1-itemset المتكررة فإن الخوارزمية ستحتاج إلى توليد أكثر من 10^7 من العناصر المرشحة المؤلفة من 2-itemset.
- كما أنها قد تحتاج إلى إعادة فحص لقاعدة البيانات عدة مرات ومقارنة مجموعة كبيرة من العناصر المرشحة بواسطة مطابقة العينات.

لذلك هل يمكننا تصميم طريقة تقوم بالتنقيب عن مجموعة كاملة من العناصر المتكررة بدون توليد العناصر المرشحة؟ إحدى أهم هذه الطرق تُدعى بـ Frequent-Pattern

Growth أو ببساطة FP-growth وتعتمد طريقة فرق تسد بالشكل التالي: ضغط قاعدة البيانات الممثلة للعناصر المتكررة إلى شجرة العينات المتكررة (FP-tree) ولكن يتم الاحتفاظ بمجموعة العناصر المرتبطة بالمعلومات ومن ثم تقسيم قاعدة البيانات المضغوطة إلى مجموعة من قواعد البيانات الشرطية وكل منها مرتبط بعنصر متكرر واحد ومن ثم التنقيب في كل قاعدة بيانات على حدى. دعنا نأخذ مثال لتوضيح ذلك.

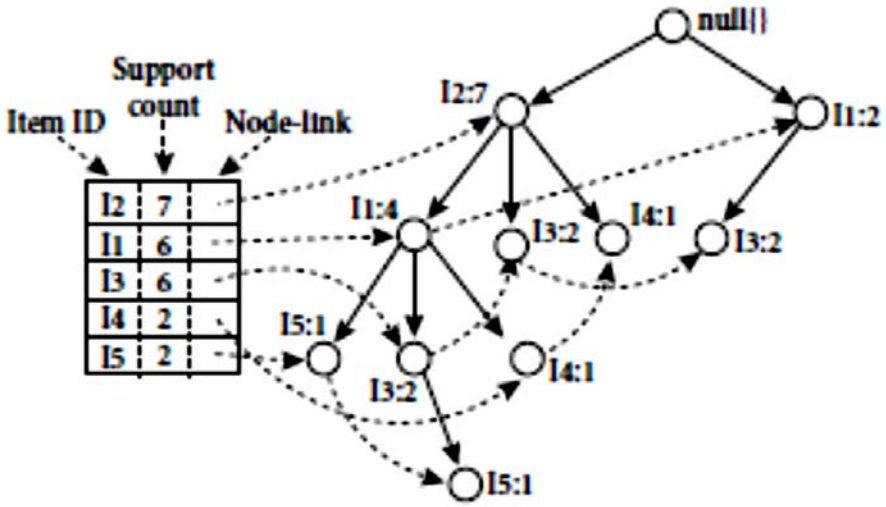
مثال 5-3: سوف نستخدم نفس قاعدة البيانات D من المثال 5-1 وذلك باستخدام طريقة FP-growth.

الفحص الأول لقاعدة البيانات هو نفس طريقة Apriori حيث يتم تشكيل مجموعة العناصر المتكررة المؤلفة من عنصر واحد فقط وعدد مرات تكرارها وإذا فرضنا أن عتبة الدعم الأصغرية 2 فإنه يتم تخزين مجموعة العناصر المتكررة بترتيب تنازلي وبرمز للمجموعة الناتجة بـ L فيكون لدينا: $L = \{I_2: 7, I_1: 6, I_3: 6, I_4: 2, I_5: 2\}$.

يتم بعد ذلك بناء شجرة FP بالشكل التالي: بداية يتم إنشاء جذر الشجرة وتعطى الاسم (null). يتم فحص قاعدة البيانات D مرة ثانية ويتم ترتيب عناصر الصفقة بحسب ترتيب L (هذا يعني ترتيبها تنازلياً) ويتم إنشاء (رسم) فرع لكل صفقة. مثلاً بفحص أول صفقة $(T_{100}: I_1, I_2, I_5)$ والتي تحتوي ثلاثة عناصر وهي (I_2, I_1, I_5) بحسب ترتيب L تؤدي إلى بناء (رسم) الفرع الأول للشجرة مع ثلاثة عقد: $(I_2: 1, I_1: 1, I_5: 1)$ حيث تعتبر I_2 كعقدة ابن لعقدة الجذر و I_1 عقدة ابن للعقدة I_2 و I_5 كعقدة ابن للعقدة I_1. الصفقة الثانية T_{200} تحتوي العناصر I_2 و I_4 بترتيب L والتي سوف تشكل الفرع الثاني مع ملاحظة أن I_2 مرتبطة بالعقدة الجذر و I_4 مرتبطة بالعقدة I_2. على كل حال سيتشارك هذا الفرع باللاحقة المشتركة I_2 مع المسار الموجود لـ T_{100}. لذلك يتم زيادة العدد للعقدة I_2 بمقدار 1 وإنشاء عقدة جديدة $I_4: 1$ التي تكون مرتبطة بالعقدة الابن $I_2: 2$.

T_{ID}	List of item_ID_s
T_{100}	I_1, I_2, I_5
T_{200}	I_2, I_4
T_{300}	I_2, I_3
T_{400}	I_1, I_2, I_4
T_{500}	I_1, I_3
T_{600}	I_2, I_3
T_{700}	I_1, I_3
T_{800}	I_1, I_2, I_3, I_5
T_{900}	I_1, I_2, I_3

لتسهيل تركيب الشجرة يتم بناء جدول الرأس (*header table*) حيث يشير كل عنصر إلى عدد مرات حدوثه في الشجرة بواسطة سلسلة من ارتباطات العقد. الشجرة التي يتم الحصول عليها بعد فحص كل الصفقات مبينة في الشكل (*5-3*) مع ارتباطات العقد الموافقة لكل عنصر. لذلك فإن المشكلة في التنقيب عن العينات المتكررة قي قواعد البيانات تحول إلى التنقيب في الشجرة *FP*.



الشكل 3-5: بناء شجرة FP

ينفذ التنقيب في شجرة FP بالشكل التالي: البداية من كل عينة متكررة ذات طول 1 وبناء قاعدة العينات الشرطية الخاصة بها (والتي تتألف من مجموعة من المسارات السابقة في شجرة FP) ومن ثم بناء شجرة FP الشرطية وتنفيذ التنقيب بطريقة عودية على تلك الشجرة ويلخص الجدول (1-5) التنقيب في شجرة FP.

item	Conditional pattern base	Conditional FP-tree	Frequent patterns generated
I_5	$\{(I_2\ I_1{:}1),(I_2I_1I_3{:}1)\}$	$\{I_2{:}2, I_1{:}2\}$	$I_2I_5{:}2, I_1I_5{:}2, I_2I_1I_5{:}2$
I_4	$\{(I_2I_1{:}1),(I_2{:}1)\}$	$\{I_2{:}2\}$	$I_2I_4{:}2$
I_3	$\{(I_2I_1{:}2),(I_2{:}2),(I_1{:}2)\}$	$\{I_2{:}4, I_1{:}2\},\{I_1{:}2\}$	$I_2I_3{:}4, I_1I_3{:}4, I_2I_1I_3{:}2$
I_1	$\{(I_2{:}4)\}$	$\{I_2{:}4\}$	$I_2I_1{:}4$

الجدول (٥-١) -التنقيب في شجرة FP

يتم التنقيب في الشجرة FP بالشكل التالي: بداية نبدأ من آخر عنصر في L وهو I_5 والذي يظهر مرتين في الشجرة كما هو موضح في الشكل (5-3) من خلال المسارين $(I_2I_1I_5{:}1)$ و$(I_2I_1I_3I_5{:}1)$ لذلك تعتبر I_5 كلاحقة أي يبقى لدينا $(I_2I_1{:}1)$ و$(I_2I_1I_3{:}1)$ والتي تشكل قاعدة العينات المشروطة لـ I_5 وتحتوي شجرة FP الشرطية مسار وحيد فقط وهو $\{I_2{:}2, I_1{:}2\}$ وتحذف I_3 لأنها لا تحقق شرط عتبة الدعم الأصغري ويولد هذا المسار كل التراكيب للعينات المتكررة:

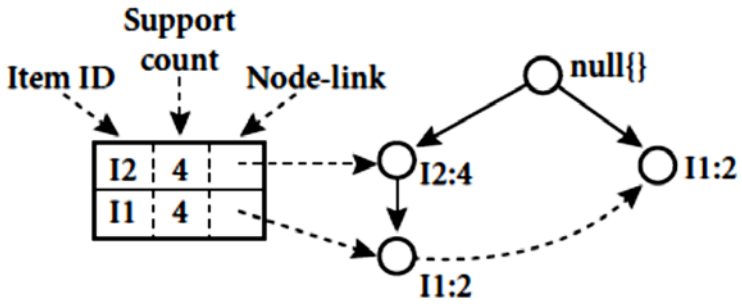
$$I_2I_5{:}2, I_1I_5{:}2, I_2I_1I_5{:}2$$

من أجل I_4 يشكل المساران السابقان قاعدة العينات الشرطية: $\{(I_2I_1{:}1),(I_2{:}1)\}$ والتي تولد بدورها شجرة FP شرطية من عقدة وحيدة وهي $\{I_2{:}2\}$ وبالتالي العينات المتكررة منها هي $I_2I_4{:}2$. لاحظ بأن I_5 تلي I_4 في الفرع الأول ولكن لا توجد ضرورة لتحليلها هنا بسبب أن أي عينة متكررة تحتوي I_5 قد تم تحليلها في اختيار I_5 ولذلك تبدأ المعالجة دوماً من نهاية L بدلاً من بدايتها.

بطريقة مشابهة قاعدة العينات الشرطية لـ I_3 هي $\{(I_2I_1{:}2),(I_2{:}2),(I_1{:}2)\}$ وشجرة FP الشرطية لها تحتوي فرعان هما $\{I_2{:}4, I_1{:}2\},\{I_1{:}2\}$ كما هو موضح في الشكل (5-4) حيث يتم توليد مجموعة العينات التالية:

$$I_2I_3{:}4, I_1I_3{:}4, I_2I_1I_3{:}2$$

وأخيراً بالنسبة لـ I_1 قاعدة العينات المتكررة هي $I_2{:}4$ وشجرة FP الشرطية هي $I_2{:}4$ والعينات المولدة منها $I_2I_1{:}4$.



الشكل 4-5: شجرة FP الشرطية الخاصة بالعقدة I_3

٥-٣-الخلاصة

يعتبر اكتشاف القواعد الارتباطية ما بين الكميات الكبيرة من البيانات مفيداً في التسويق وتحليل القرار وإدارة الأعمال والتطبيق المعروف لها هو تحليل سلة التسويق والتي تدرس العادات الشرائية للزبائن من خلال البحث في مجموعة العناصر المشتراه عادة معاً ويتألف التنقيب عن قاعدة الارتباط بداية من إيجاد مجموعة المواد المتكررة والتي تحقق شرط عتبة الدعم الأصغري من قواعد الارتباط القوية كما تحقق تلك القواعد شرط الموثوقية الأصغري.

خوارزمية Apriori هي خوارزمية فعّالة للتنقيب عن قواعد الارتباط بالاعتماد على خاصية Apriori: يجب أن تكون كل المجموعات الجزئية غير الفارغة لمجموعة العناصر المتكررة متكررة أيضاً والتكرار ذو الرقم k يشكل مجموعة k+1 من العناصر المرشحة المتكررة اعتماداً على k مجموعة عناصر متكررة ويتم فحص قاعدة البيانات مرة واحدة لإيجاد مجموعة كاملة من k+1 مجموعة من العناصر.

FP-growth أو Frequent Pattern growth هو طريقة للتنقيب عن مجموعة العناصر المتكررة بدون توليد المرشح حيث تقوم ببناء تركيب FP-tree لضغط قاعدة بيانات الصفقات الأساسية فبدلاً من توظيف استراتيجية التوليد والاختبار لطريقة Apriori فإنها تركز على العينات المتكررة والتي تتجنب توليد المرشح مما يؤدي غلى فعّالية أكبر.