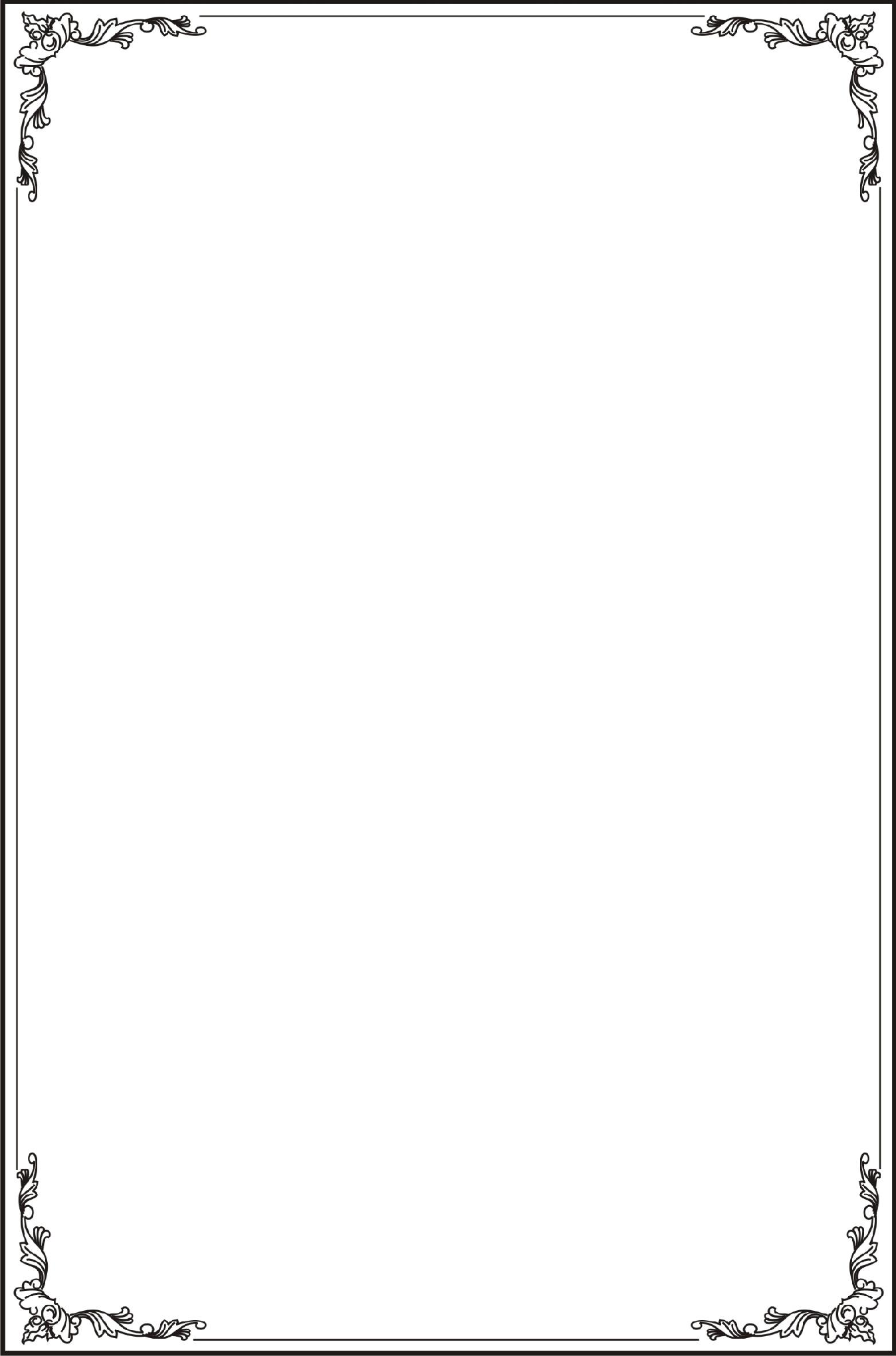
ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH



**TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN**

KHAI THÁC DỮ LIỆU

**BÀI TẬP LAP 2**

**NGUYỄN THỊ NGỌC HÀ: 17520421**

MỤC LỤC

[BÀI TẬP 1: CỬA HÀNG TẠP HÓA 3](#_Toc37744328)

[1. Tìm tập phổ biến bằng giải thuật Apriori: 3](#_Toc37744329)

[2. Tìm tập phổ biến bằng giải thuật FP-growth: 4](#_Toc37744330)

[BÀI TẬP 2: CỬA HÀNG VĂN PHÒNG PHẨM 11](#_Toc37744331)

[1. Tìm tập phổ biến bằng giải thuật Apriori: 11](#_Toc37744332)

[2. Tìm tập phổ biến bằng giải thuật FP\_growth: 12](#_Toc37744333)

[BÀI TẬP 3: CSDL VỀ NHÂN VIÊN: 22](#_Toc37744334)

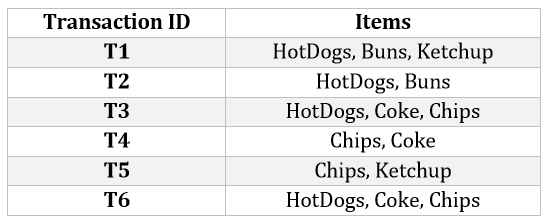
[BÀI TẬP 4: DỮ LIỆU CÔNG TY: 24](#_Toc37744335)

[1. Tìm tập phổ biến bằng giải thuật Apriori: 24](#_Toc37744336)

[2. Tìm tập phổ biến bằng giải thuật FP-growth: 25](#_Toc37744337)

## BÀI TẬP 1: CỬA HÀNG TẠP HÓA

Cho bảng dữ liệu ở một của hàng tạp hóa có 6 giao dịch như sau:



Với min\_sup = 33.33% và min\_conf = 60%

### Tìm tập phổ biến bằng giải thuật Apriori:

* Với min\_sup là 33.33% và tổng số giao dịch là 6 => Tần số xuất hiện tối thiểu của phần tử để thỏa min\_sup (min\_support\_count) là 2
* Tập các ứng viên 1 phần tử và tần số xuất hiện phần tử (support count) tương ứng:

C1 = { HotDogs: 4, Buns: 2, Ketchup: 2, Coke: 3, Chips: 4 }

* Các tập phổ biến 1 phần tử (Tập các ứng viên thỏa mãn min\_support\_count):

L1 = { {HotDogs}, {Buns}, {Ketchup}, {Coke}, {Chips} }

* Tập các ứng viên có 2 phần tử và support count tương ứng:

C2 = { {HotDogs, Buns}: 2, {HotDogs, Ketchup}: 1, {HotDogs, Coke}: 2, {HotDogs, Chips}: 2, {Buns, Ketchup}: 1, {Ketchup, Chips}: 1, {Coke, Chips}: 3 }

* Các tập phổ biến 2 phần tử thỏa min\_support\_count:

L2 = { {HotDogs, Buns}, {HotDogs, Coke}, {HotDogs, Chips}, {Coke, Chips} }

* Tập các ứng viên có 3 phần tử và support count tương ứng:

C3 = { {HotDogs, Coke, Chips}: 2 }

* Tập phổ biến 3 phần tử thỏa min\_support\_count:

L3 = { { HotDogs, Coke, Chips} }

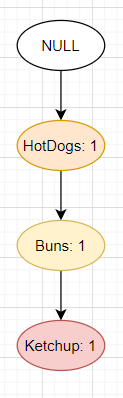
### Tìm tập phổ biến bằng giải thuật FP-growth:

* Tập phổ biến 1 phần tử và tần số xuất hiện (support count) tương ứng:
* HotDogs: 4
* Buns: 2
* Ketchup: 2
* Coke: 3
* Chips: 4
* F-List ( Danh sách giảm dần theo độ hỗ trợ (Support):

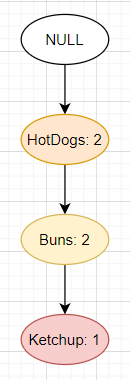
|  |  |
| --- | --- |
| ITEMS | SUPPORT COUNT |
| HotDogs | 4 |
| Chips | 4 |
| Coke | 3 |
| Buns | 2 |
| Ketchup | 2 |

* Xây cây:

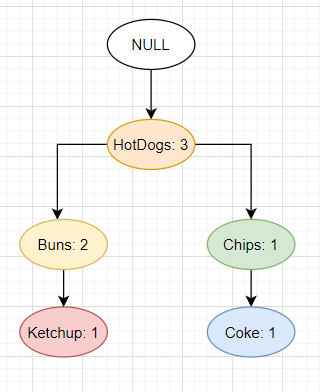
Thêm giao dịch 1 vào cây FP



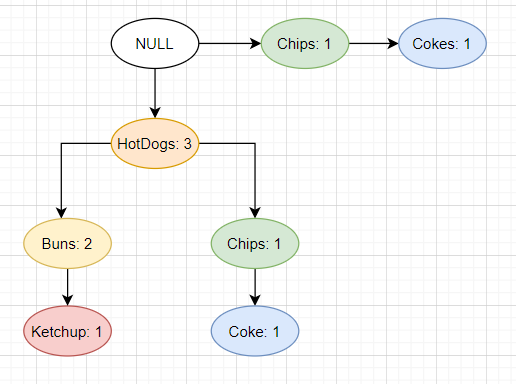
Thêm giao dịch 2 vào cây FP



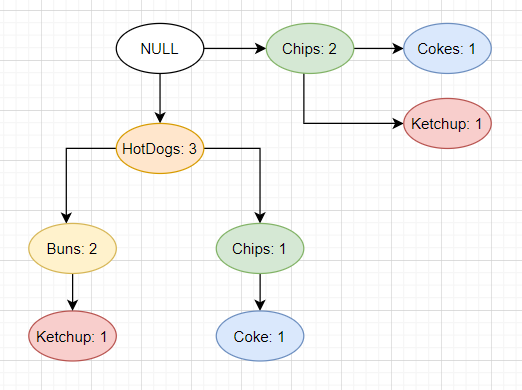
Thêm giao dịch 3 vào cây FP



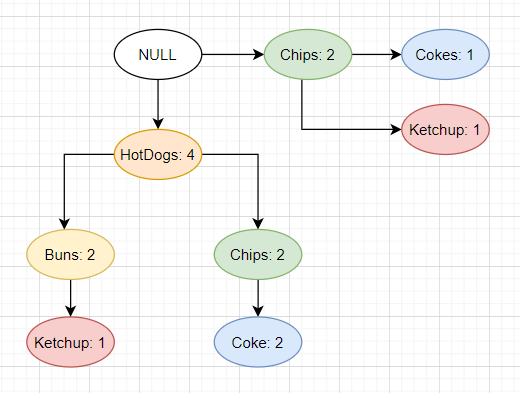
Thêm giao dịch 4 vào cây FP



Thêm giao dịch 5 vào cây FP



Thêm giao dịch 6 vào cây FP

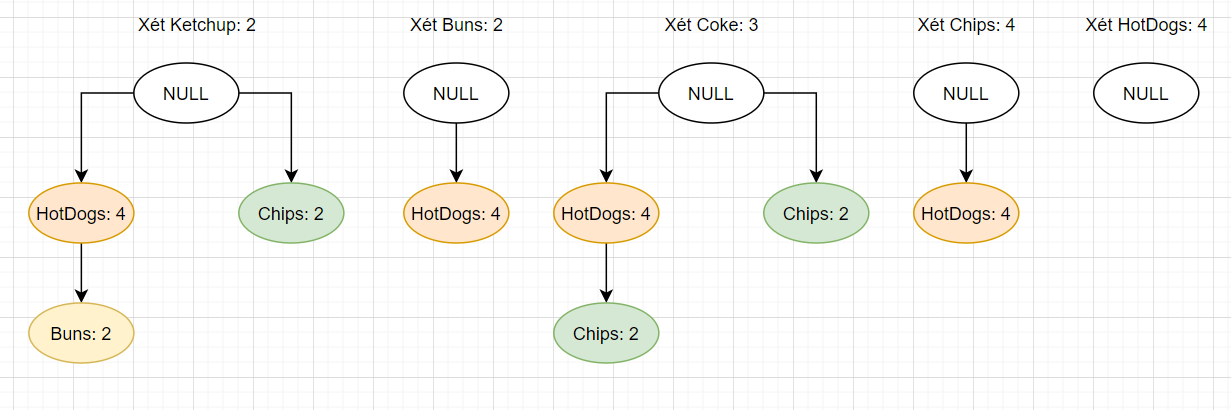


* Cơ sở mẫu điều kiện:

Duyệt bảng Header theo thứ tự từ phần tử có độ hỗ trợ thấp nhất đến cao nhất ta có:

|  |  |  |
| --- | --- | --- |
| ITEM | CONDITIONAL PATTERN BASE | SUPPORT COUNT |
| Ketchup | HotDogs, Buns | 1 |
| Chips | 1 |
| Buns | HotDogs | 2 |
| Coke | HotDogs, Chips | 2 |
| Chips | 1 |
| Chips | HotDogs | 2 |
| NULL | 2 |
| HotDogs | NULL | |

* Các cây FP điều kiện:



* Xây dựng tập phổ biến:

Xét cây FP chưa nhiều hơn 1 nhánh:

* Ketchup: 2

Ta tiến hành đệ quy trở lại để xây dựng cây như sau:

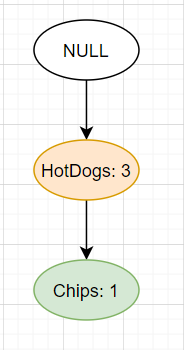
Vì HotDogs, Buns, Chips đều có support count = 1 < min\_sup\_count =2, do đó ta không xét 3 nút đó.

* Coke: 3

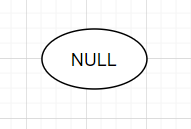
Ta tiến hành đệ quy trở lại để xây dựng cây như sau:

Vì Ketchup, Buns đều có support count = 0 < min\_sup\_count = 2, do đó ta không xét 2 nút đó.

Xét nút HotDogs, với {Coke, HotDogs}: 2



Xét nút Chips, với {Coke, Chips}: 3



Lúc này tất cả những cây FP điều kiện đều đã có dạng đường dẫn đơn. Ta tiến hành tổ hợp những phần tử để tìm ra các tập phổ biến 2 phần tử trở lên thỏa min\_sup\_count:

L2 = { {HotDogs, Buns}, {HotDogs, Coke}, {HotDogs, Chips}, {Coke, Chips} }

L3 = { { HotDogs, Coke, Chips} }

* Tìm các luật kết hợp:

Xét tập phổ biến {HotDogs, Coke, Chips}: 2 có các tập con không rỗng sau: {HotDogs}: 4, {Coke}: 3, {Chips}: 4, {HotDogs, Coke}: 2, }, {HotDogs, Chips}: 2, {Coke, Chips}: 3

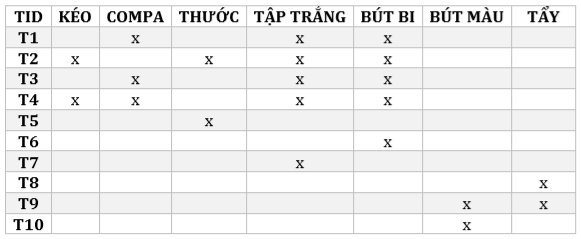
|  |  |
| --- | --- |
| HotDogs => {Coke, Chips} | 2/4 = 50% |
| Coke => {HotDogs, Chips} | 2/3 = 66,66% |
| Chips => {HotDogs, Coke} | 2/4 = 50% |
| {Coke, Chips} => HotDogs | 2/3 = 66,66% |
| {HotDogs, Chips} => Coke | 2/2 = 100% |
| {HotDogs, Coke} => Chips | 2/2 = 100% |

Với min\_conf = 60%, dựa vào bảng trên ta có các luật kết hợp thỏa yêu cầu là:

* Coke => {HotDogs, Chips}
* {Coke, Chips} => HotDogs
* {HotDogs, Chips} => Coke
* {HotDogs, Coke} => Chips

## BÀI TẬP 2: CỬA HÀNG VĂN PHÒNG PHẨM

Cho bảng dữ liệu ở một cửa hàng văn phòng phẩm như sau:



Với min\_sup = 30% và min\_conf = 80%

Ta quy đổi tên các phần tử trên như sau:

* KÉO => K
* COMPA => C
* THƯỚC => TC
* TẬP TRẮNG => TT
* BÚT BI => BB
* BÚT MÀU => BM
* TẨY => T

### Tìm tập phổ biến bằng giải thuật Apriori:

Với min\_sup là 30% và tổng số giao dịch là 10 => Tần số xuất hiện tối thiểu của phần tử để thỏa min\_sup (min\_sup\_count) là 3.

* Tập các ứng viên 1 phần tử và tần số xuất hiện (support count) tương ứng:

C1 = { K: 2, C: 3, TC: 2, TT: 5, BB: 5, BM: 2, T: 2}

* Các tập phổ biến 1 phần tử (Tập các ứng viên thỏa mãn min\_support\_count):

L1 = { {K}, {C}, {TC}, {TT}, {BB}, {BM}, {T} }

* Tập các ứng viên có 2 phần tử và support count tương ứng:

C2 = { {K,C}: 1, {K,TC}: 1, {K,TT}: 2, {K,BB}: 2, {C,TT}: 3, {C,BB}: 3, {TC,TT}: 1, {TC,BB}: 1, {TT,BB}: 4, {BM,T}: 1 }

* Các tập phổ biến 2 phần tử thỏa min\_support\_count:

L1 = { {C,TT}, {C,BB}, {TT,BB} }

* Tập các ứng viên có 3 phần tử và support count tương ứng:

C3 = { {C,TT,BB}: 3 }

* Tập phổ biến 3 phần tử thỏa min\_support\_count:

L3 = { {C,TT,BB} }

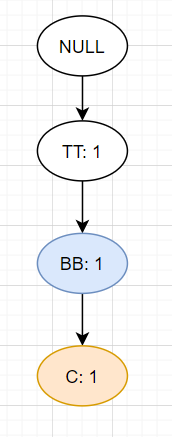
### Tìm tập phổ biến bằng giải thuật FP\_growth:

* Tập phổ biến 1 phần tử và tần số xuất hiện (support count) tương ứng:
* K: 2
* C: 3
* TC: 2
* TT: 5
* BB: 5
* BM: 2
* T: 2
* F-List (Danh sách giảm dần theo độ hỗ trợ (Support):

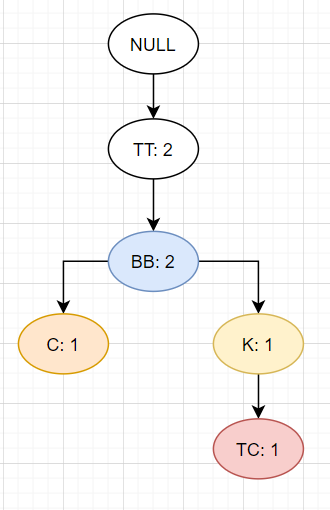
|  |  |
| --- | --- |
| ITEMS | SUPPORT COUNT |
| TT | 5 |
| BB | 5 |
| C | 3 |
| K | 2 |
| TC | 2 |
| BM | 2 |
| T | 2 |

* Xây cây:

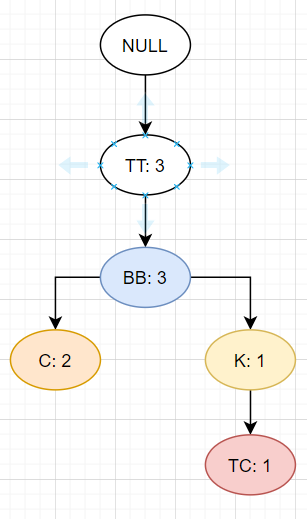
Thêm giao dịch 1 vào cây FP



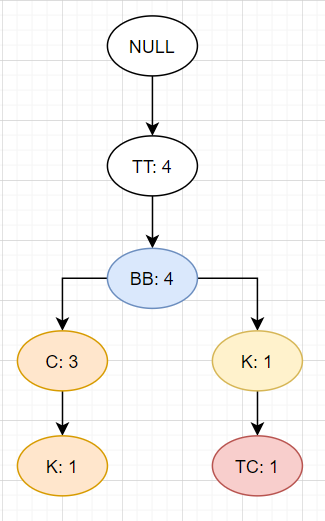
Thêm giao dịch 2 vào cây FP



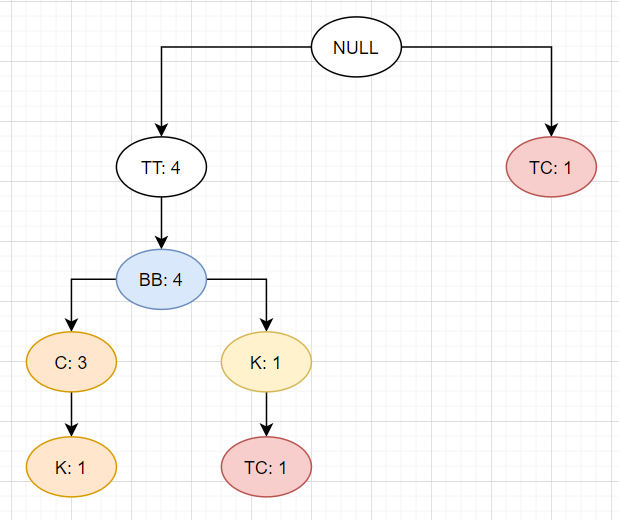
Thêm giao dịch 3 vào cây FP



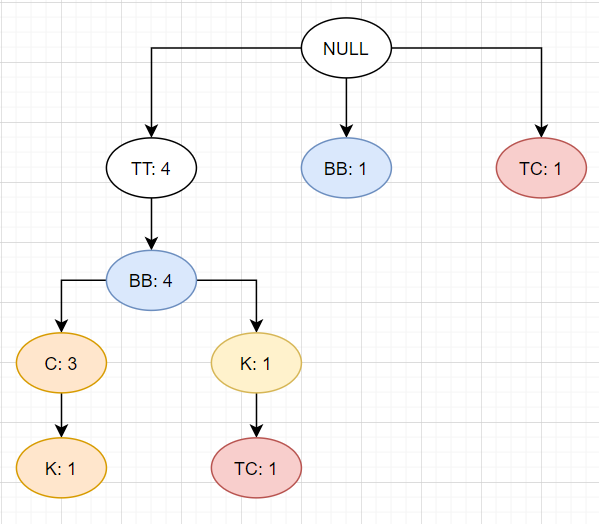
Thêm giao dịch 4 vào cây FP



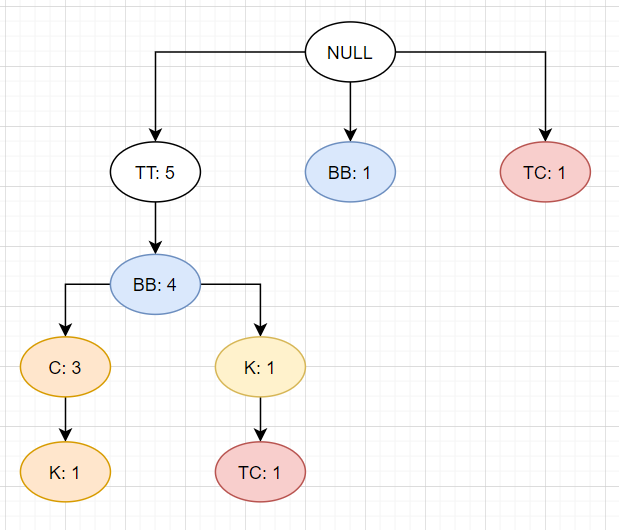
Thêm giao dịch 5 vào cây FP



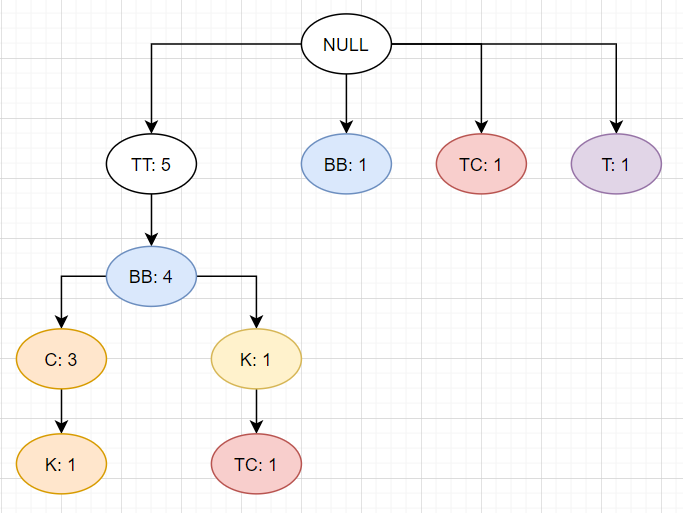
Thêm giao dịch 6 vào cây FP



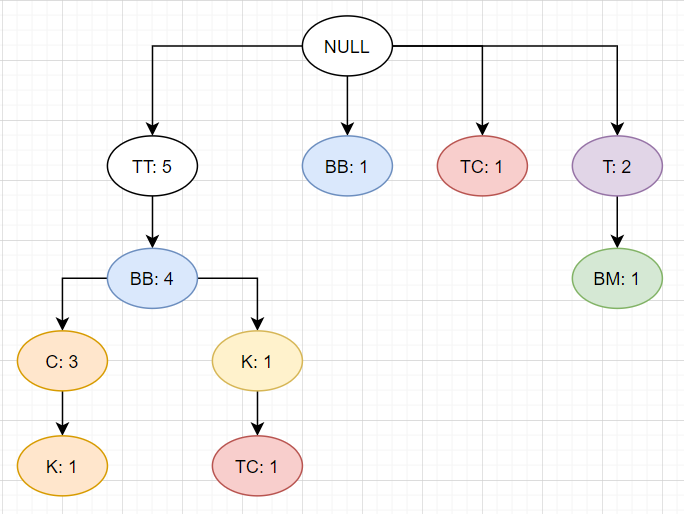
Thêm giao dịch 7 vào cây FP



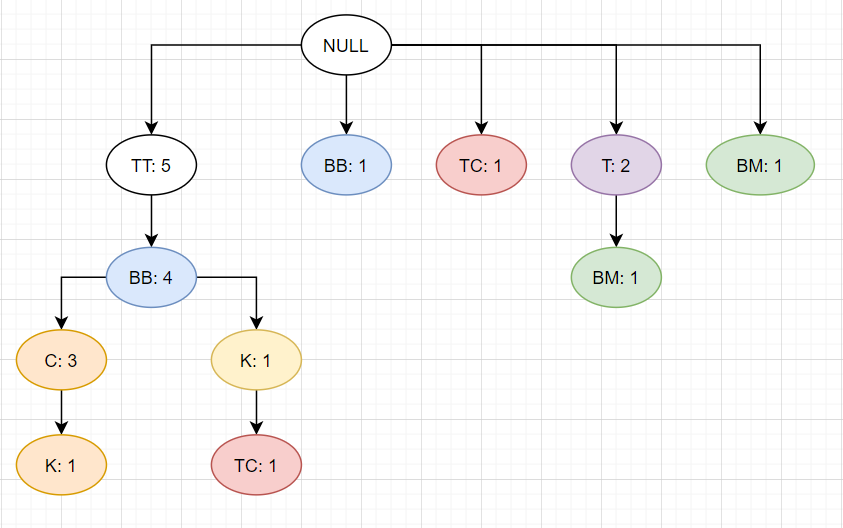
Thêm giao dịch 8 vào cây FP



Thêm giao dịch 9 vào cây FP



Thêm giao dịch 10 vào cây FP

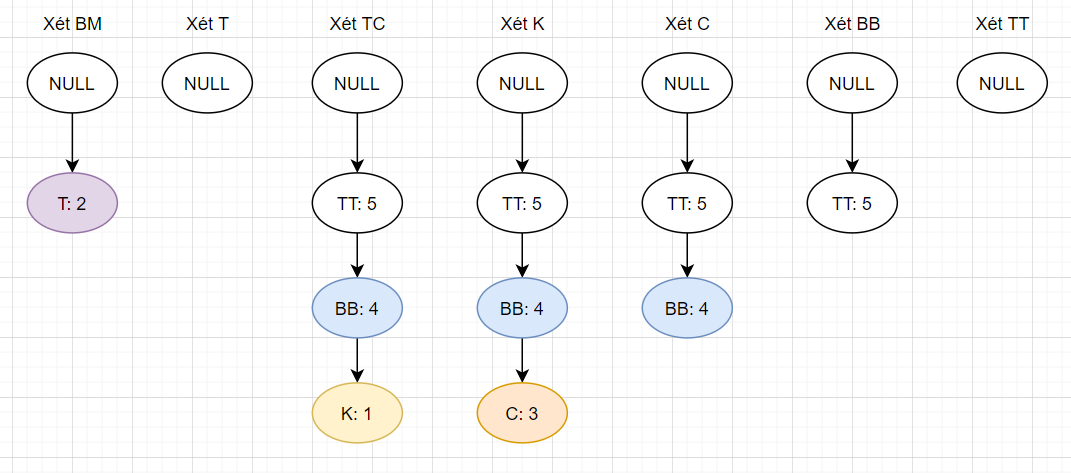


* Cơ sở mẫu điều kiện:

Dùng bảng Header theo thứ tự từ phần tử có độ hỗ trợ thấp nhất đến cao nhất ta có:

|  |  |  |
| --- | --- | --- |
| ITEM | CONDITIONAL PATTERN BASE | SUPPORT COUNT |
| T | NULL | |
| BM | NULL | 1 |
| T | 1 |
| TC | TT, BB, K | 1 |
| NULL | 1 |
| K | TT, BB, C | 1 |
| TT, BB | 1 |
| C | TT, BB | 3 |
| BB | TT | 4 |
| NULL | 1 |
| TT | NULL | |

* Các cây FP điều kiện:



* Xây dựng tập phổ biến:

Xét cây FP chứa nhiều hơn 1 nhánh:

* BB: 2

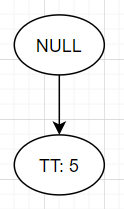
Ta tiến hành đệ quy trở lại để xây dựng cây như sau:

Vì K, TC, BM, T đều có support count < min\_sup\_count = 3, do đó ta không xét 3 nút đó.

Xét nút TT, với {BB, TT}: 4



Xét nút C, với {C, BB}: 3



Lúc này tất cả những cây FP điều kiện đều đã có dạng đường dẫn đơn. Ta tiến hành tổ hợp những phần tử để tìm ra các tập phổ biến 2 phần tử trở nên thỏa min\_sup\_count:

L2 = { {C, TT}, {C, BB}, {TT, BB} }

L3 = { {C, TT, BB} }

* K: 2

Ta tiến hành đệ quy trở lại để xây dựng cây như sau:

Vì C, TC, TT, BB đều có support count < min\_sup\_count = 3, do đó ta không xét 4 nút đó.

* TC: 2

Ta tiến hành đệ quy trở lại để xây dựng cây như sau:

Vì K, TT, BB đều có support count < min\_sup\_count = 3, do đó ta không xét 3 nút đó.

* BM: 2

Ta tiến hành đệ quy trở lại để xây dựng cây như sau:

Vì T có support count < min\_sup\_count = 3, do đó ta không xét nút này.

* Tìm các luật kết hợp:

Xét tập phổ biến {C, TT, BB}: 3 có các tập con không rỗng sau: {C}: 3, {TT}: 5, {BB}: 5, {C, TT}: 3, {C, BB}: 3, {TT, BB}: 4

|  |  |
| --- | --- |
| C => {TT, BB} | 3/3 = 100% |
| TT => {C, BB} | 3/5 = 60% |
| BB => {C, TT} | 3/5 = 60% |
| {TT, BB} => C | 3/4 = 75% |
| {C, BB} => TT | 3/3 = 100% |
| {C, TT} => BB | 3/3 = 100% |

Với min\_conf = 80%, dựa vào bảng trên ta có các luật kết hợp thỏa yêu cầu là:

* C => {TT,BB}
* {C,BB} => TT
* {C,TT} => BB

## BÀI TẬP 3: CSDL VỀ NHÂN VIÊN:

Cho bảng CSDL về nhân viên như sau:



Cho B = {Tuổi, Năng lực làm việc, Thăng chức}. Hãy tìm tất cả các luật kết hợp có vế phải chỉ gồm thuộc tính Thăng chức (TC) thỏa ngưỡng min\_sup = 30% và min\_conf = 80%

Với min\_sup = 30% và tổng số dòng dữ liệu là 10 => Tần số xuất hiện tối thiểu của phần tử để thỏa min\_sup\_count là 3

* Với thuộc tính TC có giá trị là “Có”

Các luật có thể tìm được là:

GT = Nữ => TC = Có [40%, 80%]

GT = Nam => TC = Có [10%, 20%]

T = 20..25 => TC = Có [10%, 20%]

T = 26..30 => TC = Có [30%, 60%]

T = 31..40 => TC = Có [10%, 20%]

NL = Giỏi => TC = Có [30%, 60%]

NL = Khá => TC = Có [10%, 20%]

NL = T.Bình => TC = Có [10%, 20%]

LGD = Rồi => TC = Có [20%, 40%]

LGD = Chưa => TC = Có [30%, 60%]

TN = Rất cao => TC = Có [10%, 20%]

TN = Cao => TC = Có [10%, 20%]

TN = Khá => TC = Có [20%, 40%]

TN = T.Bình => TC = Có [10%, 20%]

Dựa vào các luật trên, luật thỏa min\_conf = 80% là

GT = Nữ => TC = Có [40%,80%]

* Với thuộc tính TC có giá trị là “Không”:

Các luật có thể tìm được là:

GT = Nữ => TC = Không [30%, 60%]

GT = Nam => TC = Không [20%, 40%]

T = 20..25 => TC = Không [10%, 20%]

T = 26..30 => TC = Không [20%, 40%]

T = 31..40 => TC = Không [10%, 20%]

T > 40 => TC = Không [10%, 20%]

NL = Giỏi => TC = Không [10%, 20%]

NL = Khá => TC = Không [30%, 60%]

NL = T.Bình => TC = Không [10%, 20%]

LGD = Rồi => TC = Không [20%, 40%]

LGD = Chưa => TC = Không [30%, 60%]

TN = Rất cao => TC = Không [10%, 20%]

TN = Cao => TC = Không [10%, 20%]

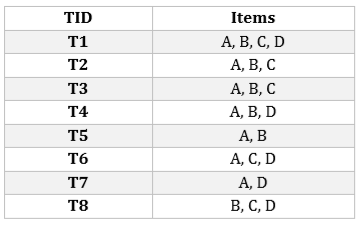
TN = Khá => TC = Không [10%, 20%]

TN = T.Bình => TC = Không [20%, 40%]

Dựa vào các luật trên, không có luật nào thỏa min\_conf = 80%

## BÀI TẬP 4: DỮ LIỆU CÔNG TY:

Cho bảng dữ liệu ở một công ty có các giao dịch sau:



Với min\_support\_count > 1

### Tìm tập phổ biến bằng giải thuật Apriori:

Với min\_sup\_count > 1 => Tần số xuất hiện tổi thiểu của phần tử là 2

* Tập các ứng viên 1 phần tử và tần số xuất hiện phần tử (support count) tương ứng:

C1 = { A: 7, B: 6, C: 5, D: 5 }

* Các tập phổ biến 1 phần tử (Tập các ứng viên thỏa mãn min\_sup\_count):

L1 = { {A}, {B}, {C}, {D} }

* Tập có ứng viên có 2 phần tử và support count tương ứng:

C2 = { {A,B}: 5, {A,C}: 4, {A,D}: 4, {B,C}: 4, {B,D}: 3, {C,D}: 3 }

* Các tập phổ biến 2 phần tử thỏa min\_support\_count:

L1 = { {A,B} , {A,C} , {A,D} , {B,C} , {B,D} , {C,D} }

* Tập có ứng viên có 3 phần tử và support count tương ứng:

C3 = { {A,B,C}: 3, {A,B,D}: 2, {A,C,D}: 2, {B,C,D}: 2 }

* Các tập phổ biến 3 phần tử thỏa min\_support\_count:

L3 = { {A,B,C}, {A,B,D}, {A,C,D}, {B,C,D} }

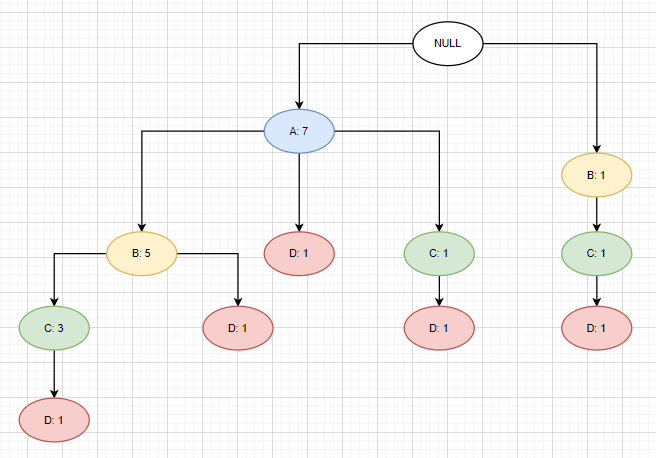
### Tìm tập phổ biến bằng giải thuật FP-growth:

* Tập phổ biến 1 phần tử và tần số xuất hiện (support count) tương ứng:
* A: 7
* B: 6
* C: 5
* D: 5
* F-List (Danh sách giảm dần theo dộ hỗ trợ (Support)):

|  |  |
| --- | --- |
| ITEM | SUPPORT COUNT |
| A | 7 |
| B | 6 |
| C | 5 |
| D | 5 |

* Xây cây:

Thêm lần lượt 8 giao dịch vào cây FP, ta được cây FP sau:

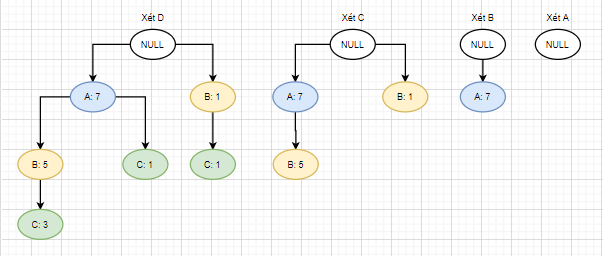


* Cơ sở mẫu điều kiện:

Dùng bảng Header theo thứ tự từ phần tử có độ hỗ trợ thấp nhất đến cao nhất ta có:

|  |  |  |
| --- | --- | --- |
| ITEM | CONDITIONAL PATTERN BASE | SUPPORT COUNT |
| D | A | 1 |
| A,B | 1 |
| A,C | 1 |
| A,B,C | 1 |
| B,C | 1 |
| C | A | 1 |
| B | 1 |
| A,B | 3 |
| B | A | 5 |
| NULL | 1 |
| A | NULL | |

* Các cây FP điều kiện:



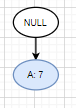
* Xây dựng tập phổ biến:

Xét cây FP chứa nhiều hơn 1 nhánh:

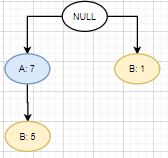
* D: 5

Ta tiến hành đệ quy trở lại để xây dựng cây như sau:

Xét nút B, với {B,D}: 3



Xét nút C, với {C,D}: 3



Xét nút A, với {A,D}: 4



Lúc này tất cả những cây FP điều kiện đều đã có dạng đường dẫn đơn. Ta tiến hành tổ hợp những phần tử để tìm ra các tập phổ biến 2 phần tử trở lên thỏa min\_sup\_count:

L2 = { {A,B} , {A,C} , {A,D} , {B,C} , {B,D} , {C,D} }

L3 = { {A,B,C}, {A,B,D}, {A,C,D}, {B,C,D} }

* Tìm các luật kết hợp:

Xét tập phổ biến {A,B,C}: 3 có các tập con không rỗng sau: {A}: 7, {B}: 6, {C}: 5, {A,B}: 5, {A,C}: 4, {B,C}: 4

|  |  |
| --- | --- |
| A => {B,C} | 3/7 = 43% |
| B => {A,C} | 3/6 = 50% |
| C => {A,B} | 3/5 = 60% |
| {B,C} => A | 3/4 = 75% |
| {A,C} => B | 3/4 = 75% |
| {A,B} => C | 3/5 = 60% |

Dựa vào bảng trên, tất cả luật kết hợp đều đạt yêu cầu.