**Bài tập con phần công nghệ KDL và bài toán khai phá luật kết hợp**

1) Giả sử một kho dữ liệu bao gồm 3 chiều: thời gian, bác sỹ và bệnh nhân, và hai độ đo là số đếm và chi phí với chi phí là tổng số tiền mà một bệnh nhân phải trả cho một bác sỹ trong một lần khám bệnh.

(a) Hãy liệt kê ba loại lược đồ dữ liệu có thể được sử dụng để mô hình hóa kho dữ liệu và sử dụng một trong số chúng để vẽ lược đồ dữ liệu cho kho dữ liệu mô tả ở trên.

(b) Xuất phát từ khối cơ bản [ngày, bác sỹ, bệnh nhân], những thao tác OLAP nào cần được thực hiện để liệt kê tổng số tiền kiếm đuợc của mỗi bác sỹ trong năm 2012?

(c) Viết các câu lệnh DMQL để định nghĩa lược đồ dữ liệu trong câu (a) và câu lệnh SQL để tính tổng tiền trong câu (b)

a) Bao gồm lược đồ hình sao, hình bông tuyết và dải ngân hà

Bài này chọn lược đồ hình sao.

DimBacSy(ma\_bac\_sy, ho\_ten, so\_dien\_thoai, chuyen\_mon)

DimBenhNhan(ma\_benh\_nhan, ho\_ten, so\_dien\_thoai, ngay\_sinh)

DimThoiGian(ma\_thoi\_gian, ngay, thang, nam)

FactKhamBenh(ma\_bac\_sy, ma\_benh\_nhan, ma\_thoi\_gian, so\_dem, chi\_phi)

So\_dem là đếm số lần khám trong 1 ngày theo ma\_thoi\_gian của 1 bác sỹ có mã là ma\_bac\_sy cho 1 bệnh nhân có mã ma\_benh\_nhan.

Chi\_phi là chi phí khám bệnh trong 1 ngày theo ma\_thoi\_gian của 1 bác sỹ có mã là ma\_bac\_sy cho 1 bệnh nhân có mã ma\_benh\_nhan.

b) **Phải vẽ các phân cấp của các chiều trước (ở đây là time). Một chiều có thể có nhiều phân cấp, phải nói rõ là theo phân cấp nào.**

Khối cơ bản [ngày, bác sỹ, bệnh nhân]

Tổng số tiền kiếm được của mỗi bác sỹ trong năm 2012 => 1 chiều bác sỹ

Roll up 1 lần để giảm số chiều [bệnh nhân] đi thành [ngày, bác sỹ]

Roll up 1 lần để đưa phân cấp của chiều thời gian từ thấp lên cao thành [tháng, bác sỹ]

Roll up 1 lần để đưa phân cấp của chiều thời gian từ thấp lên cao thành [năm, bác sỹ]

Slice 1 lần để chỉ lấy các bản ghi có nam = “2012”

c) DMQL:

define cube KhamBenh [DimBacSy, DimBenhNhan, DimThoiGian]:

so\_dem = count(\*), chi\_phi = sum(hospital\_fee)

**(Phải chú thích \* và hospital\_fee là gì ?)**

define dimension DimBacSy as (ma\_bac\_sy, ho\_ten, so\_dien\_thoai, chuyen\_mon)

define dimension DimBenhNhan as (ma\_benh\_nhan, ho\_ten, so\_dien\_thoai, ngay\_sinh)

define dimension DimThoiGian as (ma\_thoi\_gian, ngay, thang, nam)

SQL truy vấn tổng tiền cho b):

SELECT FactKhamBenh.ma\_bac\_sy, DimBacSy.ho\_ten, SUM(FactKhamBenh.chi\_phi) AS tong\_thu\_nhap

FROM FactKhamBenh

LEFT JOIN DimBacSy ON DimBacSy.ma\_bac\_sy = FactKhamBenh.ma\_bac\_sy

LEFT JOIN DimThoiGian ON DimThoiGian.ma\_thoi\_gian = FactKhamBenh.ma\_thoi\_gian WHERE DimThoiGian.nam = 2012

GROUP BY FactKhamBenh.ma\_bac\_sy, DimBacSy.ho\_ten

2) Một cơ sở dữ liệu có 5 giao dịch như dưới đây. Giả sử min sup = 60% và min conf = 80%.

|  |  |
| --- | --- |
| **Mã giao dịch** | **Mặt hàng mua** |
| T100 | A, B, C, D, E | |
| T200 | B, E, A, C, H | |
| T300 | P, A, G, E | |
| T400 | U, R, B, A, N, A | |
| T500 | G, A, U, D, I | |

1. Hãy tìm tất cả những tập mặt hàng thường xuyên sử dụng thuật toán Apriori và FP-growth. So sánh tính hiệu quả của hai phương pháp khai phá trên.

**Thuật toán Apriori:**

* + K = 1:

A với s = (A) / 5 = 5 / 5 = 1 >= minsup

B với s = (B) / 5 = 3 / 5 >= minsup

C với s = (C) / 5 = 2 / 5 < minsup

D với s = (D) / 5 = 2 / 5 < minsup

E với s = (E) / 5 = 3 / 5 >= minsup

H với s = (H) / 5 = 1 / 5 < minsup

P với s = (P) / 5 = 1 / 5 < minsup

G với s = (G) / 5 = 2 / 5 < minsup

U với s = (U) / 5 = 2 / 5 < minsup

R với s = (R) / 5 = 1 / 5 < minsup

N với s = (N) / 5 = 1 / 5 < minsup

I với s = (I) / 5 = 1 / 5 < minsup

=> Mặt hàng thường xuyên là {A}, {B}, {E}

* + K = 2:

{AB} với s = (AB) / 5 = 3 / 5 >= minsup

{AE} với s = (AE) / 5 = 3 / 5 >= minsup

{BE} với s = (BE) / 5 = 2 / 5 < minsup

=> Tập mặt hàng thường xuyên là {AB}, {AE}

* + K = 3:

{ABE} với s = (ABE) / 5 = 2 / 5 < minsup

=> Không có thêm tập mặt hàng nào

=> Các tập mặt hàng thường xuyên với thuật toán Apriori là {A, B, E, AB, AE}

**Thuật toán FP-Growth:**

Trước khi thực hiện, sắp xếp lại các mặt hàng trong mỗi giao dịch.

Riêng giao dịch có mã T400 được lược bớt 1 mặt hàng mua A vì thuật toán không quan tâm số lượng mặt hàng mua trong giao dịch:

|  |  |
| --- | --- |
| **Mã giao dịch** | **Mặt hàng mua** |
| T100 | A, B, C, D, E | |
| T200 | A, B, C, E, H | |
| T300 | A, E, G, P | |
| T400 | A, B, N, R, U | |
| T500 | A, D, G, I, U | |

Cây FP:

A diagram of a tree

AI-generated content may be incorrect.

Xét cây cơ sở điều kiện cho E: P = {(A:1, B:1, C:1, D:1, E:1), (A:1, B:1, C:1, E:1), (A:1, E:1)}. E xuất hiện 3 lần => Độ hỗ trợ s = 3 / 5 = 0.6 >= minsup => {E} là tập mặt hàng thường xuyên, cắt E khỏi cây rồi tiếp tục đệ quy trên cây này.

Xét cây cơ sở điều kiện cho D nằm trong cây điều kiện cho E: P = {(A:1, B:1, C:1, D:1)}. D xuất hiện 3 lần => Độ hỗ trợ s = 1 / 5 = 0.2 < minsup => Dừng đệ quy tại nút D.

Xét cây cơ sở điều kiện cho C nằm trong cây điều kiện cho E: P = {(A:2, B:2, C:2)}. C xuất hiện 2 lần => Độ hỗ trợ s = 2 / 5 = 0.4 < minsup => Dừng đệ quy tại nút C.

Xét cây cơ sở điều kiện cho B nằm trong cây điều kiện cho E: P = {(A:2, B:2)}. B xuất hiện 2 lần => Độ hỗ trợ s = 2 / 5 = 0.4 < minsup => Dừng đệ quy tại nút B.

Xét cây cơ sở điều kiện cho A nằm trong cây điều kiện cho E: P = {(A:3)}. A xuất hiện 3 lần => Độ hỗ trợ s = 3 / 5 = 0.6 >= minsup => {EA} là tập mặt hàng thường xuyên, cắt A khỏi cây thì hết nút, dừng đệ quy tại nút A.

Xét cây cơ sở điều kiện cho H: P = {(A:1, B:1, C:1, E:1, H:1)}. H xuất hiện 1 lần => Độ hỗ trợ s = 1 / 5 = 0.2 < minsup => Dừng đệ quy tại nút H.

Xét cây cơ sở điều kiện cho U: P = {(A:1, B:1, N:1, R:1, U:1), (A:1, D:1, G:1, I:1, U:1), }. U xuất hiện 2 lần => Độ hỗ trợ s = 2 / 5 = 0.4 < minsup => Dừng đệ quy tại nút U.

Xét cây cơ sở điều kiện cho U: P = {(A:1, B:1, N:1, R:1, U:1), (A:1, D:1, G:1, I:1, U:1), }. U xuất hiện 2 lần => Độ hỗ trợ s = 2 / 5 = 0.4 < minsup => Dừng đệ quy tại nút U.

Xét cây cơ sở điều kiện cho D: P = {(A:1, B:1, C:1, D:1), (A:1, D:1)}. D xuất hiện 2 lần => Độ hỗ trợ s = 2 / 5 = 0.4 < minsup => Dừng đệ quy tại nút D.

Xét cây cơ sở điều kiện cho R: P = {(A:1, B:1, N:1, R:1)}. R xuất hiện 2 lần => Độ hỗ trợ s = 1 / 5 = 0.2 < minsup => Dừng đệ quy tại nút R.

Xét cây cơ sở điều kiện cho P: P = {(A:1, E:1, G:1, P:1)}. P xuất hiện 1 lần => Độ hỗ trợ s = 1 / 5 = 0.2 < minsup => Dừng đệ quy tại nút P.

Xét cây cơ sở điều kiện cho I: P = {(A:1, D:1, G:1, I:1)}. I xuất hiện 1 lần => Độ hỗ trợ s = 1 / 5 = 0.2 < minsup => Dừng đệ quy tại nút I.

Xét cây cơ sở điều kiện cho C: P = {(A:2, B:2, C:2)}. C xuất hiện 2 lần => Độ hỗ trợ s = 2 / 5 = 0.4 < minsup => Dừng đệ quy tại nút C.

Xét cây cơ sở điều kiện cho N: P = {(A:1, B:1, N:1)}. N xuất hiện 1 lần => Độ hỗ trợ s = 1 / 5 = 0.2 < minsup => Dừng đệ quy tại nút N.

Xét cây cơ sở điều kiện cho G: P = {(A:1, E:1, G:1), (A:1, D:1, G:1)}. G xuất hiện 2 lần => Độ hỗ trợ s = 2 / 5 = 0.4 < minsup => Dừng đệ quy tại nút G.

Xét cây cơ sở điều kiện cho B: P = {(A:3, B:3)}. B xuất hiện 3 lần => Độ hỗ trợ s = 3 / 5 = 0.6 >= minsup => {B} là tập mặt hàng thường xuyên, cắt B khỏi cây rồi tiếp tục đệ quy trên cây này.

Xét cây cơ sở điều kiện cho A nằm trong cây điều kiện cho B: P = {(A:3)}. A xuất hiện 3 lần => Độ hỗ trợ s = 3 / 5 = 0.6 >= minsup => {BA} là tập mặt hàng thường xuyên, cắt A khỏi cây thì hết nút, dừng đệ quy tại nút A.

Xét cây cơ sở điều kiện cho A: P = {(A:5)}. A xuất hiện 5 lần => Độ hỗ trợ s = 5 / 5 = 1 >= minsup => {A} là tập mặt hàng thường xuyên, cắt A khỏi cây thì hết nút, dừng đệ quy tại nút A.

Vậy ta có các tập mặt hàng thường xuyên là {A, B, E, AB, AE}

1. Hãy liệt kê tất cả những luật kết hợp (với độ hỗ trợ s và độ tin cậy c) tuân thủ theo siêu luật được mô tả dưới đây, trong đó *X* là một biến đại diện cho khách hàng và *mhi* thể hiện các biến đại điện cho các mặt hàng (ví dụ "A", "B", v.v...): cho tất cả x trong giao dịch; mua(X, mh1) 🡪 mua(X, mh2) [s, c]

Với các tập mặt hàng thường xuyên là {A, B, E, AB, AE}, chỉ các tập có 2 phần tử trở lên mới có thể sinh luật => các tập mặt hàng {AB} và {AE}

Xét A 🡪 E: c(A 🡪 E) = s(AE) / s(A) = (AE) / (A) = 3 / 5 = 0.6 < minconf = 0.8

Xét E 🡪 A: c(E 🡪 A) = (AE) / (E) = 3 / 3 = 1 >= 0.8

Xét A 🡪 B: c(A 🡪 B) = (AB) / (A) = 3 / 5 = 0.6 < 0.8

Xét B 🡪 A: c(B 🡪 A) = (AB) / (B) = 3 / 3 = 1 >= 0.8

Vậy ta được các luật kết hợp:

mua(X, B) 🡪 mua(X, A) và mua(X, E) 🡪 mua(X, A)