Bài tập thêm \_ Câu 3

Tìm hiểu thuật toán: Apriori+, FPMax

* Apriori +:
* Apriori dùng cách tiếp cận lặp được biết đến như tìm kiếm level-wise, với các tập k item được dùng để thăm dò các tập (k+1) item.
* Đầu tiên, tập(frequent 1- itemsets) phổ biến 1 được tìm thấy ký hiệu là C1
* Bước tiếp theo là tính support có nghĩa là sự xuất hiện của các item trong cơ sở dữ liệu. Điều này đòi hỏi phải duyệt qua toàn bộ cơ sở dữ liệu.
* Sau đó, bước cắt tỉa được thực hiện trên C1 trong đó những item được so sánh với thông số minimum support. Những item thỏa điều kiện minimum support mới được xem xét cho tiến trình tiếp theo ký hiệu là L1.
* Sau đó, bước phát sinh các bộ ứng viên được thực hiện trong đó tập phổ biến 2 được tạo ra ký hiệu là C2.
* Một lần nữa, cở sở dữ liệu được duyệt để tính toán support của 2 tập phổ biến. Theo minimum support, các bộ ứng viên tạo ra được kiểm tra và chỉ những tập phổ biến nào thỏa điều kiện minimum support thì tiếp tục được sử dụng tạo ra bô ứng viên tập phổ biến 3.
* Bước trên tiếp tục cho đến khi không có tập phổ biến hoặc bộ ứng viên có thể được tạo ra.
* Mã giả thuật toán

Apriori\_Algorithm()  
{  
Ck: Candidate itemset of size k  
Lk : frequent itemset of size k  
L1 = {frequent items};  
**for** (k = 1; Lk!=0; k++)  
**{**  
          Ck+1 = candidates generated from Lk;  
**foreach** transaction t in database do  
                 increment the count of all candidates in  Ck+1                       
                    that are contained in t  
          Lk+1  = candidates in Ck+1 with min\_support  
 **}**  
return k

* FPMax:
* Giải thích cách chạy thuật toán FPMax bằng cách sử dụng thư viện khai thác dữ liệu nguồn mở SPMF.
* Nếu bạn đang sử dụng giao diện đồ họa, (1) chọn thuật toán "FPMax", (2) chọn tệp đầu vào "contextPasquier99.txt", (3) đặt tên tệp đầu ra (ví dụ: "output.txt") (4) đặt minsup thành 40% và (5) nhấp vào "Chạy thuật toán".
* Nếu bạn muốn thực hiện ví dụ này từ dòng lệnh, hãy thực hiện lệnh này:
* java -jar spmf.jar chạy FPMax contextPasquier99.txt output.txt 40% trong thư mục chứa spmf.jar và tệp đầu vào ví dụ contextPasquier99.txt.
* Đầu vào là cơ sở dữ liệu giao dịch (hay còn gọi là ngữ cảnh nhị phân) và ngưỡng có tên minsup (giá trị từ 0 đến 100%).
* Cơ sở dữ liệu giao dịch là một tập hợp các giao dịch. Mỗi giao dịch là một tập hợp các mục.
* Tập hợp các mục là một tập hợp các mục riêng biệt không có thứ tự. Hỗ trợ của một tập phổ biến là số lượng giao dịch có chứa tập hợp mục đó.
* Tập phổ biến là một tập hợp xuất hiện trong ít nhất các giao dịch minsup từ cơ sở dữ liệu giao dịch. Tập phổ biến thường xuyên đóng là tập phổ biến thường xuyên không được bao gồm trong một tập thay thế thích hợp có cùng hỗ trợ. Tập phổ biến cực đại thường xuyên là tập phổ biến thường xuyên không được bao gồm trong tập thay thế thích hợp là tập phổ biến thường xuyên. Do đó, tập hợp các tập phổ biến tối đa thường xuyên là một tập con của tập các tập phổ đóng thường xuyên, là một tập con của các tập phổ biến