BÁO CÁO MÔN KHAI THÁC DỮ LIỆU

LAB 02 FIM

THÔNG TIN SINH VIÊN:

Họ và tên: Phạm Lưu Mỹ Phúc

MSSV: 19120331

Ý TƯỞNG

PHẦN 1: THUẬT TOÁN APRIORI

Thuật toán Apriori được đề xuất bởi R. Agrawal và R. Srikant vào năm 1994 để khai thác tập phổ biến.

- Sử dụng phương pháp tiếp cận từ dưới lên (“bottom-up”), dữ liệu được kiểm tra dựa trên các tập dữ liệu.
- Sử dụng BFS để tính số lượng tập ứng viên.
- Thực hiện lưu ***non_frequent_itemset*** của tập 2_itemset để thực hiện thêm bước duyệt xem phần tử trong tập ứng viên có thuộc tập không phổ biến hay không trước khi tính support của itemset đó. Thực hiện bước này giúp giảm số lượng tập ứng viên**.**

Quy tắc Apriori:

- Nếu một tập là phổ biến thì tất cả tập con của nó phải phổ biến.
- Nếu một tập không phổ biến thì tất cả tập chứa nó không phổ biến.
- Sử dụng tập phổ biến thứ k-1 để xây dựng tập ứng viên k.

PHẦN 2: THUẬT TOÁN TREE PROJECTION

- Nén cơ sở dữ liệu vào cây, cây này thể hiện tập phổ biến và thông tin kết hợp của các tập phổ biến.
- Node của Tree Projection là các tập phổ biến và được xem như đã được sắp xếp.
- Do được biểu diễn trên cây nên với các tập có chung tập con đường đi của chúng sẽ có đoạn chung.

CÁC BƯỚC THỰC HIỆN

PHẦN 1: THUẬT TOÁN APRIORI

F là tập hợp các tập phổ biến.

Bước 1: Lấy các tập ứng viên có 1 item trong dữ liệu thêm tập ứng viên, đếm độ hỗ trợ của từng item và thêm các tập phổ biến vào F, C_1

Bước 2: k=2. Phát sinh ứng viên C_2 dựa trên C_1. Tính độ phổ biến của từng tập và thêm tập phổ biến vào F, tập không phổ biến vào ***non_frequent_itemset.***

Bước 3: k=k+1

Bước 4: Phát sinh ứng viên của tập C_k dựa trên C_{k-1}.

Bước 3: Loại bỏ các ứng viên Ck chứa tập con thuộc ***non_frequent_itemset.***

Bước 4: Thêm các tập Ck thỏa ngưỡng minsup vào F.

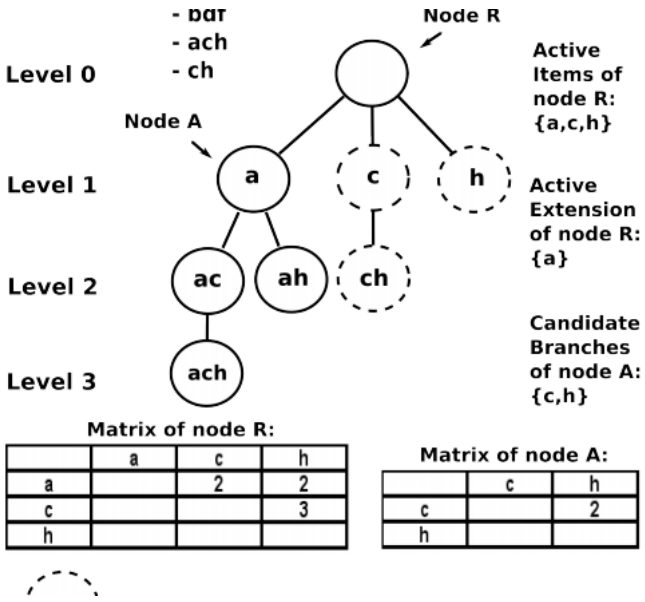
Thuật toán lặp cho đến khi tất cả các tập phổ biến được phát sinh.

PHẦN 2: THUẬT TOÁN TREE PROJECTION

Bước 1: Tạo nút gốc (null node) và các nút ở bậc 1.

Bước 2: k=1.

Bước 3: Tạo ma trận (tam giác trên) của bậc k-1 lưu độ hỗ trợ của các nút ở level k+1.



	a	c	h
a		2	2
c			3
h			

	c	h
c		2
h		

Bước 4: Kiểm tra trong ma trận ở bước 3 node nào là tập phổ biến. Tạo node là tập phổ biến ở bậc k+1.

Bước 5: Tỉa nhánh tất cả inactive node của cây.

Bước 6: k=k+1.

Bước 7: Kiểm tra bậc k của cây có null hay không. Nếu có thì ngừng, không thì quay lại bước 3.

ĐIỂM MẠNH

PHẦN 1: THUẬT TOÁN APRIORI

- Thuật toán đơn giản, dễ cài đặt.

PHẦN 2: THUẬT TOÁN TREE PROJECTION

- Đòi hỏi ít bộ nhớ hơn Apriori do thực hiện trên cấu trúc cây, hơn nữa quá trình khai phá các tập phổ biến không phải thông qua quá trình phát sinh tập ứng viên nên tiết kiệm được rất nhiều bộ nhớ.
- Giảm số lần quét trên cơ sở dữ liệu so với Apriori phải thực hiện nhiều lần quét để phát sinh tập ứng viên.
- Các đặc điểm nêu trên đồng thời giúp Tree Projection giảm thời gian thực hiện rất nhiều so với Apriori.

ĐIỂM YẾU

PHẦN 1: THUẬT TOÁN APRIORI

- Phải duyệt tập dữ liệu lại rất nhiều lần trong khi kích thước tập dữ liệu trong thực tế không hề nhỏ.
- Số lượng tập ứng viên phát sinh trong quá trình thực hiện thuật toán cũng rất lớn. Ví dụ, trường hợp tập ứng viên 1_itemset có 10^4 ứng viên thì khi phát sinh 2_itemset số lượng ứng viên lên đến hơn 10^7 ứng viên. Việc duyệt qua tất cả ứng viên này để loại bỏ ứng viên sẽ tốn rất nhiều thời gian và bộ nhớ.
- Việc tính support bị lặp đi lặp lại nhiều lần.

➔ Vừa tốn bộ nhớ vừa tốn thời gian rất nhiều.

PHẦN 2: THUẬT TOÁN TREE PROJECTION

- Chỉ chạy tốt trên các tập vừa và nhỏ

ỨNG DỤNG TRONG TÁC VỤ KHAI THÁC DỮ LIỆU

- Trong lĩnh vực y tế: phân tích cơ sở dữ liệu của bệnh nhân.
- Trong lâm nghiệp: phân tích xác suất và cường độ cháy rừng từ dữ liệu cháy rừng.
- Trong kinh doanh: xây dựng hệ thống đề xuất mua hàng cho các trang web bán hàng online hoặc sắp xếp các gian hàng một cách hợp lý nhất để thu hút người tiêu dùng và tăng doanh số.

TÀI LIỆU THAM KHẢO

- Apriori Algorithm In Data Mining: Implementation With Examples: https://www.softwaretestinghelp.com/apriori-algorithm
- Thuật toán Apriori khai phá luật kết hợp trong Data Mining: https://viblo.asia/p/thuat-toan-apriori-khai-pha-luat-ket-hop-trong-data-mining-3P0lPEv85ox
- A Tree Projection Algorithm For Generation of Frequent Itemsets: https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.103.9466&rep=rep1&type=pdf
- Tree Projection – Based Frequent Itemset Mining on Multicore CPUs and GPUs
- Khai phá dữ liệu và lớp bài toán khai thác các tập phổ biến - VIBLO