TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN
KHOA CÔNG NGHỆ THÔNG TIN

BÁO CÁO KHÓA LUẬN TỐT NGHIỆP

BẢO VỆ TÍNH RIÊNG TƯ TRONG KHAI THÁC MẪU DỰA TRÊN PHƯƠNG PHÁP TỐI ƯU NGẪU NHIÊN

Giảng viên hướng dẫn:
ThS. Nguyễn Ngọc Đức

Sinh viên thực hiện:
Trần Khắc Bình

Nội dung

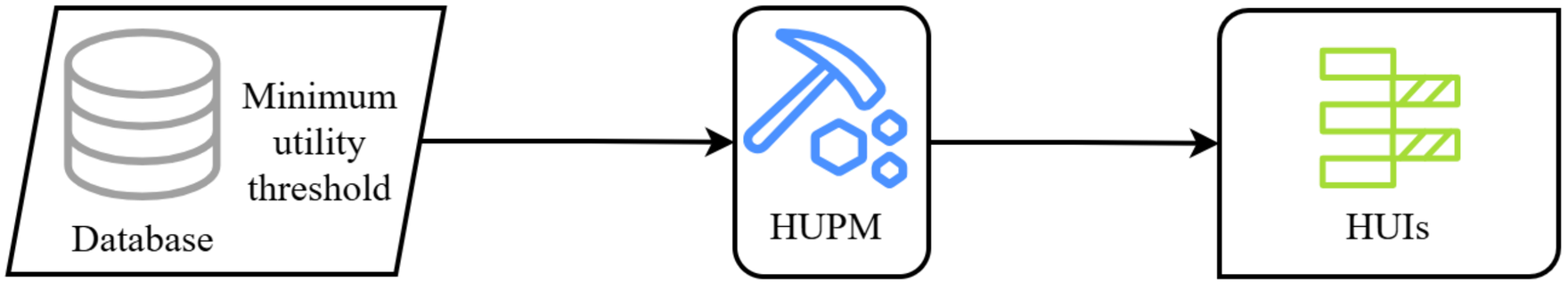
1. Giới thiệu bài toán

1.1. Tổng quan

1.2. Thách thức

1.1. Tổng quan

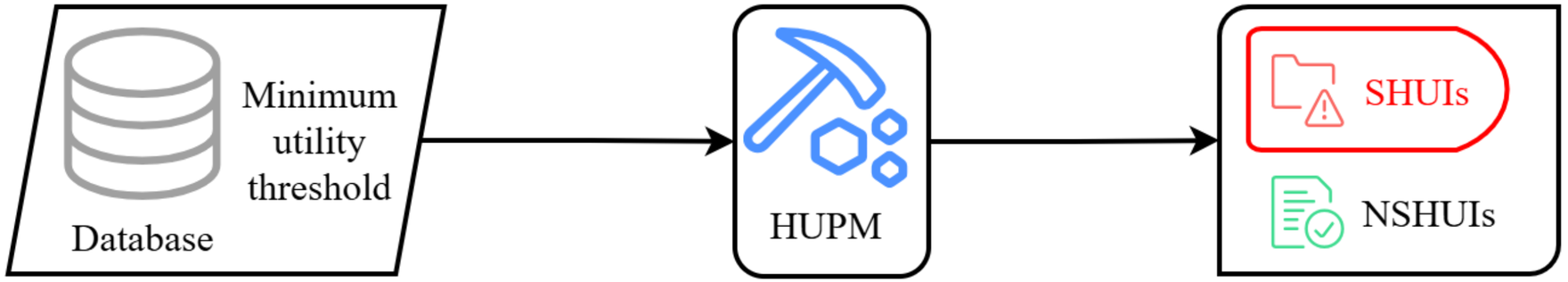
Khai thác mẫu hữu ích (High-Utility Pattern Mining - HUPM) giúp phát hiện các tri thức có giá trị và các mối tương quan thú vị tiềm ẩn trong dữ liệu.



Quy trình của HUPM

1.1. Tổng quan

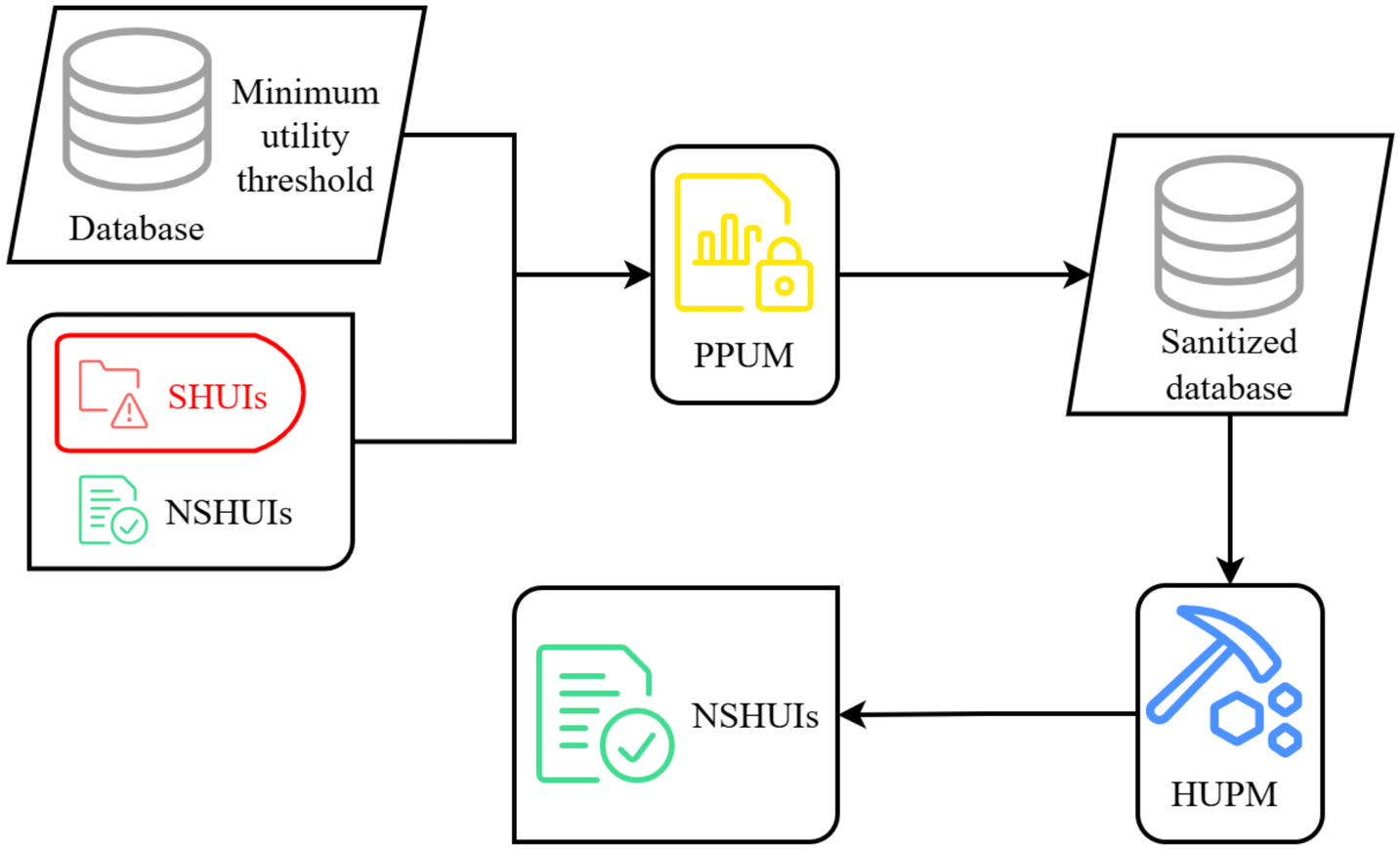
Tuy nhiên, khi triển khai HUPM, rủi ro về tiết lộ thông tin nhạy cảm luôn tồn tại, gây ra những lo ngại nghiêm trọng về bảo vệ tính riêng tư.



Quy trình của HUPM

1.1. Tổng quan

Để giải quyết vấn đề này, lĩnh vực bảo vệ tính riêng tư trong khai thác mẫu hữu ích (Privacy-Preserving Utility Mining- PPUM) đã ra đời, với mục tiêu ẩn các thông tin nhạy cảm trước khi dữ liệu được chia sẻ và khai thác.

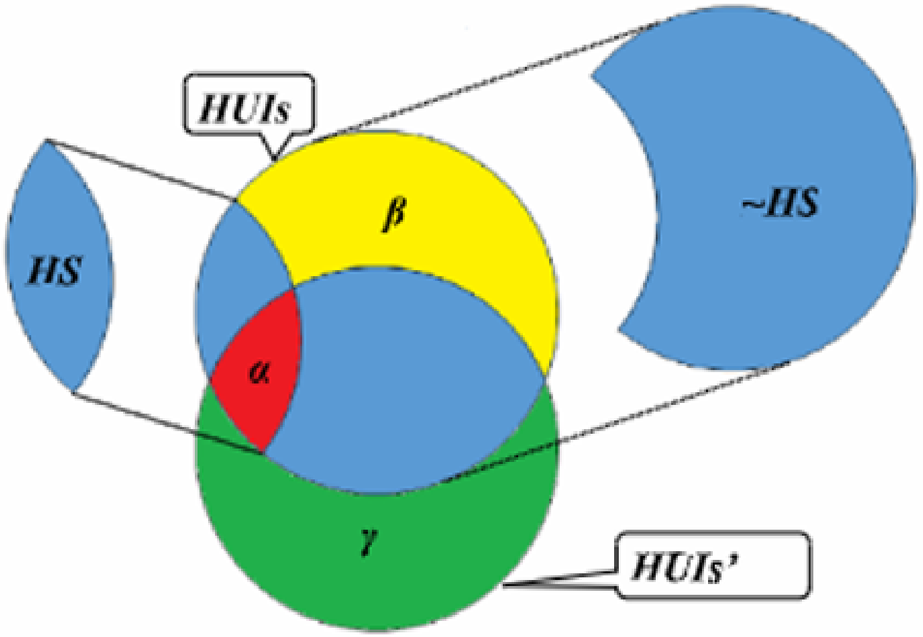


Quy trình của PPUM

1.2. Thách thức

Thách thức của PPUM là tối thiểu hóa cùng lúc cả ba tác dụng phụ sau:

- Số lượng các SHUI không bị ẩn.
- Số lượng các NSHUI bị ẩn nhầm.
- Số lượng các HUI nhân tạo được tạo mới.



Tác dụng phụ trong PPUM [1]

[1] Lin, J.-W., Hong, T.-P., Fournier-Viger, P., Liu, Q., Wong, J.-W., and Zhan, J., "Efficient hiding of confidential high-utility itemsets with minimal side effects," J. Exp. Theor. Artif. Intell., vol. 132, p. 103 360, 2017.

2. Phương pháp giải quyết

2.1. Các công trình liên quan

2.2. Thuật toán đề xuất SO2DI

2.3. Thực nghiệm và đánh giá

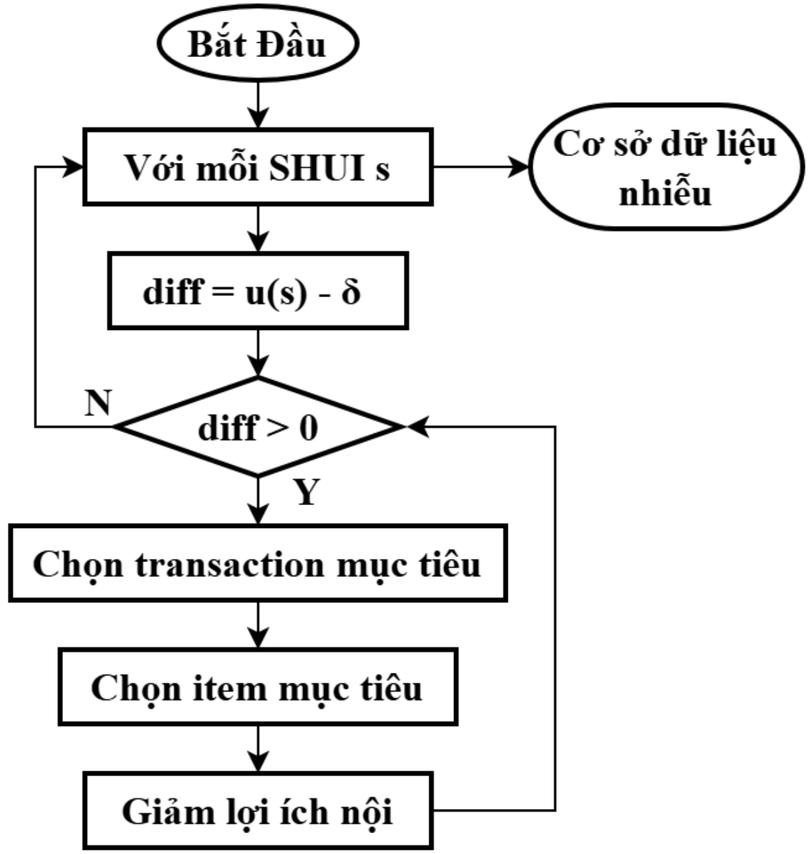
2.1. Các công trình liên quan

MSU-MAU và MSU-MIU

- Transaction mục tiêu: transaction mà s có lợi ích lớn nhất.
- Item mục tiêu:
  - MSU-MAU: item có lợi ích lớn nhất trong transaction mục tiêu.
  - MSU-MIU: item có lợi ích nhỏ nhất trong transaction mục tiêu.

Nhận xét:

- Ẩn lần lượt từng SHUI.
- Luôn đảm bảo ẩn thành công SHUI.
- Không tạo ra HUI nhân tạo.



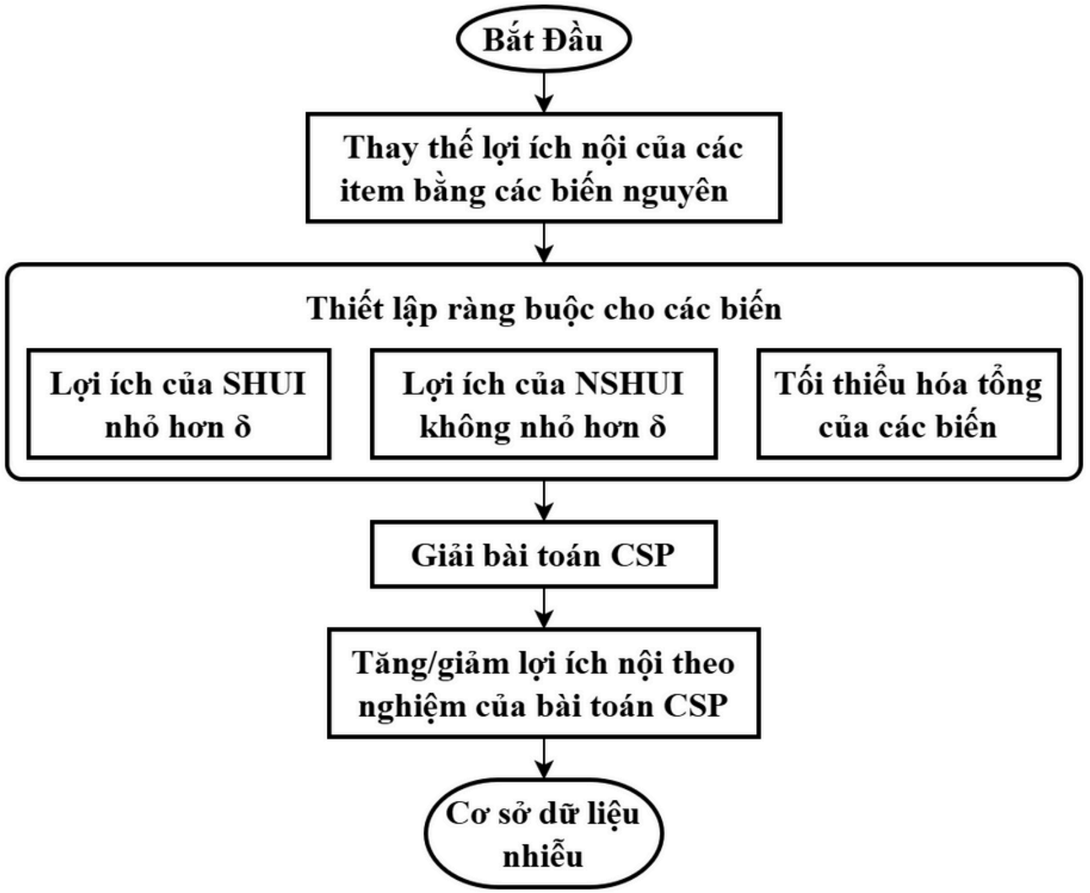
Lưu đồ thuật toán MSU-MAU/MSU-MIU

2.1. Các công trình liên quan

FILP

Nhận xét:

- Ẩn cùng lúc tất cả các SHUI.
- Không đảm bảo ẩn thành công SHUI.
- Có thể tạo ra HUI nhân tạo.



Lưu đồ thuật toán FILP

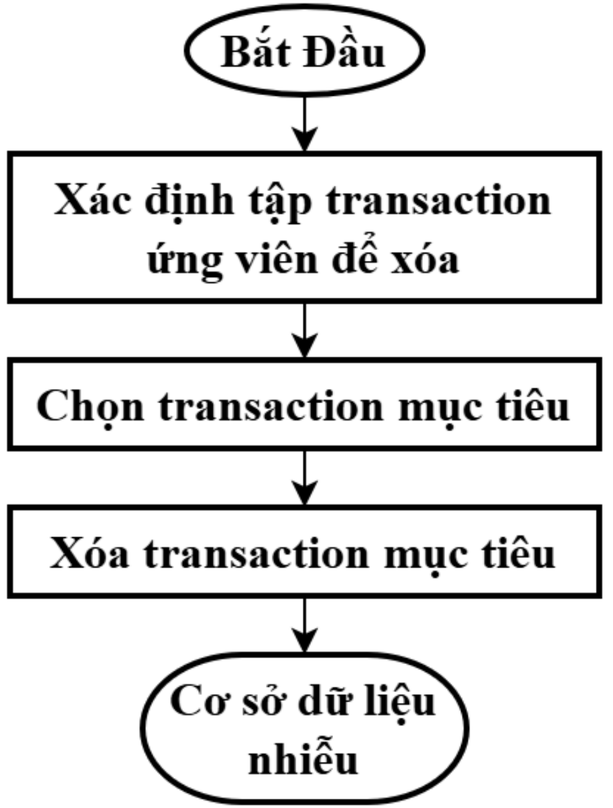
2.1. Các công trình liên quan

PPUMGAT

- Transaction ứng viên: transaction chứa ít nhất một SHUI.
- Transaction mục tiêu: transaction được chọn từ tập transaction ứng viên dựa vào thuật toán GA.

Nhận xét:

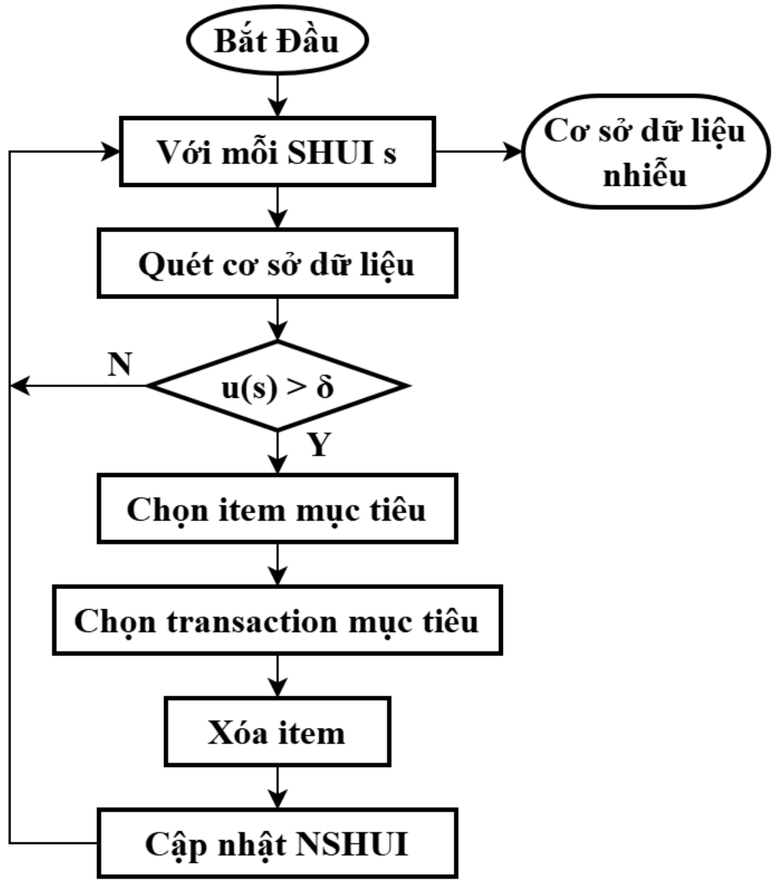
- Ẩn cùng lúc tất cả các SHUI.
- Không đảm bảo ẩn thành công SHUI.
- Không tạo ra HUI nhân tạo.



Lưu đồ thuật toán PPUMGAT

2.2. Thuật toán đề xuất SO2DI

- SO2DI ẩn lần lượt từng SHUI.
- Với mỗi SHUI s, thuật toán thực hiện năm bước chính:
    1. Quét cơ sở dữ liệu.
    2. Chọn item mục tiêu.
    3. Chọn transaction mục tiêu.
    4. Xóa item.
    5. Cập nhật NSHUI.



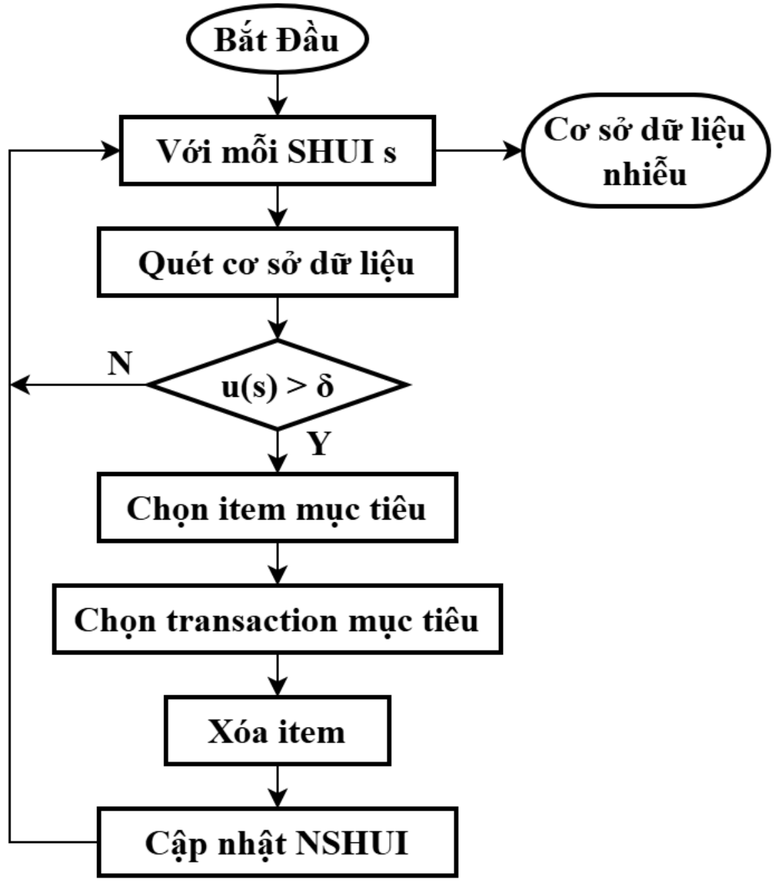
Lưu đồ thuật toán SO2DI

2.2. Thuật toán đề xuất SO2DI

1. Quét cơ sở dữ liệu.

- Xác định các transaction nhạy cảm.
- Tính lợi ích của s trong mỗi transaction nhạy cảm.
- Tính lại lợi ích của s trong cơ sở dữ liệu.

Nếu lợi ích của s lớn hơn δ, thuật toán sẽ thực hiện bước 2. Ngược lại, thuật toán sẽ bỏ qua s và chuyển sang xử lý SHUI tiếp theo.

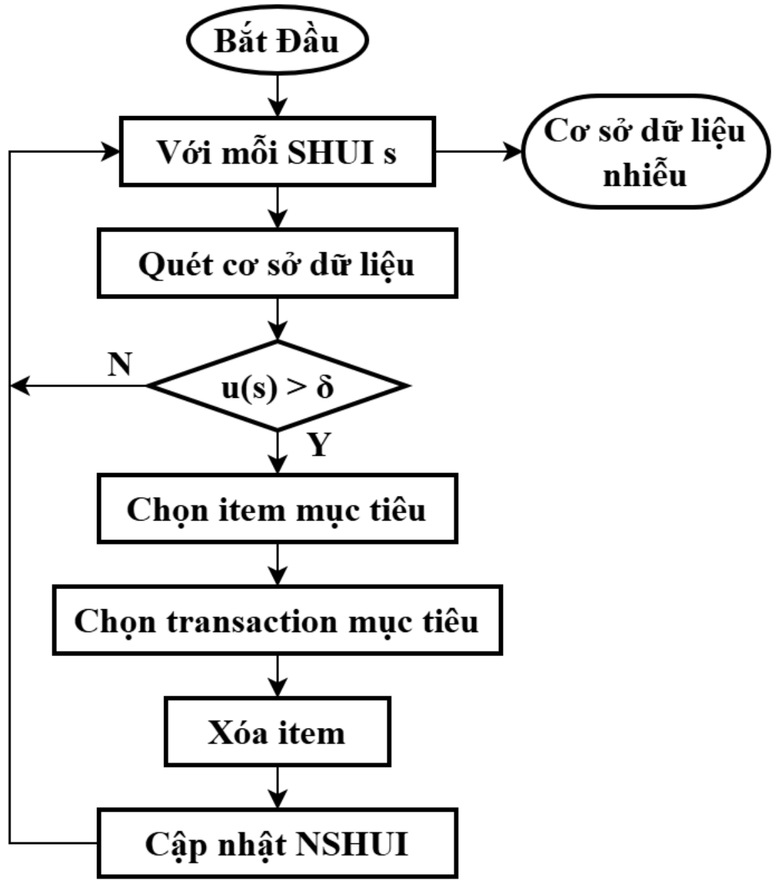


Lưu đồ thuật toán SO2DI

2.2. Thuật toán đề xuất SO2DI

2. Chọn item mục tiêu

- Item mục tiêu: Item có tần suất thấp nhất trong các NSHUI.
- Kết quả: Item mục tiêu iv và tập R chứa các NSHUI liên quan.



Lưu đồ thuật toán SO2DI

2.2. Thuật toán đề xuất SO2DI

3. Chọn transaction mục tiêu

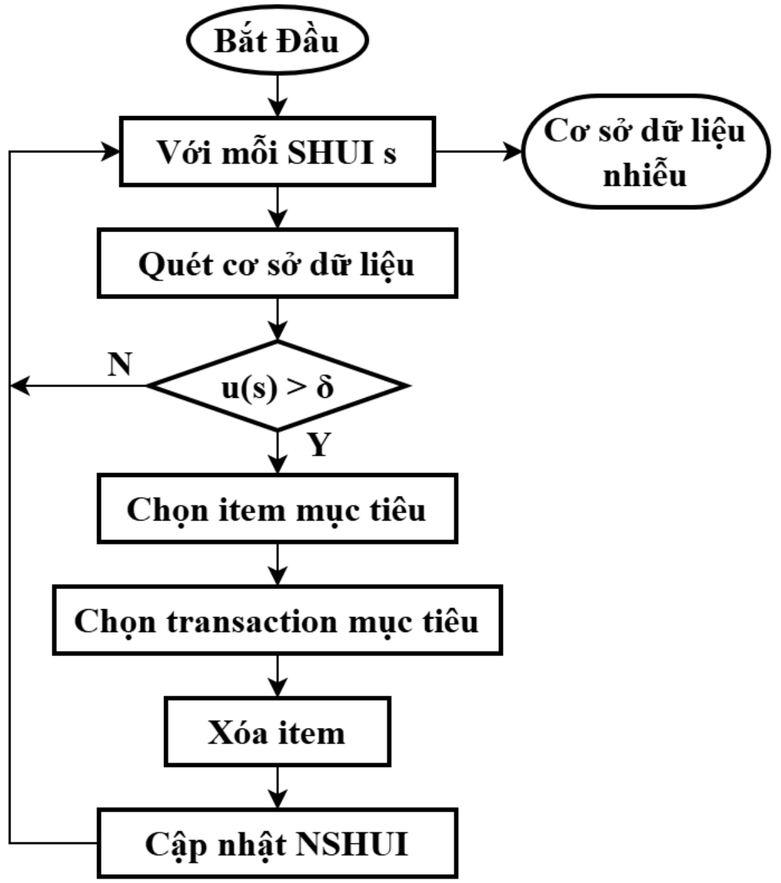
- Mô hình hóa bài toán:

$$V_{TID}^{*} = \text{argmin}_{V_{TID} \subseteq S_{TID}} f\left(V_{TID}\right)$$

- Hàm mục tiêu:

$$f_{SO2DI} = \mathbb{I}(u'(s) \geq \delta) \times (|R| + 1) + \beta$$

- Giải bài toán bằng tối ưu hóa ngẫu nhiên: GA, PSO, ACO

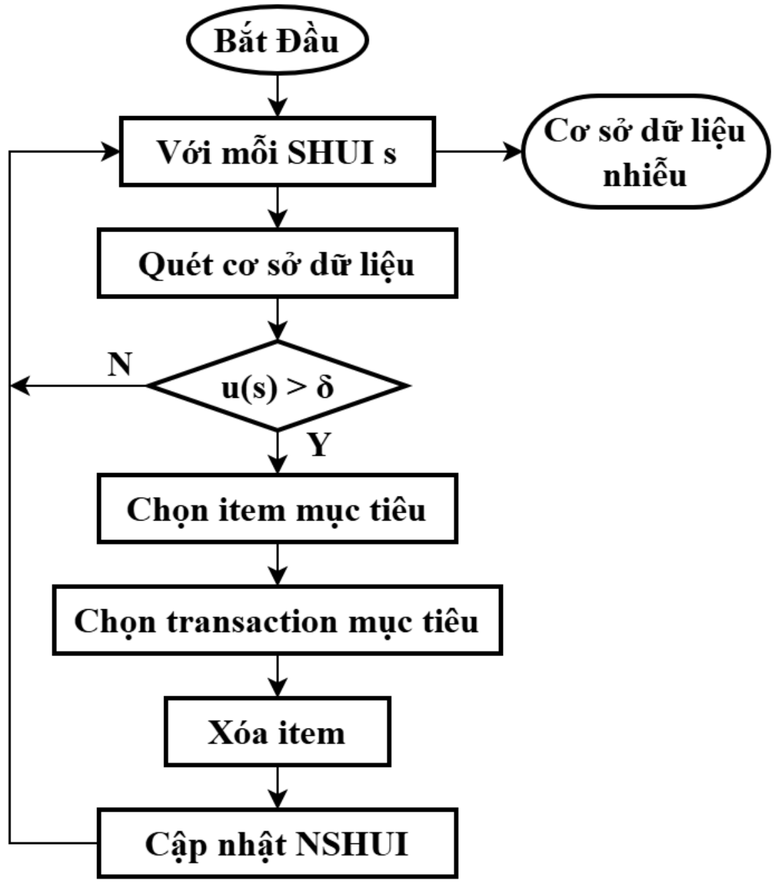


Lưu đồ thuật toán SO2DI

2.2. Thuật toán đề xuất SO2DI

4. Xóa item

- Xóa item mục tiêu iv khỏi các transaction mục tiêu.
- Kết quả: Lợi ích của s giảm xuống dưới ngưỡng lợi ích tối thiểu.

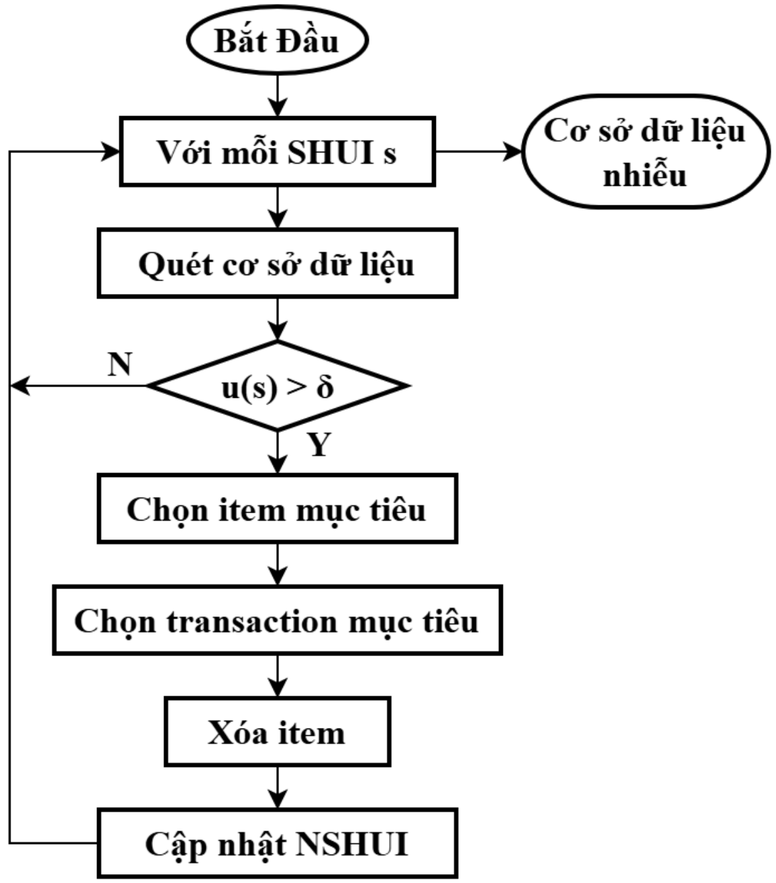


Lưu đồ thuật toán SO2DI

2.2. Thuật toán đề xuất SO2DI

5. Cập nhật NSHUI

- Tính lợi ích mới cho các NSHUI trong tập R.
- Loại bỏ các NSHUI nếu lợi ích mới của nó nhỏ hơn ngưỡng lợi ích tối thiểu.

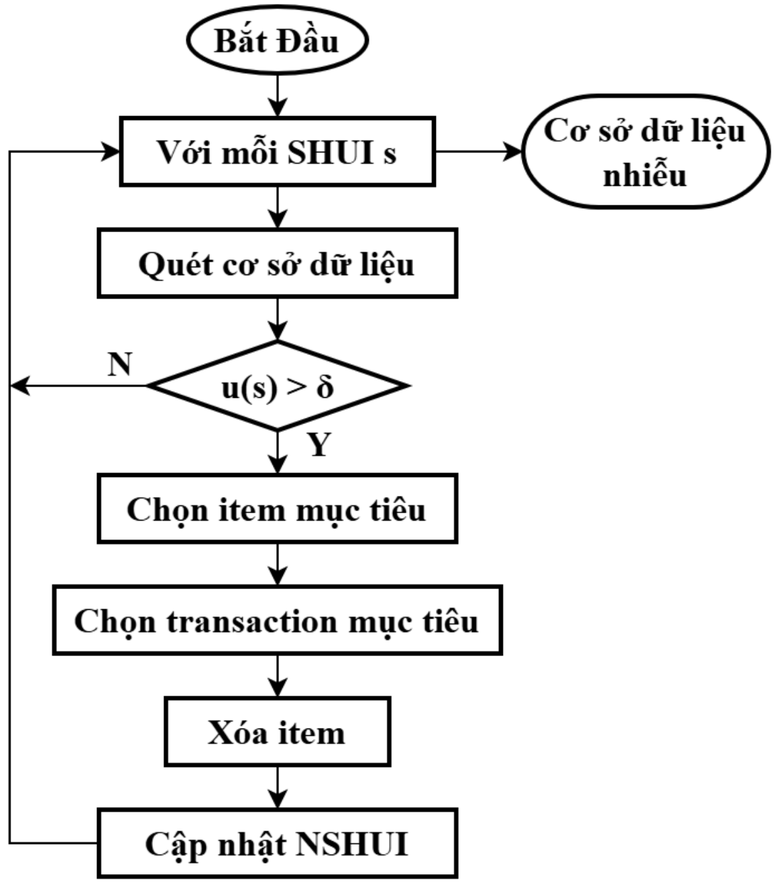


Lưu đồ thuật toán SO2DI

2.2. Thuật toán đề xuất SO2DI

Nhận xét:

- Ẩn lần lượt từng SHUI.
- Không đảm bảo ẩn thành công SHUI.
- Không tạo ra HUI nhân tạo.

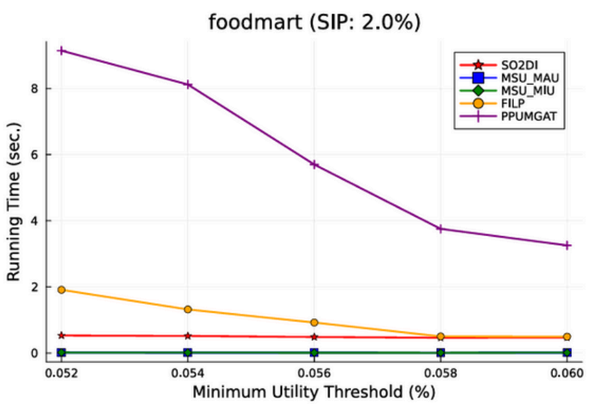


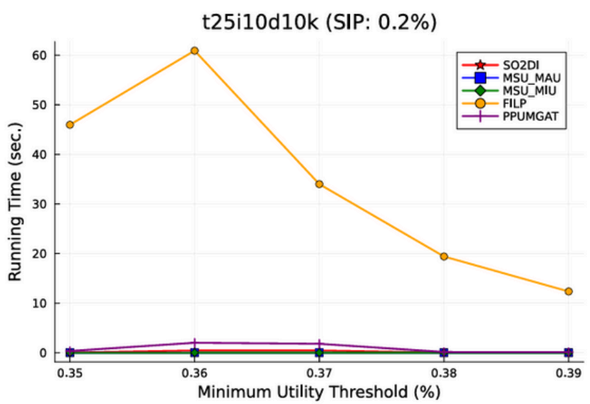
Lưu đồ thuật toán SO2DI

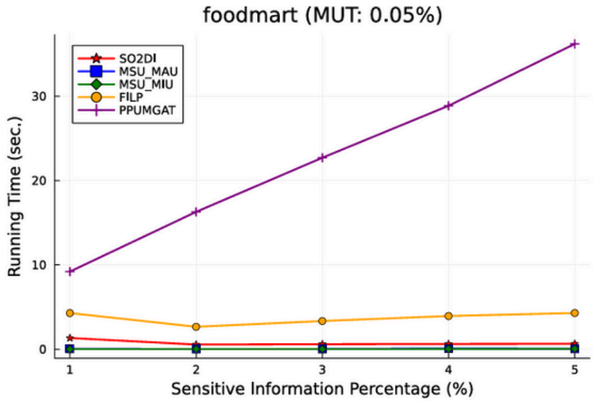
2.3. Thực nghiệm và đánh giá

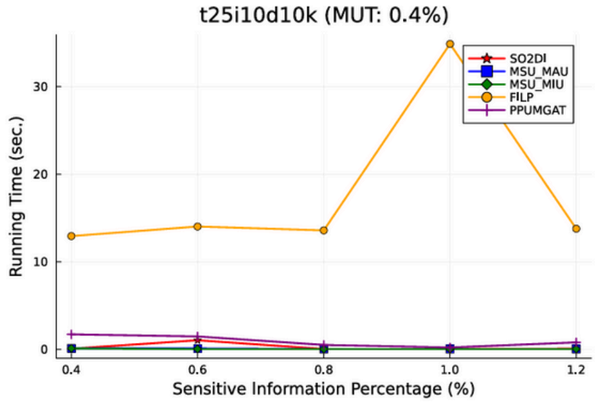
- So sánh với các thuật toán: MSU-MAU, MSU-MIU, FILP và PPUMGAT.
- Tập dữ liệu: foodmart, t25i10d10k, mushrooms, chess, retail, t20i6d100k, pumsb và connect.
- Tiêu chí đánh giá: thời gian thực thi và tác dụng phụ.

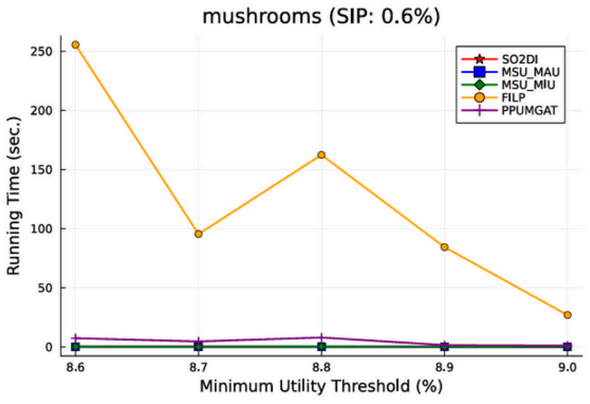
Thời gian thực thi - Tập dữ liệu nhỏ

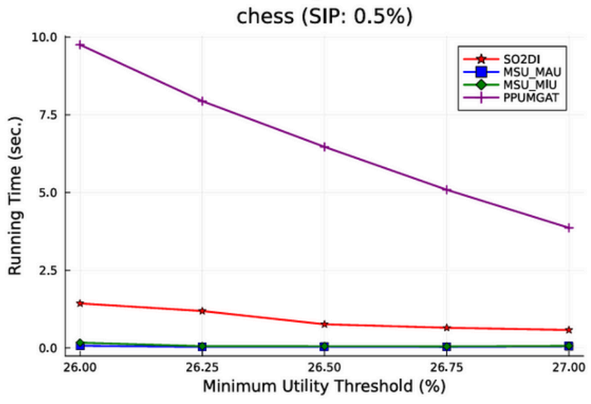


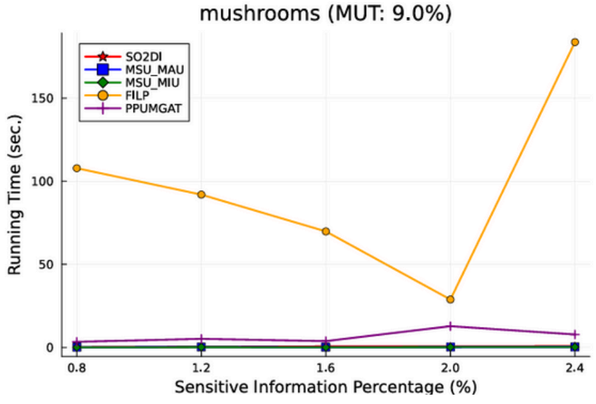


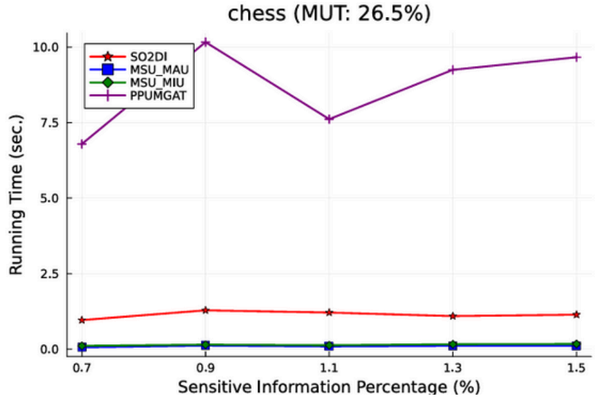








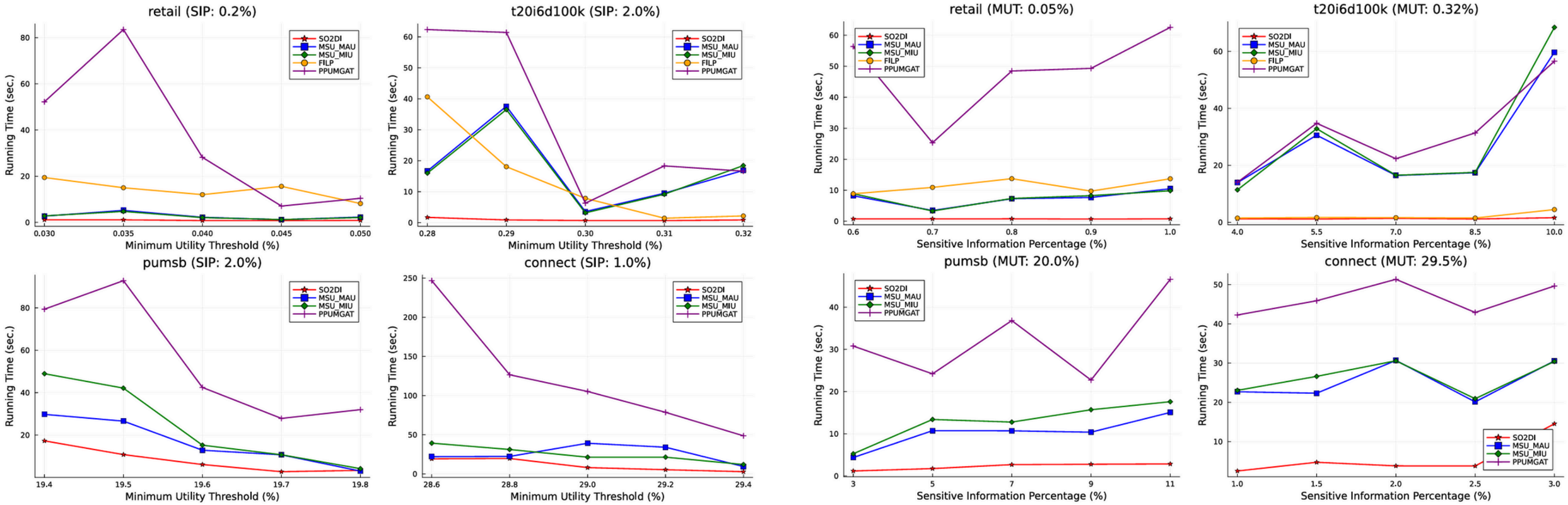




Geometric Mean Speedup:

FILP: 0.02; PPUMGAT: 0.09; SO2DI: 1.0; MSU-MIU: 7.67; MSU-MAU: 8.24

Thời gian thực thi - Tập dữ liệu lớn



Geometric Mean Speedup:

PPUMGAT: 0.05; FILP: 0.14; MSU-MIU: 0.17; MSU-MAU: 0.18; SO2DI: 1.0

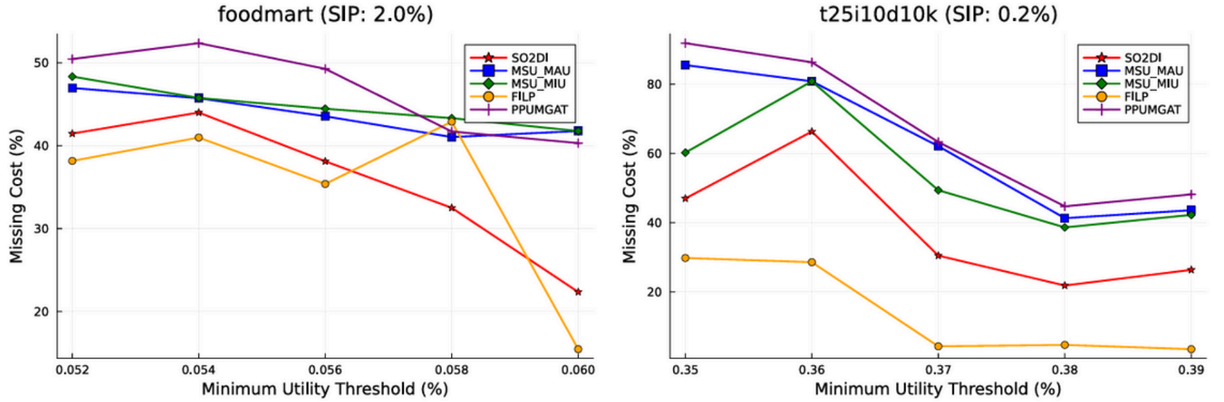
Hiding Failure

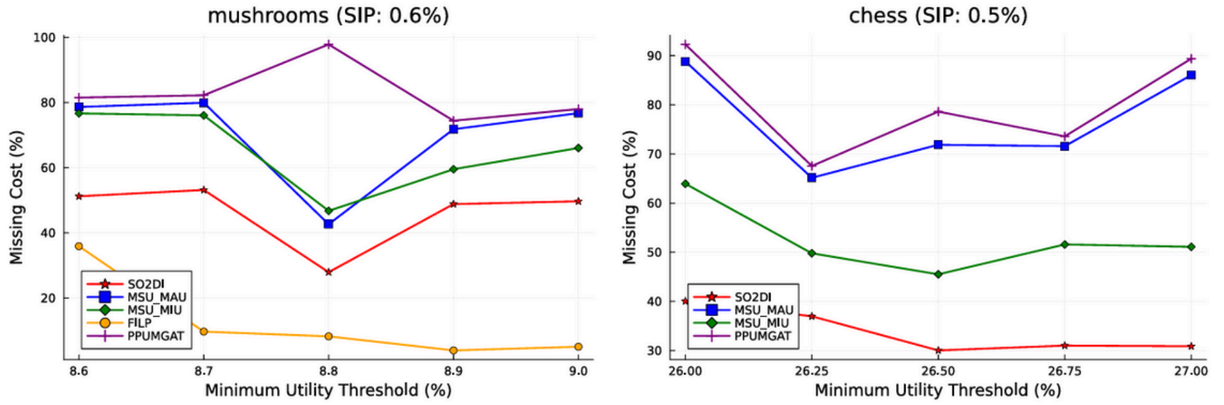
MSU-MAU, MSU-MIU, SO2DI và PPUMGAT đều ẩn giấu thành công tất cả SHUI trên cả tám tập dữ liệu thực nghiệm.

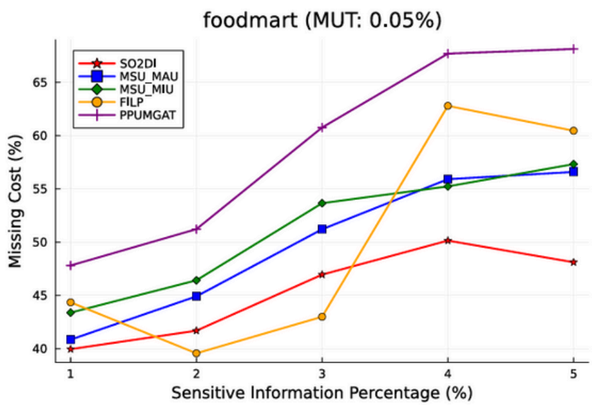
FILP thất bại trong việc ẩn hoàn toàn SHUI trên các tập dữ liệu foodmart, t20i6d100k và t25i10d10k.

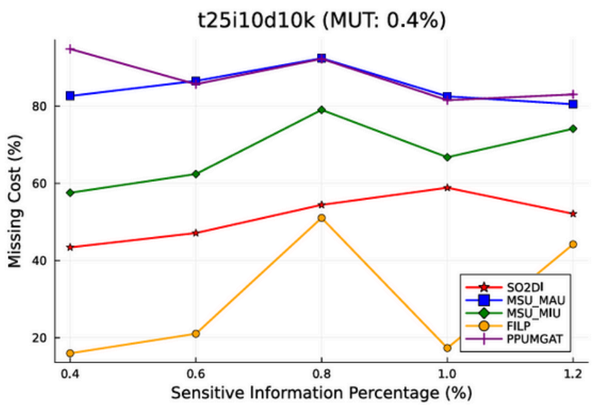
Nhận xét: FILP < MSU-MAU = MSU-MIU = SO2DI = PPUMGAT

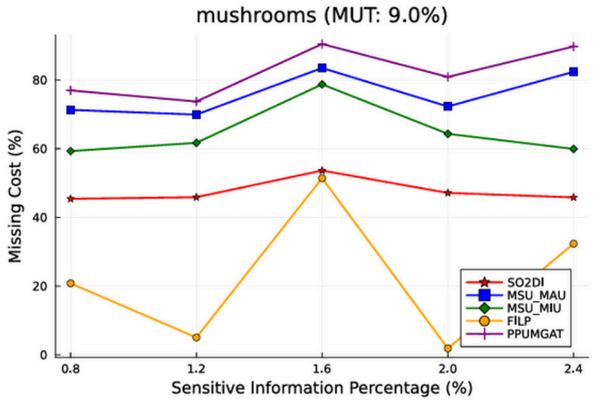
Missing Cost - Tập dữ liệu nhỏ

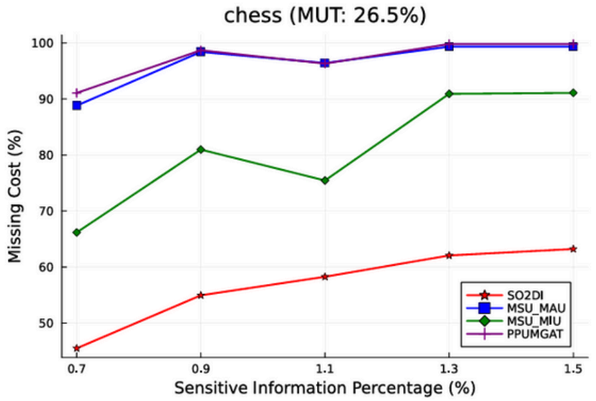








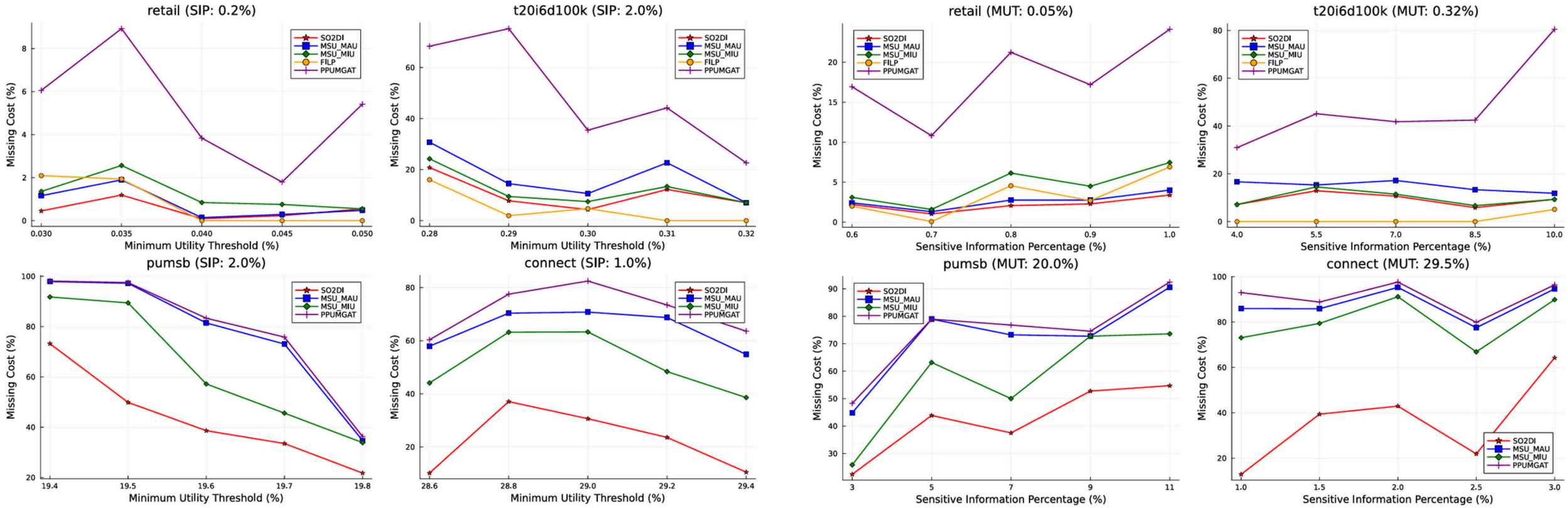




Mean Absolute Difference:

PPUMGAT: -31%; MSU-MAU: -26%; MSU-MIU: -16%; SO2DI: 0%; FILP: 16%

Missing Cost - Tập dữ liệu lớn



Mean Absolute Difference:

PPUMGAT: -34%; MSU-MAU: -21%; MSU-MIU: -14%; SO2DI: 0%; FILP: 3%

Artificial Cost

MSU-MAU, MSU-MIU, SO2DI và PPUMGAT đều duy trì artificial cost ở mức 0% trên tất cả tám tập dữ liệu thử nghiệm.

FILP sinh ra một lượng lớn HUI nhân tạo (có thể lên đến hơn 4000%) trên tất cả các tập dữ liệu mà nó chạy thành công.

Nhận xét: FILP < MSU-MAU = MSU-MIU = SO2DI = PPUMGAT

3. Kết luận

- Về thời gian thực thi, SO2DI chỉ kém MSU-MAU và MSU-MIU trên các tập dữ liệu nhỏ.
- SO2DI có khả năng che giấu thông tin nhạy cảm tốt khi đạt Hiding Failure bằng 0% trên mọi thử nghiệm.
- Về Missing Cost, SO2DI chỉ kém FILP trên các tập dữ liệu nhỏ và thưa.
- SO2DI không tạo ra HUI nhân tạo nên Artificial Cost luôn bằng 0%.

→ SO2DI có hiệu suất vượt trội so với các phương pháp khác, đặc biệt trên các tập dữ liệu lớn và dày.

4. Hướng phát triển

- Tối ưu hóa thứ tự ẩn SHUI.
- Cải tiến cách chọn item mục tiêu.

XIN CẢM ƠN
QUÝ THẦY CÔ
ĐÃ LẮNG NGHE