



fit@hcmus

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

ĐỀ CƯƠNG KHÓA LUẬN TỐT NGHIỆP

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

(Privacy-preserving in utility mining with stochastic optimization)

1 THÔNG TIN CHUNG

Người hướng dẫn:

– ThS. Nguyễn Ngọc Đức (Khoa Công nghệ Thông tin)

[Nhóm] Sinh viên thực hiện:

1. Trần Khắc Bình (MSSV: 20120437)

Loại đề tài: Nghiên cứu

Thời gian thực hiện: Từ 9/2024 đến 3/2025

2 NỘI DUNG THỰC HIỆN

2.1 Giới thiệu về đề tài

Trong thời đại dữ liệu lớn, khám phá tri thức từ cơ sở dữ liệu được ứng dụng thực tế trong nhiều lĩnh vực khác nhau. Một trong những kỹ thuật nổi bật trong lĩnh vực này là khai thác mẫu hữu ích (High-Utility Pattern Mining - HUPM), cho phép phát hiện các tri thức hữu ích và các mối tương quan thú vị tiềm ẩn trong dữ liệu. Tuy nhiên, quá trình khai thác này đang đối mặt với một thách thức lớn đó là bảo vệ tính riêng tư, khi các kỹ thuật HUPM có thể vô tình tiết lộ những thông tin nhạy cảm. 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 chia sẻ và khai thác dữ liệu.

PPUM là quá trình che giấu các itemset lợi ích cao nhạy cảm (Sensitive High-Utility Itemsets - SHUI) bằng cách làm nhiễu cơ sở dữ liệu, đồng thời duy trì chất lượng của các mẫu hữu ích khác. Tuy nhiên, việc này không hề đơn giản, vì việc làm nhiễu không hiệu quả có thể phá vỡ cấu trúc dữ liệu ban đầu, dẫn đến các tác dụng phụ không mong muốn như che giấu các mẫu không nhạy cảm (Non-Sensitive High-Utility Itemsets - NSHUI) hoặc sinh ra các mẫu nhân tạo không tồn tại trong dữ liệu gốc. Mặc dù các giải pháp PPUM hiện tại đã đạt được thành công trong việc ẩn thông tin nhạy cảm, chúng vẫn gặp hạn chế về hiệu suất xử lý và có thể gây tác động tiêu cực lên cơ sở dữ liệu.

Trong khi đó, các kỹ thuật tối ưu hóa ngẫu nhiên (Stochastic Optimization) đã chứng minh được hiệu quả vượt trội trong việc giải quyết các bài toán tối ưu phức tạp với thời gian tính toán giới hạn. Xuất phát từ tiềm năng này, khóa luận đề xuất một hướng tiếp cận mới dựa trên tối ưu hóa ngẫu nhiên nhằm khắc phục những hạn chế của các phương pháp PPUM hiện tại, từ đó nâng cao hiệu quả bảo vệ tính riêng tư và tối ưu hóa giá trị khai thác từ dữ liệu.

2.2 Mục tiêu đề tài

Trong kỷ nguyên số, dữ liệu được xem như một tài sản quý giá, mang lại lợi ích to lớn cho các tổ chức và doanh nghiệp thông qua việc khai thác các mẫu dữ liệu có lợi ích cao. Những thông tin này không chỉ giúp đưa ra các quyết định kinh doanh sáng suốt mà còn hỗ trợ điều chỉnh chiến lược một cách hiệu quả. Tuy nhiên, bên cạnh những cơ hội vượt trội, hoạt động khai thác dữ liệu cũng đặt ra thách thức lớn về bảo mật thông tin, khi các thông tin nhạy cảm có nguy cơ bị tiết lộ ngoài ý muốn, dẫn đến vi phạm quyền riêng tư của cá nhân hoặc tổ chức. Thực tế, các giải pháp hiện nay trong lĩnh vực bảo vệ tính riêng tư trong khai thác mẫu hữu ích vẫn chưa đạt được sự cân bằng giữa việc bảo vệ tính riêng tư và duy trì chất lượng của các mẫu khai thác. Điều này tạo ra động lực mạnh mẽ để nghiên cứu và phát triển các phương pháp tiên tiến hơn, nhằm giải quyết vấn đề này một cách hiệu quả. Qua đó, không chỉ giảm thiểu rủi ro mà còn nâng cao giá trị ứng dụng thực tiễn của dữ liệu.

2.3 Phạm vi của đề tài

Nội dung nghiên cứu chính của khóa luận tập trung vào việc bảo vệ tính riêng tư trong khai thác mẫu hữu ích. Các đối tượng nghiên cứu trong khóa luận bao gồm: các kỹ thuật khai thác mẫu lợi ích cao trên cơ sở dữ liệu transaction, các phương pháp bảo vệ tính riêng tư trong khai thác mẫu hữu ích, và các thuật toán tối ưu hóa ngẫu nhiên. Trong khóa luận, một thuật toán mới dựa trên phương pháp tối ưu hóa ngẫu nhiên nhằm bảo vệ tính riêng tư trong khai thác mẫu hữu ích sẽ được đề xuất. Các bộ dữ liệu giao dịch thực tế hoặc tổng hợp, như chess, connect, foodmart, mushrooms, pumsb, retail, t20i6d100k, t25i10d10k, ... sẽ được sử dụng để so sánh và đánh giá hiệu quả của thuật toán đề xuất so với các thuật toán hiện có.

2.4 Cách tiếp cận dự kiến

Hai thuật toán PPUM được giới thiệu lần đầu tiên bởi Yeh và Hsu (2010) [1], có tên là HHUIF và MSCIF. Trong cả hai thuật toán, số lượng của các item nạn nhân được giảm cho đến khi giá trị lợi ích của các SHUI giảm xuống dưới ngưỡng tiện ích tối thiểu. Sự khác biệt giữa hai thuật toán nằm ở chiến lược chọn item nạn nhân: HHUIF chọn item có lợi ích cao nhất, trong khi MSCIF chọn item xuất hiện nhiều nhất trong các itemset nhạy cảm. Về kết quả, cả hai thuật toán đều có thời gian thực thi lớn trên cơ sở dữ liệu lớn do phải quét dữ liệu nhiều lần và gây mất mát một lượng lớn các itemset hữu ích không nhạy cảm.

Nhằm cải thiện hiệu suất của HHUIF, Yun và Kim (2015) [2] đã đề xuất thuật toán FPUTT. Thuật toán này sử dụng cấu trúc cây để giảm số lần quét cơ sở dữ liệu xuống còn ba lần, giúp tăng tốc đáng kể so với HHUIF. Tuy nhiên, do vẫn giữ nguyên chiến lược chọn item nạn nhân của HHUIF, FPUTT tiếp tục gặp phải các tác dụng phụ tương tự. Bên cạnh đó, cây FPUTT đòi hỏi không gian bộ nhớ lớn và chi phí tính toán cao.

Lin và cộng sự (2016) [3] đã giới thiệu hai thuật toán mới là MSU-MAU và MSU-MIU để giảm tác dụng phụ do ẩn các SHUI. Sự khác biệt giữa hai thuật toán này cũng nằm ở chiến lược chọn item nạn nhân: MSU-MAU chọn item có lợi ích lớn nhất trong transaction nạn nhân, trong khi MSU-MIU chọn item có lợi ích nhỏ nhất. Kết quả thực nghiệm cho thấy cả hai thuật toán không chỉ giảm tác dụng phụ tốt hơn HHUIF và MSCIF mà còn cải thiện tốc độ nhờ sử dụng bảng chỉ mục để lưu trữ thông tin các transaction liên quan đến từng itemset nhạy cảm.

Ngoài các thuật toán dựa vào phương pháp heuristic ở trên, Li và cộng sự (2019) [4] lần đầu tiên áp dụng quy hoạch số nguyên tuyến tính (Integer Linear Programming - ILP) để giải quyết vấn đề PPUM, với thuật toán là PPUM-ILP. Ý tưởng là mô hình hóa quá trình ẩn các SHUI thành một bài toán thỏa mãn ràng buộc (Constraints Satisfaction Problem - CSP) và sử dụng cơ chế làm lỏng

ràng buộc để tìm lời giải. Kết quả cho thấy, PPUM-ILP giúp giảm đáng kể các tác dụng phụ nhưng thời gian thực thi không ổn định, đặc biệt có thể kéo dài trên các tập dữ liệu lớn.

Để cải thiện hiệu suất của PPUM-ILP, Nguyen và cộng sự (2022) [5] đã đề xuất thuật toán FILP, sử dụng cấu trúc dữ liệu băm để tăng tốc quá trình lọc các itemset và mô hình hóa bài toán. Nhờ đó, FILP đạt được hiệu suất cao hơn so với PPUM-ILP.

Tiếp nối, Nguyen và cộng sự (2023) [6] đã phát triển thuật toán G-ILP nhằm giảm tác dụng phụ và tối ưu thời gian thực thi trong quá trình làm nhiễu. G-ILP áp dụng lập trình song song trên GPU để rút ngắn thời gian tiền xử lý và xây dựng một mô hình thỏa mãn ràng buộc khác hiệu quả hơn. Kết quả thực nghiệm chứng minh rằng G-ILP vượt trội cả về thời gian thực thi và khả năng giảm thiểu tác dụng phụ, đặc biệt trên các tập dữ liệu lớn và thưa, so với tất cả các thuật toán trước đó.

Tất cả các thuật toán được thảo luận ở trên đều áp dụng mô hình làm nhiễu dựa trên item để ẩn SHUI, trong đó số lượng của các item nạn nhân được sửa đổi. Ngược lại, một số nghiên cứu khác tuân theo mô hình nhiễu dựa trên transaction cho nhiệm vụ PPUM. Tiêu biểu trong số đó là hai thuật toán PPUMGA+insert [7] và PPUMGAT [8] do Lin và cộng sự đề xuất, cả hai đều dựa trên thuật toán di truyền (Genetic Algorithm - GA).

PPUMGA+insert là thuật toán đầu tiên theo mô hình nhiễu dựa trên transaction, được Lin và cộng sự giới thiệu vào năm 2014. Thuật toán hoạt động bằng cách chèn thêm các transaction mới vào cơ sở dữ liệu nhằm tăng tổng lợi ích của cơ sở dữ liệu. Khi tổng lợi ích tăng, ngưỡng lợi ích tối thiểu cũng được nâng lên, từ đó làm cho các SHUI không còn được coi là lợi ích cao. Tuy nhiên, việc thêm các transaction mới có thể làm thay đổi cấu trúc dữ liệu và dẫn đến sự xuất hiện của các itemset hữu ích nhân tạo trong cơ sở dữ liệu sau khi làm nhiễu.

Tiếp nối, vào năm 2017, Lin và cộng sự đã phát triển một thuật toán khác có tên

PPUMGAT. Thuật toán này sử dụng cách tiếp cận đối lập so với PPUMGA+insert: thay vì thêm transaction, PPUMGAT thực hiện xóa các transaction hiện có trong cơ sở dữ liệu nhằm giảm lợi ích của SHUI. PPUMGAT sử dụng hàm mục tiêu để đánh giá mức độ phù hợp của các giải pháp ứng cử viên, từ đó xác định các transaction nhạy cảm cần xóa. Tuy nhiên, cũng giống như PPUMGA+insert, PPUMGAT phải đối mặt với vấn đề làm thay đổi số lượng transaction trong cơ sở dữ liệu, có thể dẫn đến mất mát thông tin và tạo ra các itemset hữu ích nhân tạo.

Trong khóa luận này, thuật toán đề xuất áp dụng mô hình nhiều dựa trên item để ẩn các SHUI. Tương tự như HHUIF, thuật toán tiến hành xử lý lần lượt từng SHUI để đảm bảo rằng lợi ích của chúng giảm xuống dưới ngưỡng lợi ích tối thiểu. Tuy nhiên, thay vì sửa đổi số lượng các item, thuật toán lựa chọn xóa item trong các transaction nạn nhân, giúp rút ngắn thời gian xử lý và tăng hiệu quả làm nhiều. Item nạn nhân được lựa chọn dựa trên tần suất xuất hiện ít nhất trong các NSHUI. Cách tiếp cận này đảm bảo rằng việc xóa item sẽ ít ảnh hưởng nhất đến các NSHUI, qua đó giảm thiểu nguy cơ làm ẩn các NSHUI sau khi thực hiện quá trình làm nhiều. Hàm mục tiêu được thiết kế để đánh giá mức độ phù hợp của các giải pháp ứng cử viên trong việc xác định các transaction nạn nhân. Hàm này cân nhắc đồng thời hai mục tiêu: đảm bảo SHUI bị ẩn và giảm thiểu số lượng NSHUI bị ẩn nhầm. Phương pháp tối ưu hóa ngẫu nhiên được sử dụng để tối ưu hóa hàm mục tiêu, cho phép tìm kiếm các giải pháp tốt nhất trong không gian tìm kiếm rộng lớn với thời gian xử lý giới hạn.

2.5 Kết quả dự kiến của đề tài

Dưới đây là các kết quả cụ thể mà khóa luận hướng đến đạt được:

- Tìm hiểu vấn đề khai thác mẫu hữu ích.
- Nghiên cứu vấn đề bảo vệ tính riêng tư trong khai thác mẫu hữu ích.

- Nghiên cứu các phương pháp và thuật toán tối ưu hóa ngẫu nhiên.
- Đề xuất một phương pháp hoặc hướng tiếp cận mới nhằm bảo vệ tính riêng tư trong khai thác mẫu hữu ích.
- Cài đặt thuật toán khai thác mẫu hữu ích EFIM.
- Cài đặt các thuật toán bảo vệ tính riêng tư trong khai thác mẫu hữu ích: MSU-MAU, MSU-MIU, FILP và PPUMGAT.
- Xây dựng chương trình thử nghiệm và đánh giá phương pháp đề xuất dựa trên kết quả thực nghiệm.


2.6 Kế hoạch thực hiện

Thời gian	Công việc	Kết quả dự kiến
9/2024	Nghiên cứu tổng quan, tìm hiểu các phương pháp hiện có	Xác định hướng tiếp cận
10/2024	Thiết kế thuật toán	Đề xuất thuật toán chi tiết
11/2024	- Cài đặt thuật toán - Thử nghiệm ban đầu	- Phiên bản đầu tiên - Kết quả thử nghiệm sơ bộ
12/2024	- Tối ưu hóa thuật toán - Thực nghiệm trên các bộ dữ liệu	- Phiên bản cải tiến - Kết quả thực nghiệm chi tiết
1/2025	- Phân tích kết quả - Viết báo cáo	- Phân tích so sánh - Bản thảo báo cáo
2/2025	- Hoàn thiện báo cáo - Chuẩn bị bảo vệ	- Báo cáo hoàn chỉnh - Slides bảo vệ
3/2025	Chỉnh sửa theo góp ý và bảo vệ khóa luận	Báo cáo cuối cùng

Tài liệu

- [1] J.-S. Yeh and P.-C. Hsu, "Hhuif and msicf: Novel algorithms for privacy preserving utility mining," *Expert Syst. Appl.*, vol. 37, no. 7, pp. 4779–4786, 2010.
- [2] U. Yun and J. Kim, "A fast perturbation algorithm using tree structure for privacy preserving utility mining," *Expert Syst. Appl.*, vol. 42, no. 3, pp. 1149–1165, 2015.
- [3] J. C.-W. Lin, T.-Y. Wu, P. Fournier-Viger, G. Lin, J. Zhan, and M. Voznak, "Fast algorithms for hiding sensitive high-utility itemsets in privacy-preserving utility mining," *Eng. Appl. Artif. Intell.*, vol. 55, pp. 269–284, 2016.
- [4] S. Li, M. Nankun, J. Le, and X. Liao, "A novel algorithm for privacy preserving utility mining based on integer linear programming," *Eng. Appl. Artif. Intell.*, vol. 81, pp. 300–312, 2019.
- [5] D. Nguyen and B. Le, "A fast algorithm for privacy-preserving utility mining," *J. Inf. Technol. Commun.*, vol. 2022, no. 1, pp. 12–22, 2022.
- [6] D. Nguyen, M.-T. Tran, and B. Le, "A new algorithm using integer programming relaxation for privacy-preserving in utility mining," *Applied Intelligence*, vol. 2023, pp. Article ID 10489–023–04913–w, 2023.
- [7] C.-W. Lin, T.-P. Hong, J.-W. Wong, G.-C. Lan, and W.-Y. Lin, "A ga-based approach to hide sensitive high utility itemsets," *Sci. World J.*, vol. 2014, pp. 1–12, 2014.
- [8] J.-W. Lin, T.-P. Hong, P. Fournier-Viger, Q. Liu, J.-W. Wong, and J. Zhan, "Efficient hiding of confidential high-utility itemsets with minimal side effects," *J. Exp. Theor. Artif. Intell.*, vol. 132, p. 103360, 2017.

XÁC NHẬN
CỦA NGƯỜI HƯỚNG DẪN
(Ký và ghi rõ họ tên)


Nguyễn Ngọc Đức

TP. Hồ Chí Minh, ngày 20 tháng 2 năm 2025
NHÓM SINH VIÊN THỰC HIỆN
(Ký và ghi rõ họ tên)

Bình
Trần Khắc Bình