**TỔNG LIÊN ĐOÀN LAO ĐỘNG VIỆT NAM**

**TRƯỜNG ĐẠI HỌC TÔN ĐỨC THẮNG**

**KHOA CÔNG NGHỆ THÔNG TIN**



**BÁO CÁO CHUYÊN ĐỀ NGHIÊN CỨU KHOA HỌC**

**WEIGHTED PROBABILISTIC MAXIMAL FREQUENT ITEMSET**

**OVER UNCERTAIN DATABASE**

*Người hướng dẫn*:  **TS. NGUYỄN CHÍ THIỆN**

*Người thực hiện*:  **TRẦN TẤN HƯNG – 52000052**

**THÀNH PHỐ HỒ CHÍ MINH, NĂM 2024**

**TỔNG LIÊN ĐOÀN LAO ĐỘNG VIỆT NAM**

**TRƯỜNG ĐẠI HỌC TÔN ĐỨC THẮNG**

**KHOA CÔNG NGHỆ THÔNG TIN**



**BÁO CÁO CHUYÊN ĐỀ NGHIÊN CỨU KHOA HỌC**

**WEIGHTED PROBABILISTIC MAXIMAL FREQUENT ITEMSET**

**OVER UNCERTAIN DATABASE**

*Người hướng dẫn*:  **TS. NGUYỄN CHÍ THIỆN**

*Người thực hiện*:  **TRẦN TẤN HƯNG – 52000052**

**THÀNH PHỐ HỒ CHÍ MINH, NĂM 2024**

**LỜI CẢM ƠN**

Chúng em xin chân thành cảm ơn khoa Công nghệ thông tin đã tạo điều kiện cho chúng em được tiếp cận và hoàn thành bài báo cáo. Chúng em xin chân thành cảm ơn thầy Nguyễn Chí Thiện đã hướng dẫn hoàn thành bài báo cáo.

Trong quá trình làm bài báo cáo, do kiến thức cũng như kinh nghiệm còn nhiều hạn chế nên bài báo cáo không thể tránh khỏi những thiếu sót, chúng em rất mong nhận được ý kiến đóng góp của thầy để chúng em có thể học hỏi được nhiều kĩ năng, kinh nghiệm và sẽ ngày càng hoàn thiện hơn.

Chúng em xin chân thành cảm ơn!

**BÁO CÁO ĐƯỢC HOÀN THÀNH**

**TẠI TRƯỜNG ĐẠI HỌC TÔN ĐỨC THẮNG**

Tôi xin cam đoan đây là sản phẩm đồ án của riêng chúng tôi và được sự hướng dẫn của thầy Nguyễn Chí Thiện. Các nội dung nghiên cứu, kết quả trong đề tài này là trung thực và chưa công bố dưới bất kỳ hình thức nào trước đây. Những số liệu trong các bảng biểu phục vụ cho việc phân tích, nhận xét, đánh giá được chính tác giả thu thập từ các nguồn khác nhau có ghi rõ trong phần tài liệu tham khảo.

Ngoài ra, trong báo cáo còn sử dụng một số nhận xét, đánh giá cũng như số liệu của các tác giả khác, cơ quan tổ chức khác đều có trích dẫn và chú thích nguồn gốc.

**Nếu phát hiện có bất kỳ sự gian lận nào tôi xin hoàn toàn chịu trách nhiệm về nội dung đồ án của mình.** Trường đại học Tôn Đức Thắng không liên quan đến những vi phạm tác quyền, bản quyền do tôi gây ra trong quá trình thực hiện (nếu có).

*TP. Hồ Chí Minh, ngày … tháng … năm 2024*

*Tác giả*

*(ký tên và ghi rõ họ tên)*

**TÓM TẮT**

Báo cáo này…

**MỤC LỤC**

**DANH MỤC KÍ HIỆU VÀ CHỮ VIẾT TẮT**

**CÁC CHỮ VIẾT TẮT**

TSP Traveling salesman problem

ACO Ant colony optimization

SOS Symbiotic optimization search

TH Trường hợp

# Introduction

## Text.

## Text.

## Text.

# Related works

Trong phần này, các công việc liên quan về khai phá các expected frequent itemset và weighted probabilistic frequent itemset được trình bày ngắn gọn:

## Khai phá expected frequent itemset:

Để giải quyết các vấn đề khai thác expected frequent itemset, có 3 phương pháp chính và cấu trúc dữ liệu được đề xuất. Đầu tiên là phương pháp apriori, Chui đề xuất thuật toán U-Apriori [], phương pháp sử dụng framework cắt dữ liệu để gia tăng hiệu quả khai phá. Sau đó, anh ấy bổ sung thêm phương pháp cắt tỉa giảm dần [] vào thuật toán để cải thiện hiệu suất. Thứ hai là phương pháp tree-based: người ta đề xuất phương pháp UF-Growth cùng với cấu trúc dữ liệu cây mới UF-tree [], phương pháp này thì có những cải tiến tốt hơn; đồng thời người ta trình bày upper bound của expected support và giới thiệu một thuật toán hiệu quả hơn gọi là BLIMP-Growth []. Thứ ba là phương pháp H-struct-based: người ta sử dụng H-struct, cấu trúc hiệu quả trong việc khai phá các dữ liệu chính xác, để thực hiện khai phá các frequent itemset trong cơ sở dữ liệu không chắc chắn người ta đề xuất UHMine [].

Bên cạnh việc phai phá các kết quả phổ biến, một vài thuật toán tập trung vào việc đạt được những kết quả chính xác và có giá trị. Phương pháp khai phá các frequent itemset rang buộc U-FPS được đề xuất [], nó sử dụng những rang buộc của người dung để thực hiện việc khai phá hiệu quả hơn. Thêm vào đó, Calders sử dụng việc lấy các mẫu để đạt được kết quả xấp xỉ []. Liu theo đuổi nghiên cứu cách để khám phá frequent itemset từ những dữ liệu không chắc chắn đơn biến.

Ngoài ra, dựa vào các phương pháp khai phá cơ sở dữ liệu, thuật toán khai phá các luồng cũng được đề xuất. Thuật toán UF-streaming [] sử dụng chế độ ngay lập tức để đạt được kết quả xấp xỉ và thuật toán SUF-growth sử dụng chế độ tạm hoãn để nhận được chính xác các frequent itemset. Leung đã nghiên cứu khai phá các frequent itemset không chắc chắn với map-reduce framework []. Những thuật toán này có thể lấy được các expected frequent itemset trong thời gian thực, nhưng chứng ẩn đi một vài kết quả lịch sử.

Đối với vấn đề này, expected support sẽ được tính toán với độ phức tạp về thời gian là và độ phức tạp về không gian là cho mỗi expected itemset, điều này có nhiều thuận lượi trong hiệu suất. Tuy nhiên, expected frequent itemset không thể hiện toàn bộ đặc điểm xác suất và độ quan trọng của từng dữ liệu, đặc biệt khi dữ liệu không có đủ trong những giao dịch.

## Khai phá weighted probabilistic frequent itemset:

Khai phá weighted probabilistic frequent itemset là một vấn đề tổng quát trong việc khai phá các tập phổ biến, nơi mà trọng số được gán cho mỗi item bởi người người để chỉ ra sự quan trọng hay quan tâm đối với mỗi item. Bởi vì trọng số được xem xét trong suốt quá trình khai phá, những item hữu ích hay được quan tâm có thể được khám phá theo sự yêu thích của người dùng.

Cai đã đề xuất tính weighted-support để giữ tính chống đơn điệu trong việc khai thác luật kết hợp khi các weighted item được xem xét. Việc tính weighted-support bao gồm nhân support của item với trung bình weighted của những item trong itemset []. Li và cộng sự đã đề xuất w-PFIs (Weighted probabilistic frequent itemsets). Phương pháp này rút ra một mô hình xác suất cho độ hỗ trợ của mỗi ứng viên w-PFIs [Efficient weighted probabilistic frequent itemset mining in uncertain databases] và đề ra ba kĩ thuật cắt tỉa nhằm mở rộng không gian tìm kiếm và loại bỏ những tập ứng viên không phù hợp. Chun-Wei Lin và cộng sự đề xuất thuật toán HEWI (high expected weighted itemset) và HEWI-Uapriori để hiệu quả hơn trong việc tìm HEWI []. Ngoài ra họ cũng giới thiệu thêm high upper-bound expected weighted downward clo sure (HUBEWDC) để cắt tỉa sớm hơn tiết kiệm không gian và loại đi những itemset không cần thiết. Yul và cộng sự đã mở rộng ý tưởng sử dụng trọng số trong khai phá các mẫu để tìm ra weighted sequential pattern. Họ đề xuất thuật toán WSpan nhầm tìm ra những weighted sequential pattern []. Thuật toán WSpan sử dụng trọng số tối đa trong tất cả item như là trọng số tới đa của mỗi chuỗi tuần tự để tìm ra weighted sequential pattern.

# Preliminaries and Problem Statements

## Preliminaries:

### Uncertain database và Possible world:

Cho 1 tập hợp các phần tử riêng biệt , , , ,…, một tập con: được gọi là **itemset.** Với mỗi item ( ) trong được liên kết với một xác xuất xuất hiện ). Chúng ta gọi X là **uncertain itemset**, biểu thị: Một **uncertain transaction** UT là một uncertain itemset đi cùng với một ID. Một **uncertain database** UD là tập hợp các uncertain transaction: . Bảng 1 bên dưới là một ví dụ về uncertain database. Bằng cách sử dụng Possible world, một uncertain database được chuyển thành nhiều exact database.

|  |  |
| --- | --- |
| **ID** | **Transaction** |
| 1 | {A 0.7} {B 0.8} |
| 2 | {A 0.1} {C 0.4} |

Bảng Ví dụ uncertain database

**Định nghĩa 1:** (Possible World) Một possible world được chuyển đổi từ uncertain database là một exact database với transaction, nơi mà mỗi transaction là một tập con của . Chúng ta ký hiệu một possible world là PW = {, ,… , }, trong đó ⊆

Nếu chúng ta giả sử rằng các uncertain transaction độc lập thì một PW có xác suất xảy ra p(PW), có thể được tính bằng cách nhân xác suất xuất hiện của từng mục x ∈ PW nếu , cũng như xác suất xuất hiện của mỗi mục nếu x thuộc , nhưng không thuộc , ký hiệu là:

Số lượng phần tử của Possible world model sẽ tăng lên theo số mũ theo kích cỡ của database gốc. Để mô tả chi tiết thì giả sử một uncertain database bao gồm m transaction, và mỗi transaction bao gồm nm item. Thì Possible world model bao gồm possible world. Mỗi phần tử trong possible world model sẽ được tính theo công thức (1). Ví dụ xem xét possible world, ta có thể tính được: p(PW6) = p{A}∈UT1∧{A}∈T1({A}) ∗p{B}∈UT1∧{B}6∈T1({B}) ∗p{A}∈UT2∧{A}∈T2({A}) ∗p{C}∈UT2∧{C}6∈T2({C}) = 0.6 \* 0.3 \* 0.2 \* 0.7 = 0.0252. Khi tính tổng các xác suất trong Possible world model, chúng ta nhận thấy tổng xác suất là 1.

### Frequent itemset trên uncertain database:

Trong exact database, khi cho 1 ngưỡng cụ thể msup (minimum support), một itemset **x** là frequent nếu không bé hơn , với là số lần xuất hiện của itemset X trong database. Trong một uncertain database, frequent itemset được định nghĩa với Possible world model. Cho 1 itemset X trong mỗi possible world PW được sinh ra từ UD, đầu tiên chúng ta cần lấy được số lần xuất hiện cùng với xác suất xuất hiện . Chúng ta biểu thị chúng với là 2-tuple . Điều này có nghĩa rằng item X có độ  với xác suất . Do đó, X trong uncertain database có tuple, chúng ta có thể tổ chức lại và biểu diễn thành summed probabilistic vector, được biểu thị .

## Problem Statements:

# Methods

Text.

Text.

Text.

## Subsection (Heading2, sentence case)

Text.

Text.

Text.

## Sub-subsection (Heading3, sentence case)

Text.

Text.

Text.

Cite figures as “Fig 1”, “Fig 2”, etc. Cite figures and tables in order. Do not cite “Fig 2” before “Fig 1”. Cite multiple figures as “Figs 1 and 2”, “Figs 1- 3”, etc. Each figure caption should appear directly after the paragraph in which they are first cited. Figure files should be saved as “Fig1.tif”, “Fig2.eps”, etc. Acceptable file formats for figures are “.tif”, “.tiff”, and “.eps”. Figures should be uploaded separately as individual files. Do not include figures in the main manuscript file.

**Fig 1. This is the figure title.**

Tables should be cited as “Table 1”, “Table 2”, etc. Cite multiple tables as “Tables 1 and 2”, “Tables 1-3”, etc. Tables must be cell-based in Microsoft Word or embedded with Microsoft Excel. Do not use empty rows to create spacing. Do not include graphic objects, images, or colored text.

**Table 1. This is the table title.**

|  |  |  |  |
| --- | --- | --- | --- |
| **Column 1 header** | **Column 2 header** | **Column 3 header** | **Column 4 header** |
| **Row 1** | valuea | value | value |
| **Row 2** | value | value | value |
| **Row 3** | value | valueb | value |

This is the Table 1 legend.

aTable footnotes belong here.

bFootnotes should have corresponding symbols in the table

Format display equations in Microsoft Equation Tools. Do not use Graphic Objects.

|  |  |
| --- | --- |
|  | (1) |

Format in regular text or as an inline equation in Microsoft Equation Tools

(). Do not use Symbol Font. Do not use Graphic Objects.

Sample citation: [1]. Cite references in brackets (for example, “[1]” or “[2-5]” or “[3,7,9]”). Using Mendeley Cite plugin for Microsoft Word to manage citations.

Style: Vancouver style

# Experiment Setup

Text.

# Experiment Results and Discussion

Text.

# Conclusion

Text.

# References

1. Ecarnot F, Seronde MF, Chopard R, Schiele F, Meneveau N. Writing a scientific article: A step-by-step guide for beginners. Eur Geriatr Med. 2015;6: 573–579. doi:10.1016/j.eurger.2015.08.005