**TRANG THÔNG TIN VỀ LUẬN ÁN**

Tên đề tài luận án: **CẤU TRÚC CỦA CÁC CHUỖI PHỔ BIẾN VỚI RÀNG BUỘC VÀ CÁC THUẬT TOÁN KHAI THÁC HIỆU QUẢ**

Chuyên ngành: Khoa học máy tính

Mã số: 62 48 01 01

Họ tên nghiên cứu sinh: Dương Văn Hải Khóa đào tạo: 2016

Người hướng dẫn khoa học: PGS. TS. Lê Hoài Bắc, TS. Trương Chí Tín

Cơ sở đào tạo: Trường Đại học Khoa học Tự nhiên, ĐHQG-HCM.

**1. TÓM TẮT NỘI DUNG LUẬN ÁN**

Luận án đề cập đến bài toán khai thác tập $FS^{C}$ gồm tất cả các chuỗi phổ biến thỏa mãn một ràng buộc C cho trước từ các cơ sở dữ liệu chuỗi (SDB). Các phương pháp truyền thống thường khai thác $FS^{C}$ trực tiếp từ SDB nên hiệu quả của chúng không cao do phải thường xuyên kiểm tra ràng buộc trên một số lượng lớn các chuỗi ứng viên trong suốt quá trình khai thác và phải đọc lại SDB mỗi khi ràng buộc thay đổi. Để khắc phục, luận án đã đưa ra một mô hình mới cho phép sinh ra tập $FS^{C}$ từ tập các chuỗi đóng phổ biến $FCS$ và tập các chuỗi sinh phổ biến $FGS$. Mô hình này dựa vào ba bài toán chính $BT\_{1-3}$ dưới đây:

Bài toán đầu tiên ($BT\_{1}$) là khai thác hiệu quả $FCS$ và $FGS$ từ SDB. Vì hai tập $FCS$ và $FGS$ có ý nghĩa quan trọng trong nhiều ứng dụng thực tế, nên nhiều thuật toán đã được đề xuất để khai thác chúng, nhưng hiệu quả chưa cao do còn tồn tại nhiều hạn chế. Luận án đã đưa ra một số đo mới, các điều kiện tỉa sớm dựa trên số đo này và các chiến lược tỉa địa phương nhằm loại bỏ nhanh nhiều chuỗi ứng viên trong cây tiền tố, qua đó rút gọn nhanh không gian tìm kiếm. Bốn thuật toán đã được đề xuất để khai thác tuần tự và song song $FCS$ và/hoặc $FGS$.

Về mặt lý thuyết, hai tập $FCS$ và $FGS$ là các biểu diễn súc tích và không mất thông tin của $FS$ (tập tất cả các chuỗi phổ biến), nghĩa là hai tập này mặc dù có lực lượng thường bé hơn rất nhiều so với $FS$ nhưng chúng có thể được kết hợp với nhau để sinh ra tập $FS$ mà không cần đọc lại SDB. Tuy nhiên, cho đến nay chưa có nghiên cứu nào đề xuất một phương pháp tường minh để sinh ra nhanh $FS$ từ $FCS$ và $FGS$. Vì vậy, trong bài toán $BT\_{2}$ luận án tập trung nghiên cứu cấu trúc của tập $FS$ dựa trên $FCS$ và $FGS$. Cấu trúc ở đây có nghĩa là mối quan hệ tường minh giữa các chuỗi trong $FS$ và các chuỗi (đại diện của chúng) trong $FCS$ và $FGS$. Phương pháp phân hoạch đã được sử dụng để chia $FS$ thành các lớp tương đương được đại diện bởi một số chuỗi đóng và chuỗi sinh phổ biến trong $FCS$ và $FGS$. Từ cấu trúc lý thuyết này, luận án đã đưa ra một phương pháp để sinh ra nhanh, đầy đủ và không trùng lặp tất cả các chuỗi trong mỗi lớp dựa trên các chuỗi đại diện của lớp đó.

Bài toán thứ ba ($BT\_{3}$) nhằm chỉ ra cách sinh nhanh tập $FS^{C}$ từ $FCS$ và $FGS$. Phương pháp phân hoạch cũng đã được sử dụng trong bài toán này để chia tập lời giải $FS^{C}$ thành các lớp tương đương. Luận án đã sử dụng tính chất đơn điệu giảm hoặc tăng của các lớp ràng buộc, tính tối tiểu và tối đại của các phần tử đại diện để loại bỏ sớm nhiều lớp chuỗi ứng viên bằng việc kiểm tra ràng buộc trên chỉ các phần tử đại diện của chúng. Mối quan hệ tường minh giữa $FS$ và cặp ($FCS$ và $FGS$) trong bài toán $BT\_{2}$ đã được áp dụng để sinh ra nhanh tất cả các chuỗi trong các lớp còn lại trong $FS^{C}$.

**2. NHỮNG KẾT QUẢ MỚI CỦA LUẬN ÁN**

*Bài toán* $BT\_{1}$*.* Luận án đã chỉ ra sự không chính xác trong trường hợp tổng quát của một số kết quả lý thuyết đã được đưa ra năm 2003 và đã được trích dẫn và sử dụng nhiều trong các thuật toán sau đó để khai thác các tập $FCS$ và $FGS$. Một số đo $SE$ mới cùng với các kết quả lý thuyết chính xác và tổng quát hơn đã được đề xuất trong luận án. Chúng là cơ sở lý thuyết tin cậy để đưa ra các điều kiện tỉa sớm và các chiến lược tỉa địa phương nhằm loại bỏ nhanh các chuỗi ứng viên không là đóng và/hoặc sinh phổ biến trên cây tìm kiếm tiền tố. Luận án đã đề xuất bốn thuật toán mới: FGenSM (khai thác $FGS$), FCloSM (khai thác $FGS$), FGenCloSM (khai thác tuần tự cả hai $FCS$ và $FGS$) và Par-GenCloSM (khai thác song song đồng thời $FCS$ và $FGS$). Sự chính xác và hiệu quả của các thuật toán này đã đóng góp nhiều vào việc giải quyết hiệu quả bài toán $BT\_{3}$.

*Bài toán* $BT\_{2}$. Một cấu trúc của $FS$ dựa trên hai tập biểu diễn súc tích $FCS$ và $FGS$ của nó đã được đưa ra trong luận án. Nó cho thấy mối quan hệ tường minh giữa $FS$ và hai tập biểu diễn súc tích của này. Đó cũng chính là cơ sở lý thuyết của một phương pháp sinh nhanh $FS$ từ $FCS$ và $FGS$ được chỉ ra trong luận án.

*Bài toán* $BT\_{3}$. Luận án đã đề xuất một mô hình mới cho phép khai thác $FS^{C}$ từ $FCS$ và $FGS$, thay vì khai thác nó trực tiếp từ SDB như các tiếp cận truyền thống. Dựa trên những kết quả đóng góp của luận án cho hai bài toán $BT\_{1}$ và $BT\_{2}$, tính hiệu quả của mô hình này ít chịu ảnh hưởng khi các ràng buộc được thay đổi thường xuyên bởi người dùng. Hai thuật toán tổng quát MFS-AC và MFS-MC đã được đề xuất để khai thác hiệu quả các chuỗi phổ biến với nhiều loại ràng buộc khác nhau thuộc hai lớp ràng buộc thông dụng $C^{A}$ và $C^{M}$ (hai lớp ràng buộc có tính đơn điệu giảm và tăng). Lưu ý là, đa số các thuật toán trước đây thường được thiết kế cho một loại ràng buộc cụ thể và chúng thường gặp nhiều khó khăn trong việc rút gọn không gian tìm kiếm khi đối mặt với lớp ràng buộc $C^{M}$.

**3. CÁC ỨNG DỤNG/ KHẢ NĂNG ỨNG DỤNG TRONG THỰC TIỄN HAY NHỮNG VẤN ĐỀ CÒN BỎ NGỎ CẦN TIẾP TỤC NGHIÊN CỨU**

Trong thời gian tới, chúng tôi sẽ áp dụng mô hình được đưa ra trong luận án để giải bài toán khai thác các chuỗi phổ biến với những loại ràng buộc phức tạp hơn. Chúng tôi đồng thời sẽ áp dụng mô hình này để giải bài toán tổng quát hơn, khai thác các chuỗi có lợi ích cao trên cơ sở dữ liệu chuỗi lượng hóa với ràng buộc.

|  |  |
| --- | --- |
| **CÁN BỘ HƯỚNG DẪN** | **NGHIÊN CỨU SINH****Dương Văn Hải** |

**XÁC NHẬN CỦA CƠ SỞ ĐÀO TẠO**

**PHÓ HIỆU TRƯỞNG**

**THESIS INFORMATION**

Thesis title: **STRUCTURE OF FREQUENT SEQUENCES WITH CONSTRAINTS AND EFFICIENT MINING ALGORITHMS**

Speciality: Computer Science

Code: 62 48 01 01

PhD student: Duong Van Hai Academic year: 2016

Supervisors: Prof. Le Hoai Bac, Dr. Truong Chi Tin

At: University of Science, VNU-HCMC, Vietnam.

**1. SUMMARY OF THE THESIS**

This thesis considers the problem of mining the set $FS^{C}$ of all frequent sequences satisfying a given C constraint from sequential databases (SDB). Traditional methods discover $FS^{C}$ directly from an SDB, and thus their performance is not high since they must regularly check constraints on a large number of candidate sequences during the mining process, and read an SDB again when a constraint is changed by the user. To overcome this issue, this thesis proposes a novel model that allows the quick generation of $FS^{C}$ from the sets $FCS$ of frequent closed sequences and $FGS$ of frequent generator sequences. This model is based on the following three main problems $BT\_{1-3}$.

The first problem ($BT\_{1}$) is to effectively mine $FCS$ and $FGS$ from an SDB. Because these two sets have many real-life applications, many algorithms have been proposed to discover them. However, their performance is poor in terms of runtime and memory usage due to certain limitations. This thesis introduces a new measure, early pruning conditions based on this measure and local pruning strategies to quickly eliminate many candidate sequences on the prefix search tree. Four algorithms are proposed to mine $FCS$ and/or $FGS$.

Theoretically, the two sets $FCS$ and $FGS$ are concise and lossless representations of the set $FS$ of all frequent sequences, i.e., although their cardinalities are often much smaller than that of $FS$, they can be combined together to generate $FS$ without reading an SDB again. However, up to now no studies have proposed an explicit method for quickly generating $FS$ from $FCS$ and $FGS$. Thus, in the problem $BT\_{2}$, the thesis focuses on the structure of $FS$ based on $FCS$ and $FGS$. The structure indicates an explicit relationship between frequent sequences in $FS$ and their representative sequences in $FCS$ and $FGS$. The partitioning method is used to divide $FS$ into equivalence classes represented by frequent closed and generator sequences. The thesis then proposes a method to quickly generate all sequences in each class based on its representative sequences.

The third problem ($BT\_{3}$) is to quickly generate the set $FS^{C}$ from $FCS$ and $FGS$. The partitioning method is also used in this problem to divide the solution set $FS^{C}$ into equivalence classes. The thesis uses anti-monotonic and monotonic properties of constraint classes, and the minimum and maximum of representative sequences to eliminate many classes of candidate sequences at an early stage by checking constraints on only their representative elements. To quickly generate all sequences in the remaining classes in $FS^{C}$, the explicit relationship between $FS$ and the pair ($FCS$ and $FGS$) in the problem $BT\_{2}$ is used.

**2. CONTRIBUTIONS OF THE THESIS**

*Problem* $BT\_{1}$*.* This thesis shows the incorrectness of some theoretical results (in the general case) that were proposed in 2003, then cited and used in typical algorithms to mine the sets $FCS$ and $FGS$. A novel SE measure and more accurate and general theoretical results are proposed in this thesis. They are the basis for proposing early pruning conditions and local pruning strategies to quickly eliminate candidate sequences that are not closed and/or generators. This thesis proposes four new algorithms FGenSM, FCloSM, FGenCloSM, and Par-GenCloSM for mining $FGS$ and/or $FCS$. The correctness and efficiency of the algorithms greatly contribute to efficiently solving the problem $BT\_{3}$.

*Problem* $BT\_{2}$*.* The structure of $FS$, based on $FCS$ and $FGS$, is introduced in the thesis. It indicates an explicit relationship between $FS$ and these two sets, and it is also the theoretical basis for quickly generating $FS$ from $FCS$ and $FGS$.

*Problem* $BT\_{3}$*.* The thesis proposes a novel model for mining $FS^{C}$ from $FCS$ and $FGS$, instead of mining it directly from an SDB as in traditional approaches. Based on the contributions of the thesis to the problems $BT\_{1}$ and $BT\_{2}$, the efficiency of this model is not much influenced when constraints are regularly changed by users. Two general algorithms MFS-AC and MFS-MC are proposed to effectively mine frequent sequences with different types of anti-monotonic and monotonic constraints, respectively. Note that most previous algorithms were designed for a specific constraint, and they often face challenges in reducing the search space with monotonic constraints.

**3. FUTURE WORK**

In the future, we will apply the model proposed in this thesis to solve the problem of mining frequent sequences with more complicated constraints. We will also apply this model to solve the more general problem of discovering high utility sequences on quantitative sequential databases with constraints

|  |  |
| --- | --- |
| **SUPERVISORS** | **PhD STUDENT****Duong Van Hai** |

**CONFIRMATION OF THE UNIVERSITY OF SCIENCE**

**VICE PRESIDENT**