THESIS INFORMATION

Thesis title: **Mining Frequent Sequential Patterns based on Constraints**
Major: Computer Science
Major Code: 62 48 01 01
PhD Student: VAN THI THIEN TRANG
Supervisor: Prof. Dr. Le Hoai Bac
Training place: University of Information Technology, Vietnam National University - Ho Chi Minh City

1. ABSTRACT OF THE THESIS

The thesis presents an overview and explores the theoretical basis of mining sequential patterns based on constraints. The aim of the thesis is to study the problem of sequential pattern mining with Itemset constraints and apply the constraint-satisfied pattern set in sequential rule mining with itemset constraints. Besides, it studies a specific application area of this problem which is mining web access patterns with super-pattern constraints. The thesis has completed the goals of proposing effective methods for the given problems so that it is possible to find the sequential patterns exactly, shorten the mining time and reduce the memory usage.

Results of this thesis are combined from 4 articles published in journals: Knowledge and Information Systems (1 article), Vietnam Journal Computer Science (Springer, 1 article), Applied Intelligence (2 articles) and presented in the following 3 main problems:

Problem 1: mining sequential patterns with an itemset constraints is to find all frequent patterns in the database which contain one of the itemsets indicated by the user in order to reduce the number of redundant patterns. The mining focuses on the user interests, reduces the mining time and the memory usage

Problem 2: from the result of Problem 1, the thesis considers the problem of mining sequential rules with itemset constraints in the left parts of the rules by utilizing the constraint-satisfied pattern set in rule generating process.

Problem 3: mining web access patterns with super-pattern constraint is to find all frequent patterns in the database which contain any pattern indicated by the user as subsequence.

2. NEW RESULTS OF THE THESIS

Thesis contains some new results, which have been published on prestigious scientific journals. The novelty of thesis can be mentioned as follows:

Proposing the *MSPIC-DBV* algorithm, an efficient algorithm for mining sequential patterns with itemset constraints. The *MSPIC-DBV* algorithm uses a new data structure, called Dynamic Bit Vector to compress candidate information. In particular, it has two strategies: one for pruning the search space via the use of prefixes and subsequences both at the beginning and during the mining process, and reducing the calculations needed for

pattern extension; and one for skipping the checking constraint step for a large amount of candidates.

Proposing three algorithms for mining sequential rules with itemset constraints, namely *MSRIC-B*, *MSRIC-R* and *MSRIC-P*. *MSRIC-B* is a naïve strategy putting the itemset constraints into the post-processing step and two remaining algorithms integrate the itemset constraints into the actual mining process in which *MSRIC-R* put the constraints in the period of rule generating and *MSRIC-P* performs in the pattern mining by using the constraint-satisfied patterns. *MSRIC-P* algorithm is the main contribution of this study and it is more efficient than the others.

Proposing two algorithms for mining web access patterns with super-pattern constraint, namely *MWAPC* and *EMWAPC*. Noticeably, the latter is the main contribution and it is an improved algorithm of *MWAPC*. *EMWAPC* uses the data structure and the techniques similar to the method of mining with itemset constraint. However, based on the characteristics of the web access patterns and the super-pattern constraints, it is able to prune the search space quickly and minimizes the constraint checking.

3. APPLICATIONS/POSSIBILITIES OF APPLICATION IN PRACTICE OR QUESTION ISSUES TO CONTINUE THE RESEARCH

In the future, we will continue to develop the methods for mining sequential patterns based on constraints which can achieve more optimal speed and memory.

Study the problem of mining sequential patterns based on constraints in the distributed databases in order to find the solutions for big data.

Apply the proposed methods for mining sequential patterns and rules with other constraints such as super-pattern constraints or incorporating constraints.

SUPERVISOR	PhD STUDENT
Prof. Dr. Le Hoai Bac	**Van Thi Thien Trang**