Effective Algorithms for Sequential Pattern Mining

Zhenglu YANG* Yitong WANG* Masaru KITSUREGAWA ♦

Sequential pattern mining is very important because it is the basis of many applications. Although there has been a great deal of effort on sequential pattern mining in recent years, its performance is still far from satisfactory because of two main challenges: large search spaces and the ineffectiveness in handling dense data sets. To offer a solution to the above challenges, we have proposed a series of novel algorithms, called the LAst Position INDuction (LAPIN) sequential pattern mining, which is based on the simple idea that the last position of an item, \( \sigma \), is the key to judging whether or not a frequent \( k \)-length sequential pattern can be extended to be a frequent \( (k+1) \)-length pattern by appending the item \( \sigma \) to it. LAPIN can largely reduce the search space during the mining process, and is very effective in mining dense data sets. Our experimental data and performance studies show that LAPIN outperforms PrefixSpan by up to an order of magnitude on long pattern dense data sets.

1. Introduction

* Student Member Institute of Industrial Science, the University of Tokyo yangzl@tkl.iis.u-tokyo.ac.jp
* Computer Science Department, Fudan University, China yitongw@fudan.edu.cn
* Regular Member Institute of Industrial Science, the University of Tokyo kitsure@tkl.iis.u-tokyo.ac.jp

1.1 Overview of Our Algorithm

日本データベース学会 日本データベース学会
2. LAPIN: Design and Implementation

The following sections describe the design and implementation of the LAPIN algorithm. The algorithm is designed to efficiently discover frequent itemsets in a database. The design is based on a combination of existing techniques and new optimizations to handle large datasets.

2.1 LAPIN LCI

The LAPIN LCI algorithm is a modified version of the LAPIN algorithm that focuses on discovering frequent itemsets with a length of exactly L. The algorithm is implemented as a modified version of the LAPIN algorithm.

LAPIN Algorithm:

Input: A sequence database, and the minimum support threshold, δ
Output: The complete set of sequential patterns

Function: GenPatterns(S, CandL, CandL+1)
Parameters: n = length of frequent sequential pattern; S = prefix border position set of (S(n-1)-length sequential pattern; CandL = candidate sequence extension item list of length k+1 sequential pattern; CandL+1 = candidate sequence extension item list of length k+1 sequential pattern
Goal: Generate (k+1)-length frequent sequential pattern
Main:
1. Scan DB once to do
   1.1 P1 — Compute the prefix list representation of the 1-length SE sequences
   1.2 P2 — Find the frequent 1-length SE sequences
   1.3 P3 — Obtain the item list position list of the 1-length SE sequences
   1.4 P4 — Find the frequent 2-length SE sequences
   1.5 P5 — Construct the position lists of the frequent 2-length SE sequences
   1.6 P6 — Obtain the item list-position list of the frequent 2-length SE sequences
2. For each frequent SE sequence α_l in D
   2.1 Call GenPatterns(α_l, 0, 0, 0)
   2.2 Call GenPatterns(α_l, 0, 0, 0)

Function GenPatterns(S, CandL, CandL+1)
4. S_l — Find the prefix border position set of S based on S
5. PrefixItems_l — Obtain the SE item list of α based on CandL and S_l
6. PrefixItems_m — Obtain the SE item list of α based on CandL and S_m
7. For each item γ_l in PrefixItems_l, γ_l ∈ SE results in S and output
8. Call GenPatterns(S, S_l, PrefixItems_l, PrefixItems_m)
論文

2.2 LAPIN Suffix

3. Performance Study
4. Conclusion

Ph.D student of Graduate School of Information Science and Technology, the University of Tokyo. He received the Master degree in the above graduate school. His research interests include sequence mining and data mining.

Associate professor at Fudan University, China. She received Ph.D degree in computer science in 1999 from the above university. Her research interests include data mining and parallel processing.

Professor and the director of center for information at the University of Tokyo. He received the Ph.D degree in information engineering in 1983 from the University of Tokyo. His research interests include parallel processing and database engineering. He is a member of steering committee of IEEE ICDE, PAKDD and WAIM, and has been a trustee of the VLDB Endowment. He was the chair of data engineering special interest group of Institute of Electronic, Information, and Endowment. He was the chair of data engineering special interest group of Institute of Electronic, Information, and Endowment.

4. Conclusion

Ph.D student of Graduate School of Information Science and Technology, the University of Tokyo. He received the Master degree in the above graduate school. His research interests include sequence mining and data mining.

Associate professor at Fudan University, China. She received Ph.D degree in computer science in 1999 from the above university. Her research interests include data mining and parallel processing.

Professor and the director of center for information at the University of Tokyo. He received the Ph.D degree in information engineering in 1983 from the University of Tokyo. His research interests include parallel processing and database engineering. He is a member of steering committee of IEEE ICDE, PAKDD and WAIM, and has been a trustee of the VLDB Endowment. He was the chair of data engineering special interest group of Institute of Electronic, Information, Communication, Engineering, Japan, the chair of ACM SIGMOD Japan, and has been a trustee of DBS.

Japanデータベース学会 40周年記念号