Data Mining

ICTU Student 160
Thai Nguyen University of Information and Communication Technology
Thai Nguyen University, Thai Nguyen, Vietnam
student160@ictu.edu.vn

***Abstract*—This thesis explores data mining techniques, focusing on the Apriori algorithm and its improvements, particularly the FP-growth algorithm. It provides an overview of data mining, its applications, and the limitations of the Apriori algorithm, then introduces the FP-growth algorithm for more efficient frequent pattern mining. The thesis details the implementation of the FP-growth algorithm on a sample dataset, demonstrating its efficiency in finding frequent patterns compared to the Apriori algorithm. The results show that FP-growth requires fewer database scans and is more efficient in terms of computational resources. The discussion highlights the advantages of the FP-growth algorithm, particularly in handling large datasets and reducing computational costs. The conclusion suggests that the FP-growth algorithm could be applied in various fields, including retail, finance, and healthcare.**

***Index Terms*—data mining, Apriori algorithm, FP-growth, frequent pattern mining, database, algorithm efficiency**

I. Introduction

The thesis begins with an introduction to data mining, highlighting its importance and applications in various fields. It discusses the challenges faced by the Apriori algorithm, such as high computational costs and the need for multiple database scans, and introduces the FP-growth algorithm as a more efficient alternative. The introduction also covers the basic concepts of data mining, including the definition, types of data, and the process of data mining. The thesis further explores the various approaches and techniques used in data mining, as well as its applications in different sectors. The limitations of the Apriori algorithm are discussed, emphasizing the need for a more efficient method to handle large datasets [1]. Related earlier ICTU undergraduate theses provide context for this English summary [5]. Published articles on the methods and subject of this project provide context for this English summary [6], [7], [8], [9], [10], [11].

II. Methods

The thesis describes the Apriori algorithm and its limitations, then details the FP-growth algorithm. It includes a step-by-step explanation of the FP-growth algorithm, including the construction of the FP-tree and the mining of frequent patterns without the need for multiple database scans. The implementation of the FP-growth algorithm on a sample dataset is described, with a focus on the process of counting the frequency of each item, sorting items based on their frequency, and constructing the FP-tree. The thesis also explains the process of mining frequent patterns from the FP-tree, which involves recursively building conditional FP-trees and extracting frequent patterns [1].

Supervisor: ICTU Teacher 102, Thai Nguyen University of Information and Communication Technology.

III. Results

The thesis presents the results of applying the FP-growth algorithm to a sample dataset, demonstrating its efficiency in finding frequent patterns compared to the Apriori algorithm. The results show that FP-growth requires fewer database scans and is more efficient in terms of computational resources. The implementation of the FP-growth algorithm on a sample dataset, as shown in Table 2.7 and Figure 2.2 to 2.6, illustrates the process of counting the frequency of each item, sorting items based on their frequency, and constructing the FP-tree. The thesis also provides a detailed example of the FP-growth algorithm's application, highlighting its effectiveness in identifying frequent patterns without the need for multiple database scans [1].

IV. Discussion

The discussion focuses on the advantages of the FP-growth algorithm over the Apriori algorithm, particularly in handling large datasets and reducing computational costs. The thesis also mentions potential areas for further research, such as optimizing the FP-tree construction process. The FP-growth algorithm is shown to be more efficient in scenarios with large datasets, as it requires fewer database scans and is less computationally intensive. The thesis suggests that this algorithm could be applied in various fields, including retail, finance, and healthcare, where frequent pattern mining is crucial [1].

V. Conclusion

The thesis concludes that the FP-growth algorithm is a more efficient method for frequent pattern mining, especially in scenarios with large datasets. It suggests that this algorithm could be applied in various fields, including retail, finance, and healthcare, where frequent pattern mining is crucial. The conclusion also highlights the need for further research to optimize the FP-tree construction process and improve the overall efficiency of the algorithm [1].

References

[1] ICTU Student 160, "Data Mining," internal undergraduate thesis (original in Vietnamese), Thai Nguyen University of Information and Communication Technology, 2010.
[2] Vũ Đức Thi, Cơ sở dữ liệu – Kiến thức và thực hành, Nhà xuất bản thống kê, 1997

[3] Vũ Đức Thi, Lê Hải Khôi, Một số nguyên lý hoạt động của kho dữ liệu, Tạp chí Tin học và Điều khiển, 1999
[4] Phan Đình Diệu, Lô gic trong các hệ tri thức, Hà Nội, 1999
[5] ICTU Student 10, "Data Mining: An Overview and Application of Association Rule Mining," internal undergraduate thesis, Thai Nguyen University of Information and Communication Technology, 2007.
[6] Qinghua Zou, Wesley Chu, David Johnson, Henry Chiu, "A Pattern Decomposition Algorithm for Data Mining of Frequent Patterns," Knowledge and Information Systems, vol. 4, no. 4, pp. 466-482, 2002. DOI: 10.1007/s101150200016.
[7] Jiawei Han, Hong Cheng, Dong Xin, Xifeng Yan, "Frequent pattern mining: current status and future directions," Data Mining and Knowledge Discovery, vol. 15, no. 1, pp. 55-86, 2007. DOI: 10.1007/s10618-006-0059-1.
[8] Jiawei Han, Jian Pei, Yiwen Yin, Runying Mao, "Mining Frequent Patterns without Candidate Generation: A Frequent-Pattern Tree Approach," Data Mining and Knowledge Discovery, vol. 8, no. 1, pp. 53-87, 2004. DOI: 10.1023/b:dami.0000005258.31418.83.
[9] Moonjung Cho, Jian Pei, Haixun Wang, Wei Wang, "Preference-Based Frequent Pattern Mining," International Journal of Data Warehousing and Mining, vol. 1, no. 4, pp. 56-77, 2005. DOI: 10.4018/jdwm.2005100103.
[10] Jian Pei, Jiawei Han, "Constrained frequent pattern mining," ACM SIGKDD Explorations Newsletter, vol. 4, no. 1, pp. 31-39, 2002. DOI: 10.1145/568574.568580.
[11] Qiong ZHU, "Algorithm of frequent-patterns mining in data stream," Journal of Computer Applications, vol. 28, no. 6, pp. 1463-1466, 2008. DOI: 10.3724/sp.j.1087.2008.01463.