Examiner's report of doctoral thesis

Author: Huy Minh Huynh
Title: Efficient Methods for Mining Clickstream Patterns

Examiner: prof. RNDr. PaedDr. Eva Volná, PhD.
University of Ostrava

a) Topicality of the dissertation topic

The topic of the thesis, Clickstream Pattern Mining (CPM), is topical due to the growing volume of digital data from the Internet and online shopping. Manual analysis of this data is difficult, which has led to the development of CPM. Although it has often been treated as part of Sequential Pattern Mining (SPM), it has not yet been sufficiently studied on its own. This offers opportunities for improving performance and efficiency, confirming the importance and relevance of the topic.

b) Objectives of the thesis and their fulfilment

The dissertation set five objectives, all of which were successfully achieved.

1. **Propose approaches that can exploit certain clickstreams' characteristics for mining frequent clickstream patterns.** This goal was achieved by designing the CUP (Clickstream pattern mining Using Pseudo-IDList) algorithm. CUP uses Pseudo-IDList to reduce duplicate data and solve memory issues, and DUB (Dynamic Intersection Upper Bound) for efficient computation.
2. **Extend the proposed algorithm to mine sequential patterns in which each itemset can include multiple items/events.** The CUP algorithm has been extended to the SUI (Sequential pattern mining Using IDList) algorithm, which effectively solves the problem of sequential pattern mining.
3. **Integrate weight factor to give another option for ranking patterns other than the general support measure (i.e., the frequency of which the patterns appear).** The goal was achieved by integrating weighting factors and parallelism into pattern mining. New algorithms, Compact-SPADE, DPCompact-SPADE, and APCompact-SPADE, were proposed, which use average weight as an alternative metric for pattern evaluation.
4. **Improve performance by integrating parallelism into existing clickstream pattern mining algorithms.** The goal was achieved by implementing several parallel strategies for greater efficiency. These include DPCompact-SPADE and APCompact-SPADE, which optimize the performance of mining weighted clickstream patterns.
5. **Propose methods to tackle the situation where new clickstream data is being added on a daily basis.** This goal was achieved by proposing three new algorithms: PF-CUP, PSB-CUP, and PSB-CUP+, which solve the problem of incremental clickstream pattern mining.

c) Results of the dissertation

The thesis primarily focuses on the performance of algorithms rather than on evaluating the quality of the patterns found. The student focused on the efficient mining of clickstream and sequential patterns. He designed and implemented a number of innovative algorithms.

- **CUP** – a new algorithm for mining clickstream patterns that reduces memory requirements and speeds up computation thanks to Pseudo-IDList and DUB techniques.
- **SUI** – CUP extension for general sequence patterns with optimization for different types of pattern endings.
- **Compact-/DP-/APCompact-SPADE** – Weighted algorithms with new structures (WICList, WCMAP) and parallel strategies that improve performance and scalability.
- **PF-/PSB-/PSB-CUP+** – incremental algorithms for processing new data with an emphasis on saving memory and computing time.

The student's contribution lies in the design of new methods, data structures, and optimization techniques that improve upon existing approaches in speed, memory efficiency, and scalability.

d) Benefits in the field of knowledge

Optimized and incremental algorithms streamline clickstream data analysis, improve user experience, support personalization, and reduce costs. The work extends the less explored field of CPM by introducing new data structures, heuristics, and parallel techniques, thus pushing the boundaries of pattern mining.

e) Formal arrangement

Formally, the thesis includes all standard requirements for a dissertation. The thesis is written in English.

f) Publication of the doctoral student

The student has a good publication record – nine articles in impact journals (including Q1/D1) and four conference papers, one of which received an award. The outputs reflect a significant contribution to the field and active involvement in the scientific community.

Questions and comments

1. In your work, you mentioned various ways of storing click information, such as "Pseudo-IDList," "Data-IDList," and "WICList." Could you explain the main advantages and disadvantages of each of these data storage methods?
2. The thesis presents several algorithms optimized for different aspects of clickstream pattern mining (frequency, sequence, weighted, incremental). However, the performance of the algorithms can be different depending on the type of database and the set thresholds. Could you explain how to choose the most suitable algorithm for

specific clickstream pattern mining needs to optimize both runtime and memory consumption?

Conclusion

The submitted thesis fulfils the requirements for a doctoral thesis, both in terms of theoretical - methodological level, so the usefulness in practice. The thesis contains the original results.

I recommend the thesis to the defence before the relevant commission. Based on the thesis, I suggest the academic and scientific degree "Doctor Philosophiae" (Ph.D. abbreviation) to confer to Huy Minh Huynh after successfully defending of his thesis.

Ostrava, 6 August 2025

prof. RNDr. PaedDr. Eva Volná, PhD.

Radek Silhavy, PhD
Associate Professor in System Engineering and Informatics

Review of Dissertation Thesis

Dissertation topic: Efficient Methods for Mining Clickstream Patterns

PhD Candidate: Huy Minh Huynh
Supervisor: prof. Zuzana Komínková Oplatková, Ph.D.

The submitted dissertation thesis addresses a current and practically significant issue related to mining patterns from clickstream data. In today's digital era, where user interactions with websites and online services generate vast amounts of data, the ability to efficiently analyse these sequences of clicks is crucial. Applications range from personalising the user experience, optimising website design, and building recommendation systems, to detecting anomalous behaviour and conducting marketing analyses.

The author correctly identifies that Clickstream Pattern Mining is often viewed merely as a subset of the more general problem of Sequential Pattern Mining. However, this perspective overlooks the specific characteristics of clickstream data (e.g., that each event is atomic and does not contain a set of items), which can lead to suboptimal performance when using general SPM algorithms.

The thesis focuses on an actual topic, and its relevance is enhanced by the following aspects:

1. The work proceeds in a very logical and systematic manner. It begins with the fundamental problem of frequent clickstream patterns (Chapter 3), then extends and generalises the proposed solution to the broader SPM (Chapter 4), and subsequently tackles advanced and practical variants—weighted and parallel mining (Chapter 5) and incremental mining (Chapter 6). This progression demonstrates a deep understanding of the subject matter and the ability to build a complex solution step-by-step.

2. The author introduces not just one, but a whole suite of new algorithms for the individual problems addressed.

3. Each proposed algorithm is meticulously validated through experiments on several real-world and synthetic datasets. The performance is compared against established state-of-the-art algorithms (e.g., PrefixSpan, CM-SPADE). The measured metrics (runtime, maximum memory consumption, scalability) are standard in the field, and the results are clearly presented and analysed.

4. The quality of the work is confirmed by numerous publications based on its chapters. The author has published the results in prestigious, high-impact journals (Q1, D1). This is evidence of the originality, contribution, and acceptance of the results by the scientific community.

Although the thesis is of a good standard, a few points can be identified that may serve as topics for discussion during the defence, rather than as criticisms of major shortcomings.

1. The author explicitly states in the thesis that it "mainly focuses on the performance of algorithms and does not judge the patterns' quality." This is a legitimate scoping of the work. However, for practical implementation, the quality and interpretability of the patterns are key. A faster algorithm is only beneficial if the patterns it finds lead to better business outcomes. A discussion on how different approaches (e.g., weighted vs. unweighted) might affect the relevance of patterns for a specific application would be valuable.

2. The thesis concentrates on the algorithmic and experimental aspects. Although applications are mentioned, a more in-depth case study demonstrating the deployment of one of the proposed algorithms on a real-world problem (e.g., A/B testing of website improvements based on patterns from PSB-CUP+) would further underscore the practical impact of the work.

Contribution of the Thesis

The main contribution lies in successfully defining CPM as a specific problem and designing a family of algorithms that effectively leverage its characteristics. The key innovation is the Pseudo-IDList concept, which demonstrably reduces memory requirements and improves runtime compared to state-of-the-art methods that process clickstream data with more general procedures.

The thesis clearly delineates itself from existing solutions by directly addressing their inefficiency when applied to clickstream data. The experimental comparison with algorithms like PrefixSpan and CM-SPADE proves the superiority of the proposed methods on the tested datasets in the context of CPM.

Further contributions lie in extending the core solution for advanced scenarios (weighted, parallel, incremental mining), which makes the results of the work a comprehensive and practically applicable toolkit for clickstream data analysis.

Questions for the Defence

1. From the perspective of a business stakeholder, how would you explain the added value of your algorithms (e.g., PSB-CUP+) compared to an existing solution like PrefixSpan? How could your results translate into specific business metrics (e.g., increased conversion rates, longer time on site)?

2. In Chapter 5, you proposed a weighted mining model. Could you provide a specific example of a scenario where standard (unweighted) mining would fail or provide misleading results?

3. Your APCompact-SPADE algorithm uses heuristic sampling for adaptive selection of a parallelisation strategy. What are the main risk factors of this heuristic? Could you provide a practical example of a situation where a data sample is unrepresentative?

4. In recent years, deep learning-based models (e.g., RNNs, Transformers) have become increasingly prominent for sequential analysis. Where do you see the main advantages

and disadvantages of your combinatorial approaches compared to these modern methods for analysing clickstream data?

Final Evaluation

The submitted dissertation thesis by Huy Minh Huynh represents an original scientific work. The author has demonstrated an ability to identify a relevant research problem, design innovative and efficient algorithmic solutions, and validate them through experiments. The thesis is logically structured, clearly written, and its results have been confirmed by publications in the most prestigious international journals in the field.

The thesis fully meets all the requirements for a doctoral dissertation in the field of Information Technologies. The author has demonstrated the ability for independent scientific work.

Based on the above, I recommend that the thesis be accepted for defence. Upon successful defence, I propose that the degree of Doctor of Philosophy be awarded to the candidate.

In Zlín, August 6, 2025,

Radek Silhavy, Ph.D.
Dept. of Computer and Communication Systems
Faculty of Applied Informatics
Tomas Bata University in Zlín

VSB TECHNICAL UNIVERSITY OF OSTRAVA | FACULTY OF ELECTRICAL ENGINEERING AND COMPUTER SCIENCE | DEPARTMENT OF COMPUTER SCIENCE

Dissertation thesis review

The dissertation (doctoral) thesis entitled „**Efficient Methods for Mining Clickstream Patterns** ", submitted by **M.Eng. Huy Minh Huynh**, summarizes author's research and main results in the field of the data analysis and mining. The work aims at clickstream, a type of sequential data associated with the interactions of agents (users and systems) on the World Wide Web, in particular in the context of e-commerce. The topic is up-to-date and under active development by a variety of researchers worldwide. The author presents in total 8 original algorithms focusing on clickstream and sequential pattern mining. This is an impressive number of novel contributions to science, although it is slightly increased by including preliminary (PSB-CUP) and extended (PSB-CUP+) versions of the same principal approach.

The manuscript is organized into 7 sections. The structure is logical and very well describes the problem(s), the state-of-the-art, and the contributions of the author. The related work section (Sec. 2) is comprehensive and provides a complex summary of the problems and challenges associated with the target field as well as state-of-the-art methods and algorithms. The sections dedicated to individual novel algorithms (Sec. 3-6) follow similar outline (problem definition, set goals, description of the proposed approach, experimental evaluation), which contributes to the clarity of the manuscript. Problem descriptions are correct and the goals are clearly structured, valid and complement the state of the art.

The author extended the well-known Sequential Pattern Discovery Using Equivalence Class (SPADE) algorithm by several improvements such as efficient data structures, new pruning approaches, parallel implementation, and applied the new methods to clickstream (regular, weighted, and incremental) and sequential pattern mining. The descriptions of the proposed concepts are essentially understandable and the proposed approaches credible. The experiments illustrate the validity of the proposed methods and their application potential. The descriptions of experimental results are good and the drawn conclusions logical and convincing. However, there is one problem occurring through the manuscript: the lack of description of the experimental (measurement) methodology, in particular in relation to time measurements, which need to be clearly described (how many times repeated, statistical analysis of results, etc.). This is in particular important given the strong claims, e.g., on p. 36 (smthg. outperforms smthg. else) and the lack of theoretical analysis of time and space complexity (i.e., asymptotic analysis) that would support such claims. Additionally, the figures with results should be also complemented by tables clearly showing the numbers visualized in the figures.

There are also several minor problems with the manuscript. Although written well and easy to follow, some problems can be found. For example, some abbreviations are defined multiple times (e.g., CPM on p. 12, then on page 15, etc.), essentially identical definitions of horizontal and vertical DBs on page 16, etc. Nevertheless, these problems are not critical in the context of the thesis.

17. listopadu 2172/15
708 00 Ostrava-Poruba
Czech Republic

phone: +420 597 325 898
attendant: +420 597 321 111
ID data mailbox: d3kj88v

IČO: 61989100
VATIN: CZ61989100

email: pavel.kromer@vsb.cz
www.vsb.cz

In summary, it can be concluded that the topic of the research is in line with current research trends, the candidate presents a good number of interesting results published in relevant scientific journals (including Q1 and D1 ranked ones) . Therefore, it is my pleasure to **recommend the thesis for defense** and Mr. Huy Minh Huynh **for the academic degree of *doctor philosophiae* (Ph.D.)**.

Prof. Ing. Pavel Krömer, Ph.D.
10. 8. 2025

17. listopadu 2172/15
708 00 Ostrava-Poruba
Czech Republic

phone: +420 597 325 898
attendant: +420 597 321 111
ID data mailbox: d3kj88v

IČO: 61989100
VATIN: CZ61989100

email: pavel.kromer@vsb.cz
www.vsb.cz