



Review of the Doctoral Thesis

TITLE: MINING MULTI-LEVEL SEQUENTIAL PATTERNS

AUTHOR: ING. MICHAL ŠEBEK

The focus of the thesis is the mining of the sequential patterns using the multi-level approach. The thesis has three main parts that corresponds with the defined goals. The first part provides an insight into problematics of pattern mining, sequential pattern mining, multi-levels pattern, taxonomies, etc. State of the art is also good and contains all relevant sources. I appreciate that all examples use the same data and the results may be compared. The second part consists of the description of the main problem that is studied in the thesis. And the third part is focused on the experimental evaluation and comparison of the algorithm with others.

The formal style of the thesis is good. The quality of the English is also excellent. I did not find any typo or error and only very few strange sentences. Due to the high frequency of the definition, mathematic equation it is not easy to read the thesis, but all information are well defined and used appropriately. I found a mistake in the tables where the results for 1000k dataset are presented as a result for 100k.

The description of the main problem is well written. The problems well as the proposed solution are described and explained. The proposed novel algorithm is based on the interesting presumptions, but they are described in detail, and the following experiment proves that the setting of the algorithm is valid.

The experimental evaluation of the algorithm is really large and the artificial dataset used are sufficient. The maximal size of the dataset is 1M of transactions which is not much for real-world application. The experimental evaluation on the real dataset is tiny.

The overall evaluation of the thesis:

The thesis deals with the mining sequential pattern with extension o the multi-level organization and taxonomies. The quality of the thesis from the formal point of view is very good. The description of the algorithm that is the main goal of the thesis is well described, and all aspects are discussed in detail. The experimental evaluation of the algorithm is done on artificial dataset as well as on the real-world dataset. The results prove that the algorithm is well designed and applicable on such datasets. The size of the dataset is small, and the experiments with larger data will be better to prove the usability of the algorithm in today's big data world.



Fakulta elektrotechniky a informatiky VŠB-TUO
17. listopadu 15, Ostrava - Poruba, 708 33
ID datové schránky d3kj88v, tel.: +420 597 325 252, fax: +420 596 919 597
spojovatelka:+420 597 321 111, e-mail: sekretariat.feii@vsb.cz, www.feii.vsb.cz
IČ: 61989100, DIČ: CZ 61989100

The publication activities of the student were small. The list of publication includes ten publication where only one is in the journal that is not indexed in Scopus or WoS journal index. The other publication is in conference and books.

When I take into account all aspects of the presented work and include all my remarks, I think that the work done by the students is very interesting and promising for the future research and

I recommend the thesis for defense.

Questions:

1. May you please explain why the processing times for 1000k dataset are smaller than for 750k?
2. Because all algorithm were implemented by the author (from the description in the experimental part), was the algorithm GSP and PrefixSpan optimized in the same manner as the proposed algorithm?
3. Is there a real application for your algorithm?
4. What is the limitation of your algorithm for practical implementation?
5. What was the largest bottleneck of your algorithm?

Ostrava, 19.3.2017

doc. Ing. Jan Platoš, Ph.D.