Research of Occupancy-Based Skyline Pattern Mining

Kai Zhang[1,2] *, Kun Hu[1,3]
[1]Key Laboratory of Technology on Data link
[2]China Electronics Technology Group Corporation (CETC), 20th Institute, Xi'an, China
[3]School of Computer Science, Northwestern Polytechnical University, Xi'an, China
*E-mail: brantson@126.com

Objectives:

In recent years, it has been proposed to use occupancy to measure the completeness of the pattern in the transaction database. The results of DOFRA mining are frequent and highly complete patterns, but the minimum occupancy threshold is not easy to set. An effective way to solve this problem is to study an occupancy-based skyline pattern mining algorithm to design an effective occupancy-based skyline pattern mining algorithm. To solve the above limitations, this article proposes a method based on an efficient PEL structure Skyline pattern mining algorithms (SOPM and SOPM-occumax).

Methods:

This paper proposes a skyline mining algorithm based on SOPM-occubound by designing an efficient pattern extended list (PEL) structure to replace the occubound calculation in DOFRA algorithm. Two PEL structure-based Skyline algorithms, SOPM(candidate set) and SOPM-occumax (no-candidate set), are used to mine occup-based Skyline patterns (OBSPs).The advantage of SOPM is that the unreasonable threshold setting leads to the huge difference of mining data. It can effectively find the most representative pattern in the two dimensions of support and occupation, and dominate other patterns.The results of some tests on real and simulated data sets show that SOPM based on PEL structure is superior to SOPM-occubound in running time and memory usage, and SOPM-occumax is not more efficient and effective in running time than candidate sets in improving the efficiency and scalability of the algorithm.

Modeling:

- Given a transaction database **D**

TID	*Transactions*	*Length*
T1	A B C	3
T2	A B C D	4
T3	A C	2
T4	C D	2
T5	E	1

- Construct a pattern extended list (PEL) structure

{C}	
sup_count=4	
length=1	
Tid	SURPLUS
T1	2
T2	3
T3	1
T4	1

{A}	
sup_count=3	
length=1	
Tid	SURPLUS
T1	1
T2	2

{B}	
sup_count=2	
length=1	
Tid	SURPLUS
T2	1

{D}	
sup_count=2	
length=1	
Tid	SURPLUS

{E}	
sup_count=1	
length=1	
Tid	SURPLUS

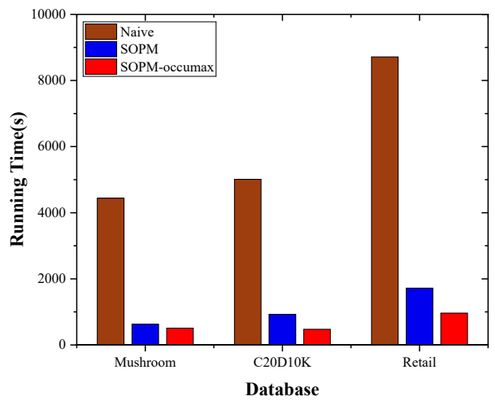
Figure 1. The PEL structure includes the support of the pattern, the supporting database corresponding to the Itemset, and the schema residual value for each transaction of the surplus, namely (tid,surplus). The PEL structure of the Itemset is successfully built in descending order of support.

- A pruning theorem called ExVaule is proposed, which can predict the value of the parent itemset

$$ExValue(X)=\frac{|D_X|\times|X|+\phi(X)}{\sum_{T_p\in D_X}T_p}$$

Results:

- Runing Time



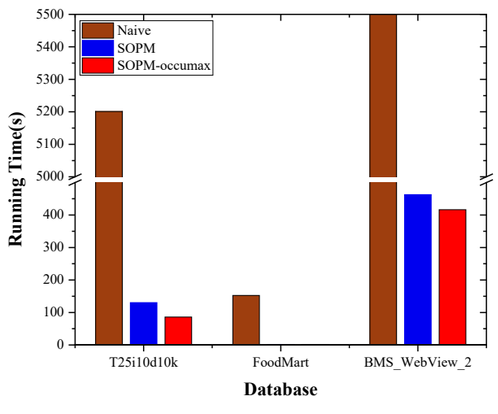


Figure 2. The proposed algorithm SOPM-occumax has the best running time of the three algorithms. Because the algorithm of SOPM-occumax hits the pruning algorithm ahead of time, unnecessary itemset searches are reduced. In addition, SOPM-occumax does not generate a candidate itemset, so SOPM-occumax has better mining efficiency.

- Memory Usage

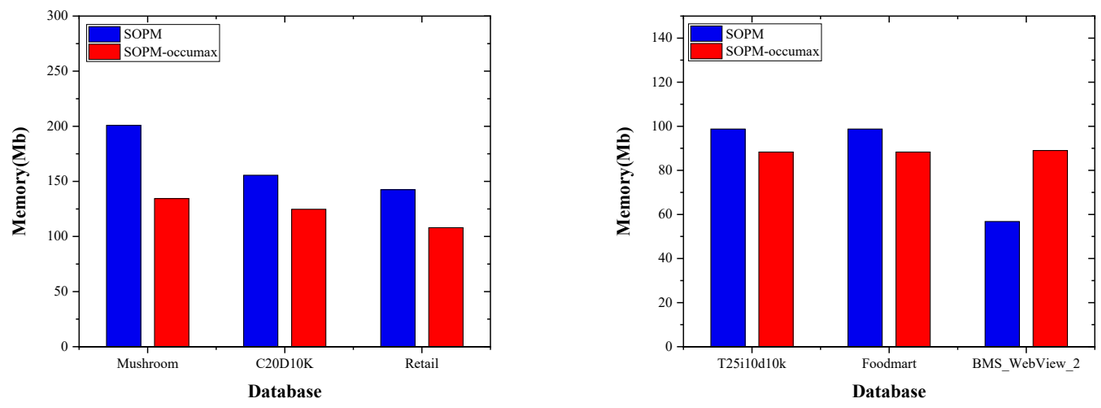


Figure 3. We can see that the memory usage of SOPM in mushroom, c20d10k, retail, t25i10d10k and FoodMart is higher than that of SOPM-occumax algorithm. In mushroom, the memory usage of SOPM is about twice that of SOPM-occumax. Similarly, the memory usage gap between the two algorithms is larger in the dense dataset than in the sparse dataset. In BMSWebView2, SOPM uses less memory than the latter, but the gap is always within an order of magnitude, and SOPM-occumax runs faster than the former, so some small fluctuations are acceptable..

Conclusion:

This paper designs a PEL structure and proposes SOPM-Occumax algorithm. After experimental tests, it is proved that this algorithm is an effective occupancy-based Skyline pattern mining algorithm.

Acknowledgements:

The work presented here was supported financially by the Data link Technology Key Laboratory open fund program(CLDL-20201102)