Incorporating Occupancy into Frequent Pattern Mining for High Quality Pattern Recommendation

Lei Zhang

stone@mail.ustc.edu.cn

Lab. of Semantic Computing and Data Mining

University of Science and Technology of China

Outline

- Problem statement
- Motivating application
  - Demo
- Problem solution
  - Upper Bound Estimation
- Experiments
- Conclusion

Problem Statement

- Frequent pattern mining
  - *Support*
  - *Frequent patterns*

Tran. No.	Items
1	A B C D E
2	A B C
3	A B C
4	A B C
5	A C D E F

The support of {ABC} = 4/5
{ABC} is frequent when α = 0.5

Problem Statement

- Propose a new interestingness measure
- *Occupancy*

Tran. No.	Items	Occupancy in this Trans.
1	A B C D E	3/5
2	A B C	3/3
3	A B C	3/3
4	A B C	3/3
5	A C D E F	

$$occu(ABC) = \frac{1}{4} \times (\frac{3}{5} + \frac{3}{3} + \frac{3}{3} + \frac{3}{3}) = \frac{9}{10}$$

Problem Statement

- The property of occupancy
  - *Neither monotone nor anti-monotone*

Transaction No.	Items
1	A B C D E
2	A B C
3	A B C
4	A B C
5	A C D E F

$$occu(ABCD) = \frac{4}{5} < occu(ABC) = \frac{9}{10}$$

$$occu(ABCDE) = \frac{5}{5} > occu(ABC) = \frac{9}{10}$$

Problem Statement

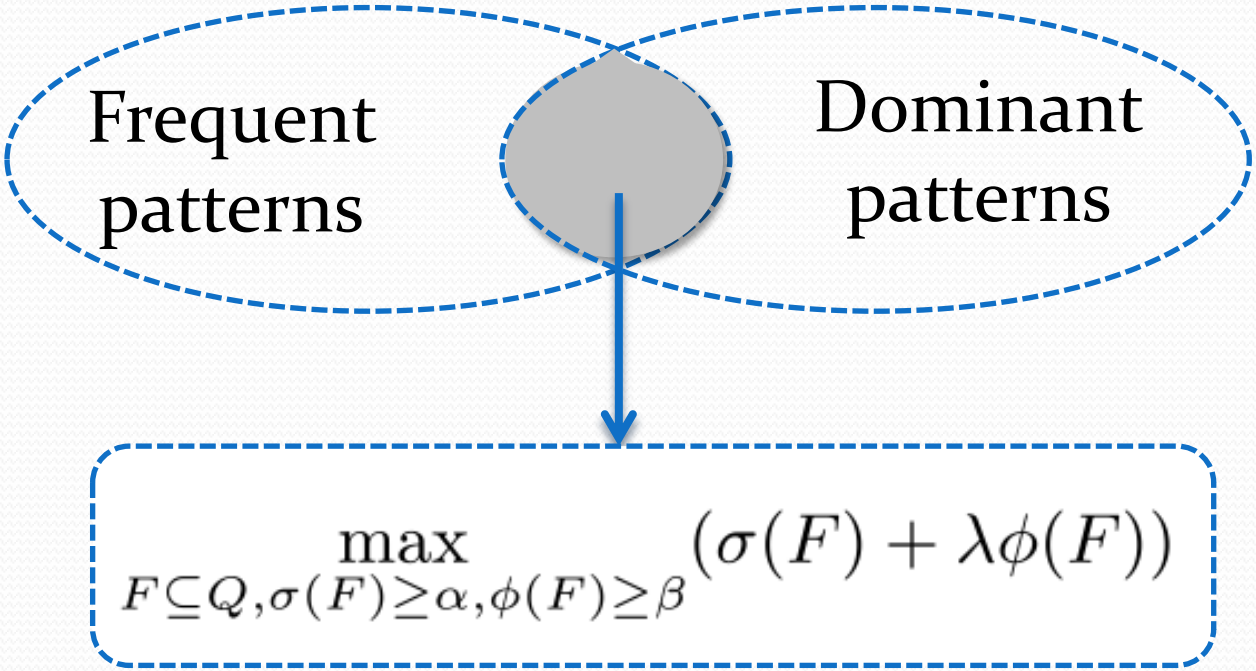
- *Dominant patterns*, whose occupancy is bigger than a parameter β

Transaction No.	Items
1	A B C D E
2	A B C
3	A B C
4	A B C
5	A C D E F

The occupancy of {ABC} ≈ 0.85
{ABC} is dominant when β = 0.5

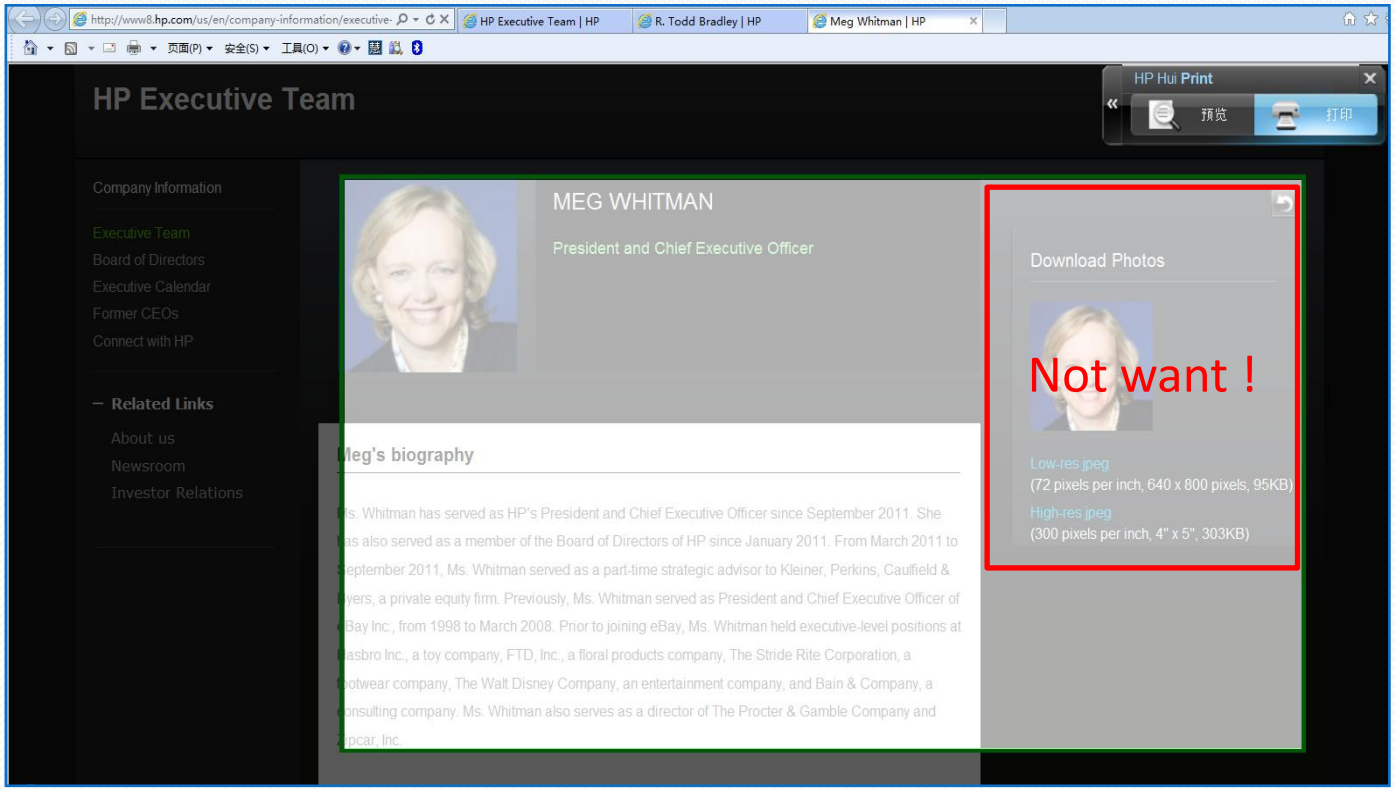
Problem Statement

- Task: **Top qualified pattern mining** for recommendation
  - *Qualified patterns*, which are both frequent and dominant
  - *Quality value of a pattern,* $q(X) = \sigma(X) + \lambda\phi(X)$



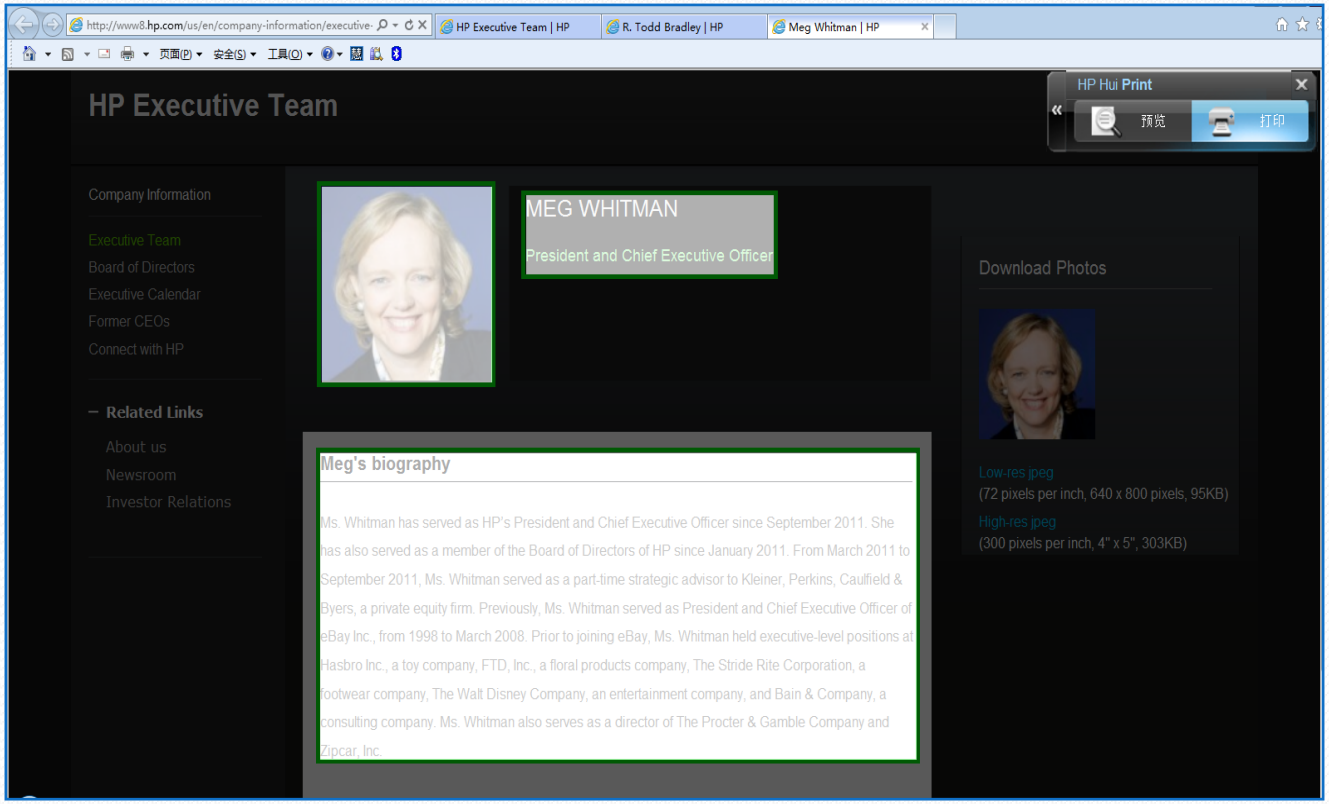
Motivating Application

- Web page printing: Smart Print
  - Manually adjustment on the print-areas required after the first-round recommendation



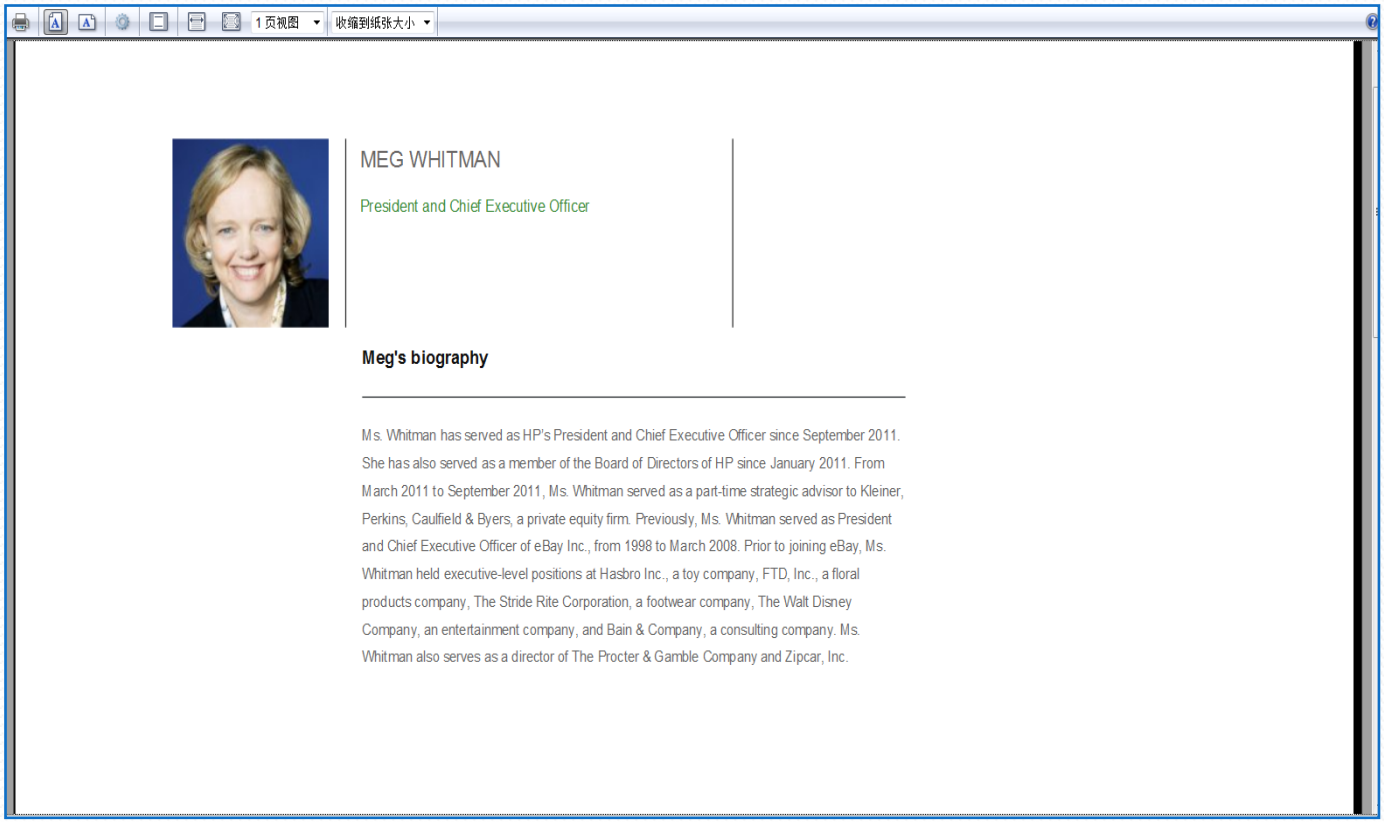
Motivating Application

- Web page printing: Smart Print
  - Tedious selections



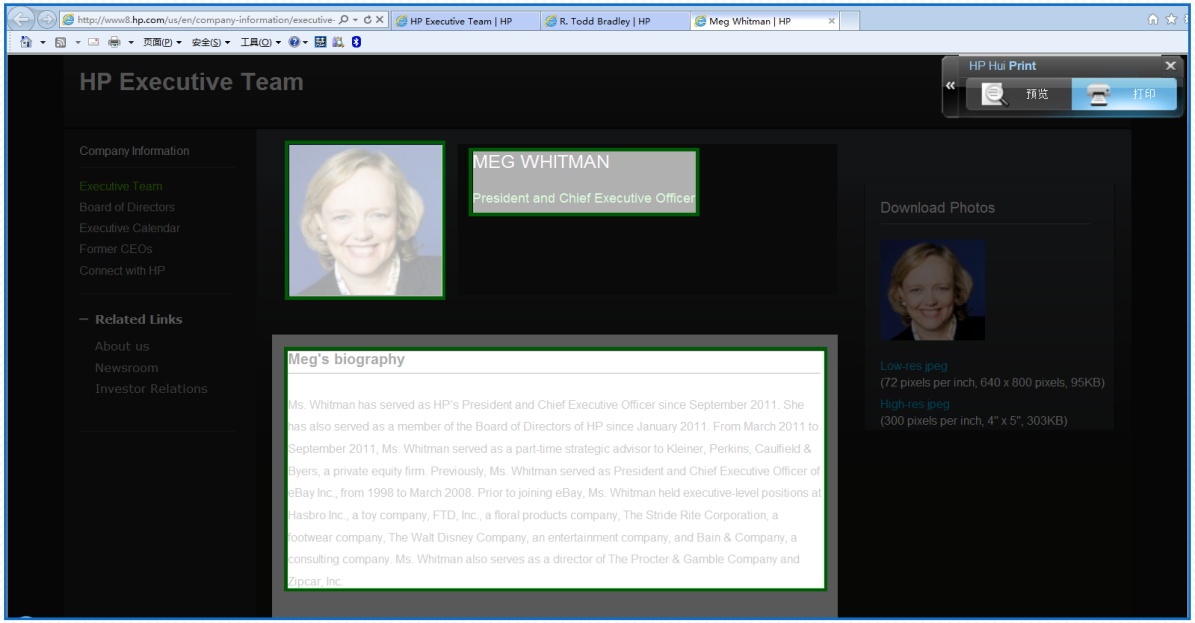
Motivating Application

- Web page printing: Smart Print
  - Print preview



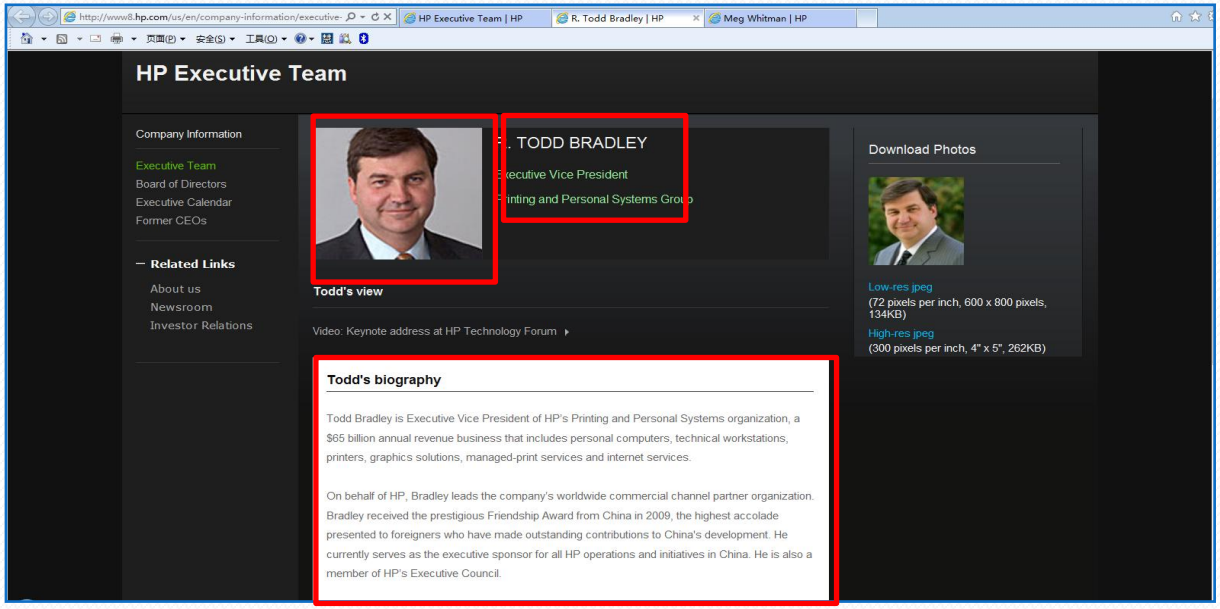
Motivating Application

- Goal
  - Only one click to get the exact print areas by providing more accurate print-area recommendation in Smart Print
- Solution
  - Leverage the print logs from all the users for more accurate print-area recommendation



Motivating Application

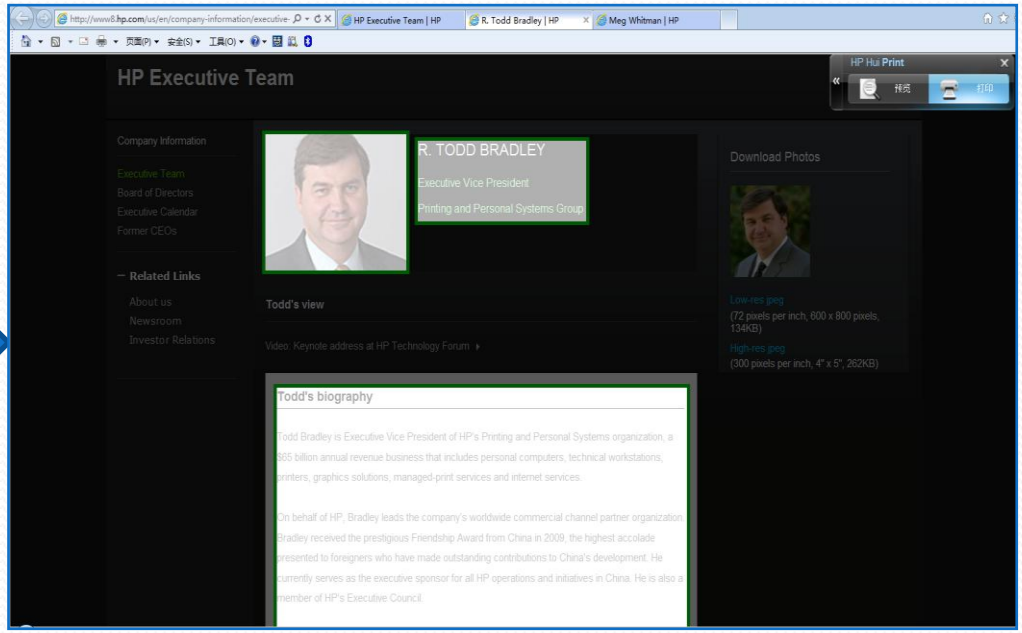
- Goal
  - Only one click to get the exact print areas by providing more accurate print-area recommendation in Smart Print
- Solution
  - Leverage the print logs from all the users for more accurate print-area recommendation



Motivating Application

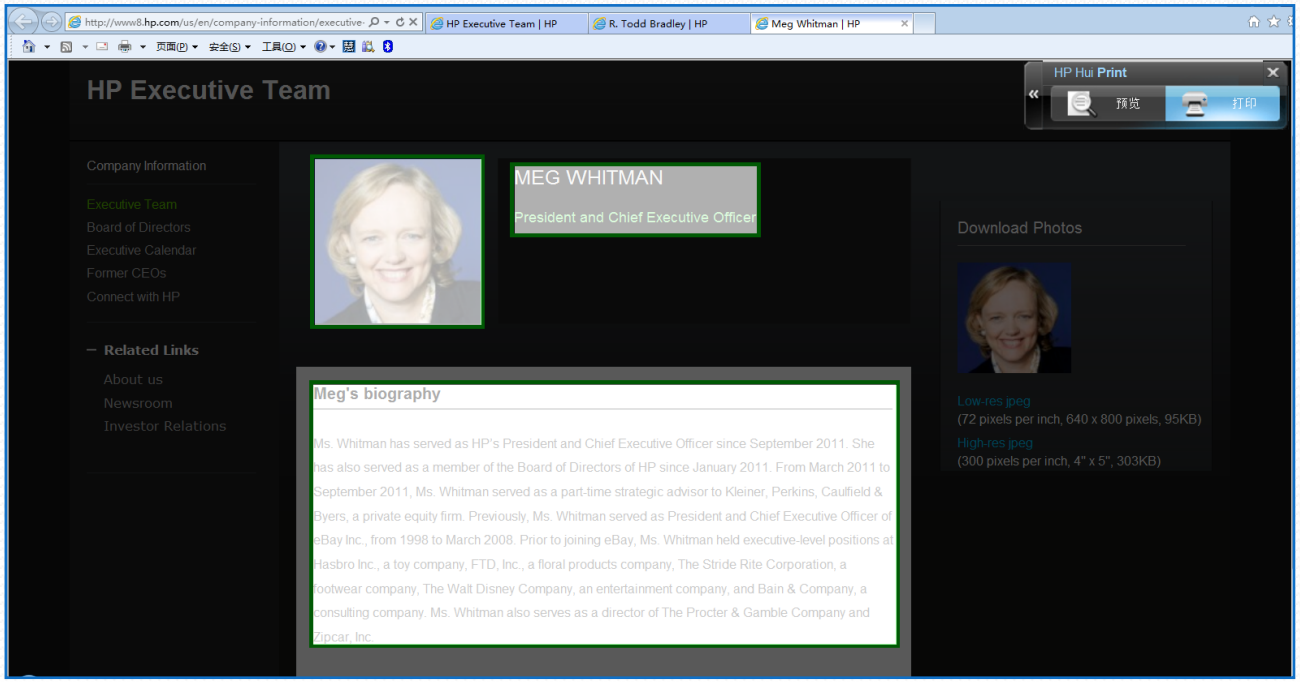
- Demo

Print Log Database



Motivating Application

- One piece of print log



- A selected print-area = a clip = an item
- A piece of print log = a set of clips

Motivating Application

- Print log database

Print Log Database

Transaction No.	Items
1	A B C D E
2	A B C
3	A B C
4	A B C
5	A C D E F

Transaction Database

Motivating Application

- Formulate the task of print-area recommendation as a pattern mining problem

Given the transaction database of print logs and a query Web page

- Identify all the candidate clips inside the query page, denoted by Q
- Find a subset of Q for recommendation

The interestingness measure for patterns

- ***Support***
- ***Occupancy***

Motivating Application

- Support

Transaction No.	Items
1	A B C D E
2	A B C
3	A B C
4	A B C
5	A C D E F

The support of {ABC} = 4/5

The more frequently a pattern appears in the database, the more number of users select this set of clips for printing.

Motivating Application

- Occupancy

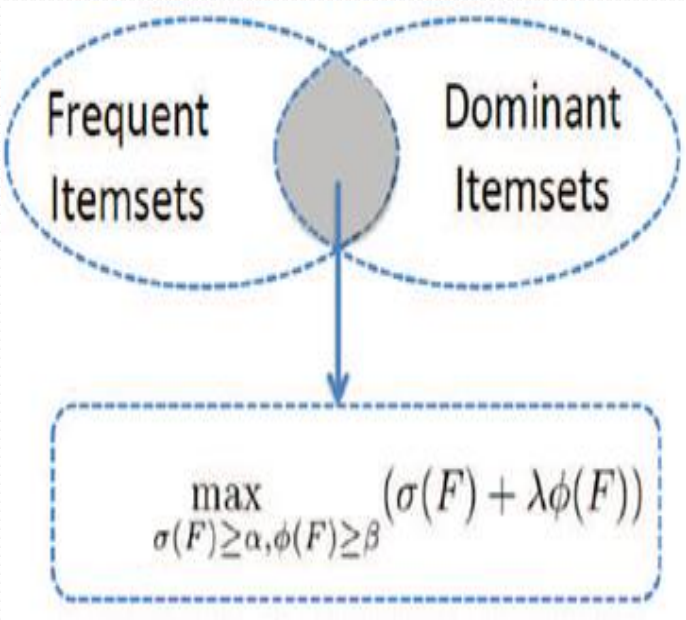
Transaction No.	Items
1	A B C D E
2	A B C
3	A B C
4	A B C
5	A C D E F

The occupancy of {ABC} = $\frac{1}{4} \times (\frac{3}{5} + \frac{3}{3} + \frac{3}{3} + \frac{3}{3}) = \frac{9}{10}$

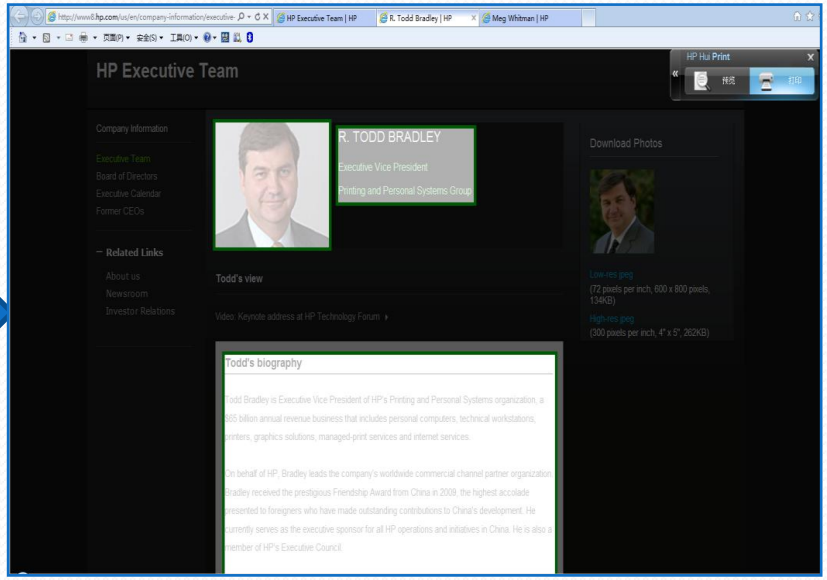
The bigger the occupancy is, the more complete the recommendation is.

Motivating Application

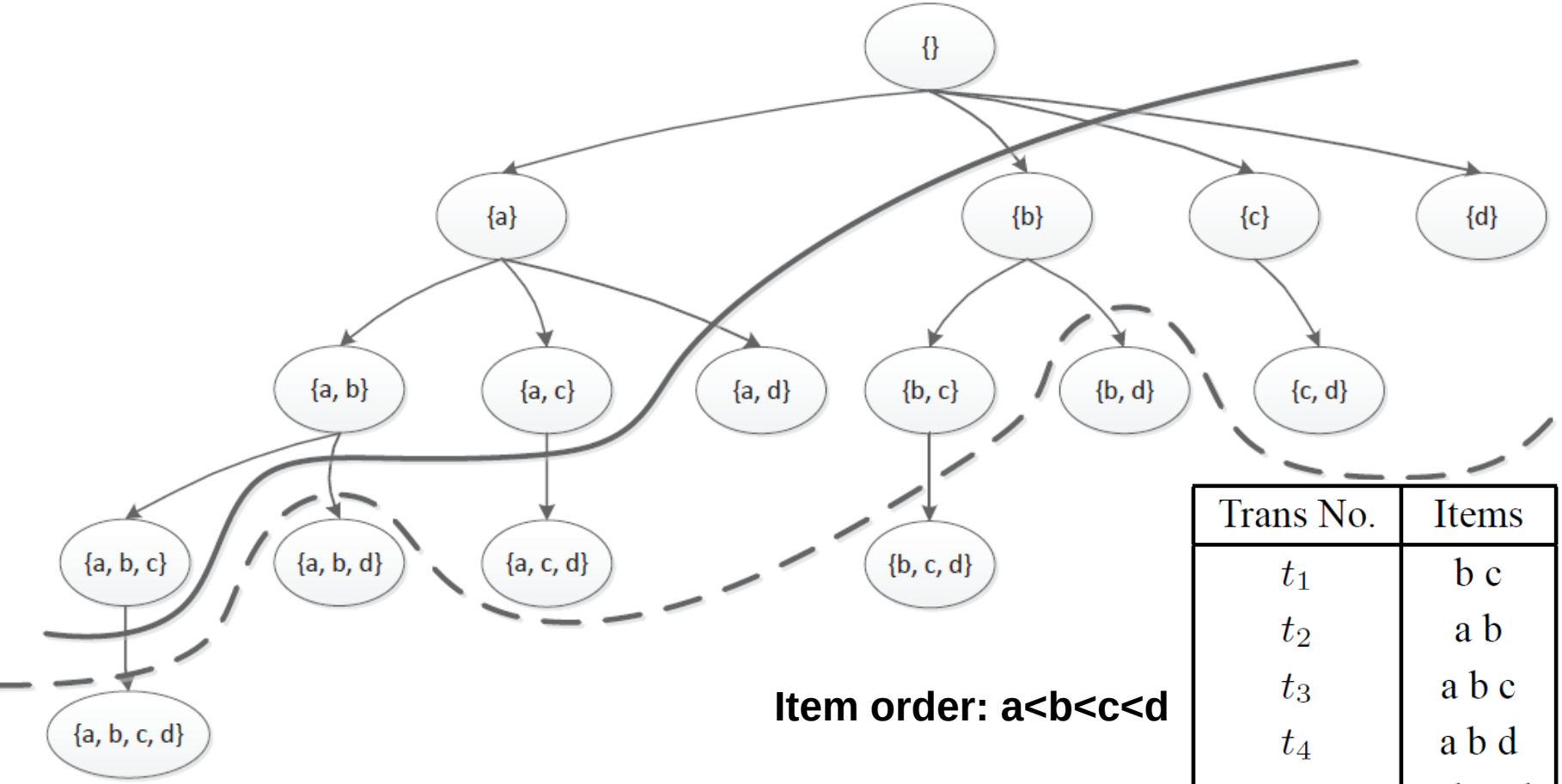
- **Mining top qualified pattern** for recommendation



3 clips



Solution

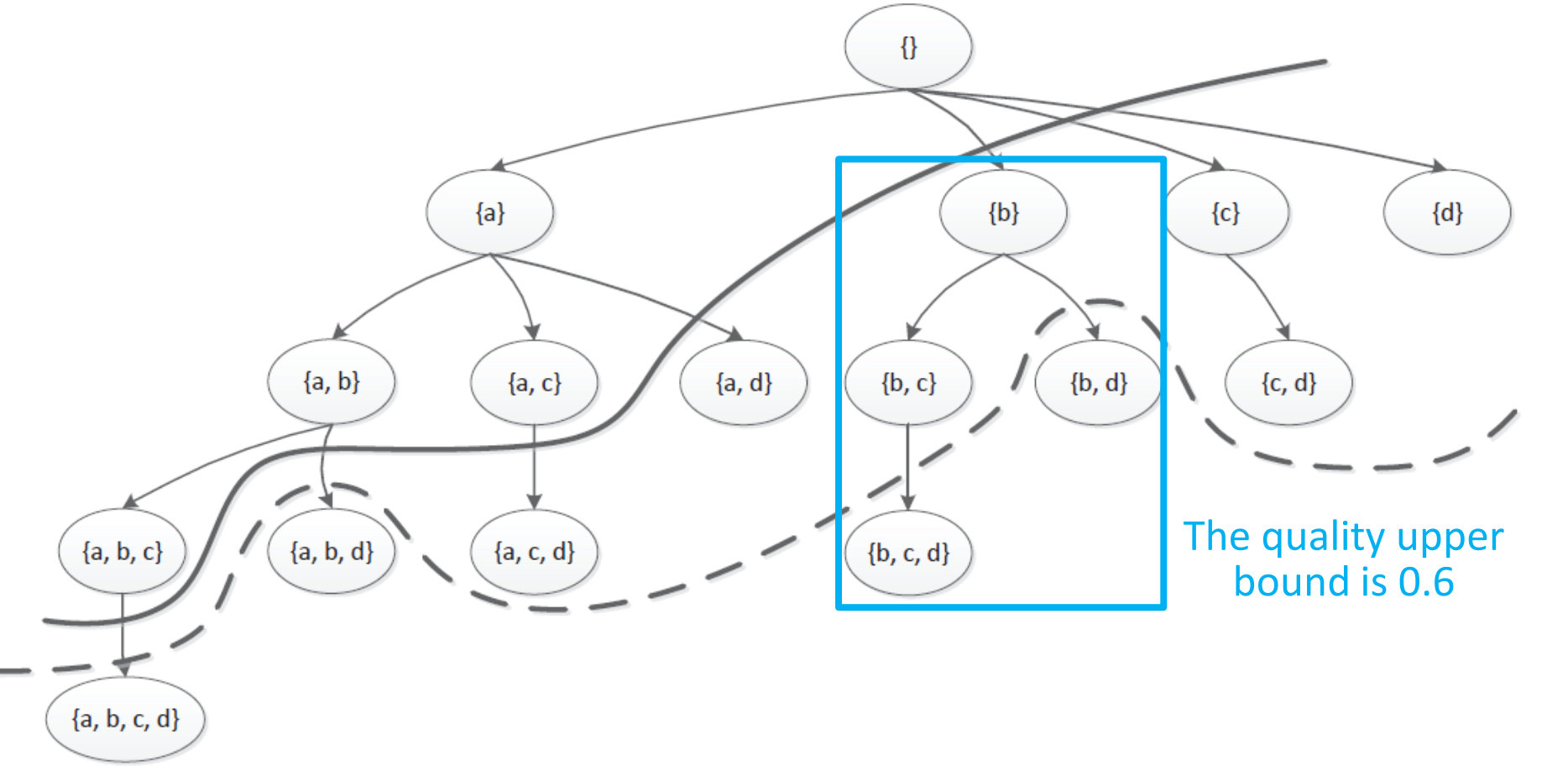


Trans No.	Items
t_1	b c
t_2	a b
t_3	a b c
t_4	a b d
t_5	a b c d

Item order: a<b<c<d

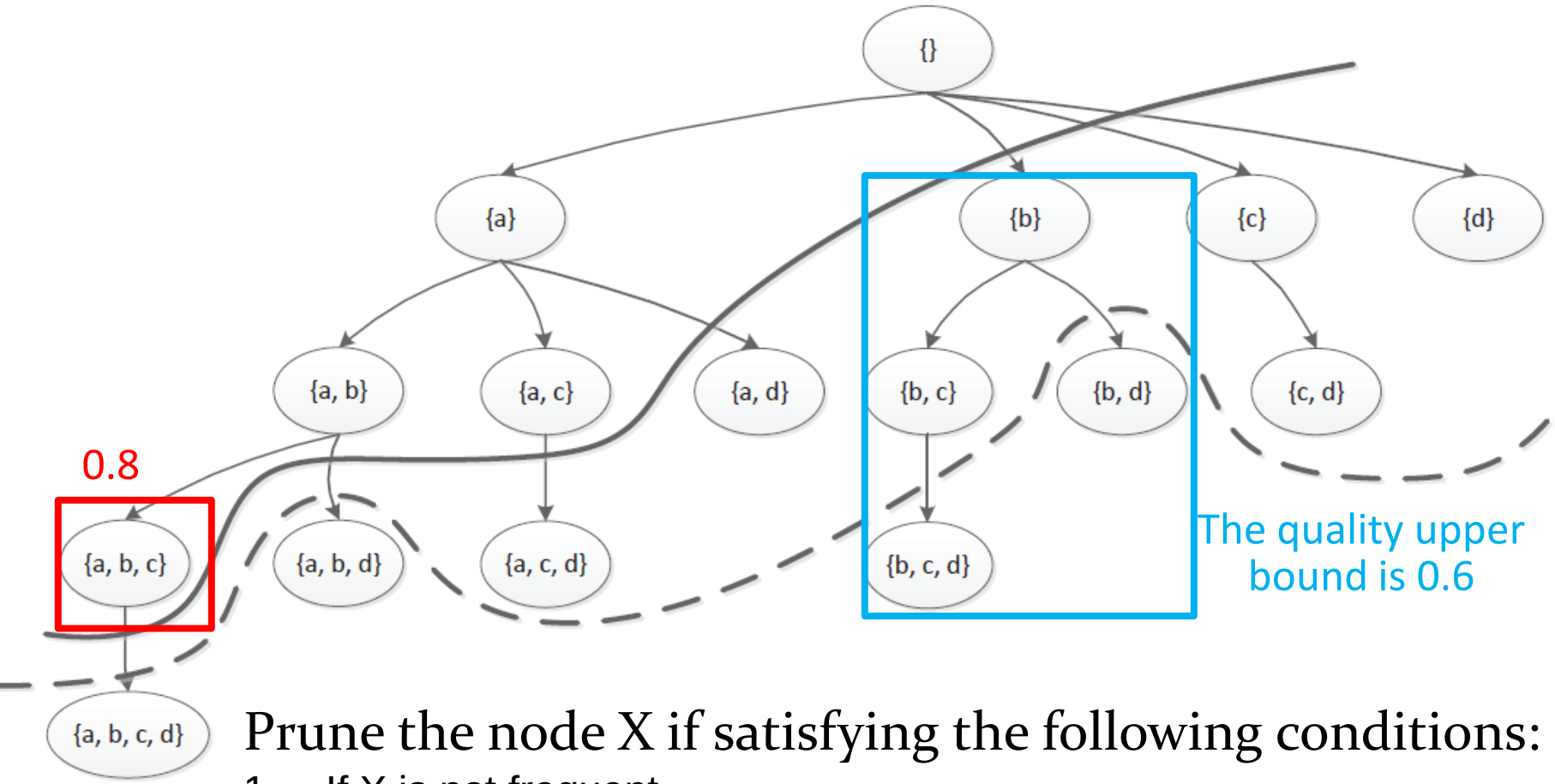
Set-enumeration tree

Solution



For a subtree we can give the upper bounds of the occupancy and quality values for all the nodes in this subtree.

Solution



Prune the node X if satisfying the following conditions:

1. If X is not frequent,
2. If the upper bound on **occupany** for the subtree rooted at **X** is smaller than β
3. If the upper bound on **quality** for the subtree rooted at X is smaller than the current maximal quality value in the search process

Upper Bound Estimation

- Problem formulation
  - Input: an itemset **X** and the supporting transactions of **X**
  - The task: estimate the upper bound of **occupancy** for all frequent supersets of **X**

Upper Bound Estimation

- EL,TL to represent the supporting transactions of X

t_{id}	Items
t_1	b c e f
t_2	a b e
t_3	a b c
t_4	a b d f
t_5	a b c d

t_{id}	X	Extension
t_1	b	c e f
t_2	b	e
t_3	b	c
t_4	b	d f
t_5	b	c d

t_{id}	EL	TL
t_1	3	4
t_2	1	3
t_3	1	3
t_4	2	4
t_5	2	4

Upper Bound Estimation

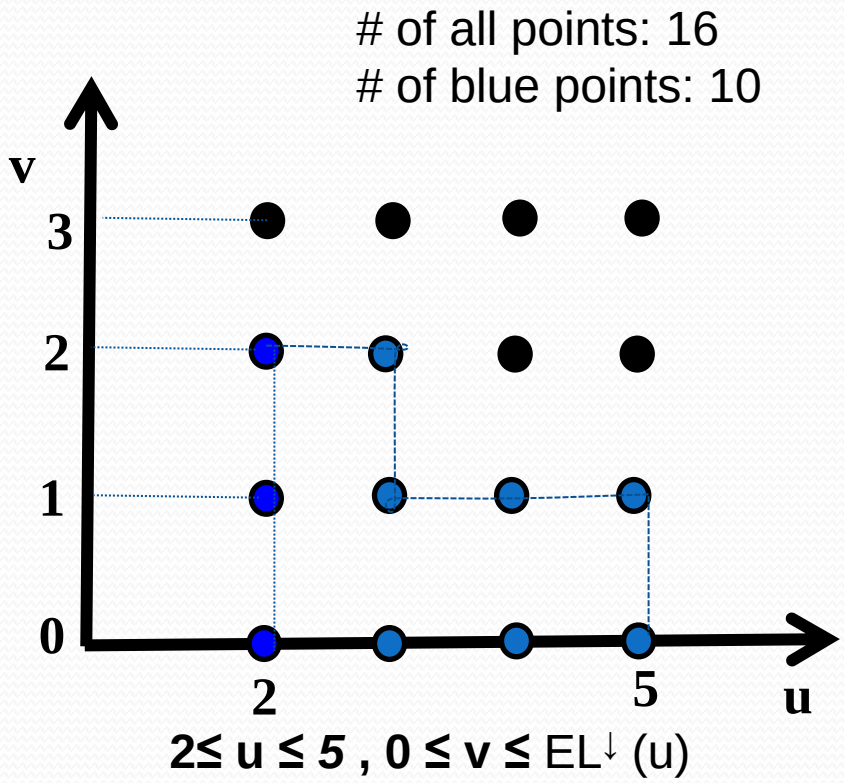
- Problem reformulation
  - Input : an itemset **X** and **EL, TL** of **X**,
  - The task is to estimate the upper bound of **occupancy** for all frequent supersets of **X**
- The basic idea
  - First, Let **X'** be any frequent superset of **X**, suppose $|T_{X'}|=u$ and the extension length $|X'|-|X|=v$, then we will propose **F(u, v, EL,TL)** such that $occu(X') <= F(u,v,EL,TL)$
  - Next, we can enumerate all posible **u,v** for the above F function and then get the upper bound for any superset of X, $occu(X') <= max_{u,v} F(u,v,EL,TL)$

Upper Bound Estimation

- The relationship between u and v
  - When u is fixed, the range of v is [o, $EL^{\downarrow}$ (u)] and u ϵ [fre_min, $|T_X|$]
  - EL = <3,1,1,2,2>
  - $EL^{\downarrow}$ = <3,2,2,1,1>

t_{id}	EL	TL
t_1	3	4
t_2	1	3
t_3	1	3
t_4	2	4
t_5	2	4

$EL^{\downarrow}$ (1)=3
$EL^{\downarrow}$ (2)=2,
$EL^{\downarrow}$ (3)=2,
$EL^{\downarrow}$ (4)=1,
$EL^{\downarrow}$ (5)=1
f re_min=2
$|T_{\{b\}}|$=5



$$\max_{u,v} F(u,v,EL,TL) =$$

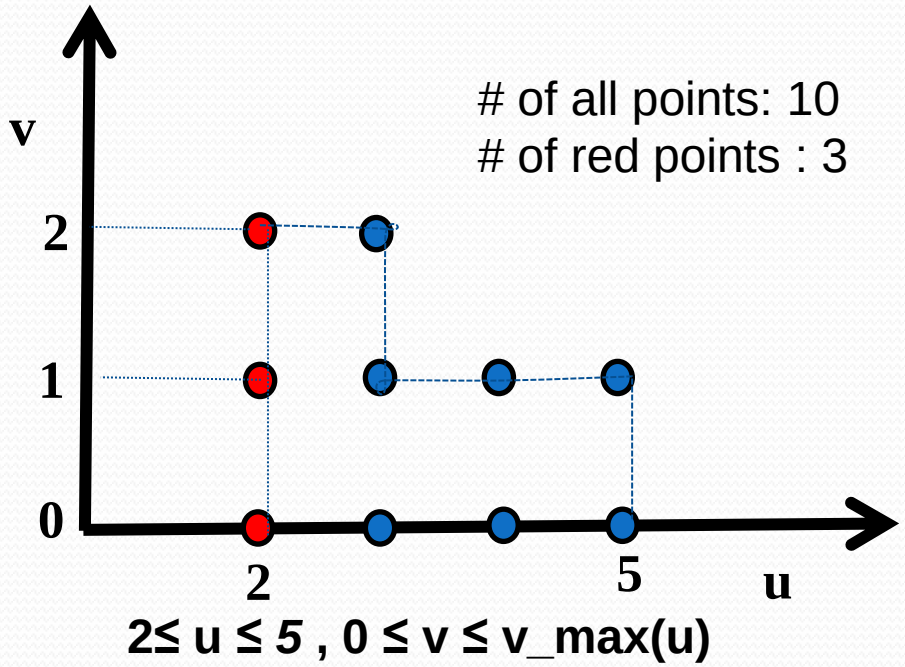
$$\max_{fre_{\min} \le u \le |T_X|, 0 \le v \le EL\downarrow(fre_{\min})} F(u,v,EL,TL)$$

Upper Bound Estimation

- The property of F with respect to u
  - When the F function is monotonically decreasing with respect to u, i.e., F(u+1,v, EL,TL) <= F(u,v,EL,TL), we have

$$\max_{fre_{\min} \le u \le |T_X|, 0 \le v \le EL\downarrow(fre_{\min})} F(u,v,EL,TL)$$

$$= \max_{0 \le v \le EL\downarrow(fre_{\min})} F(fre_{\min},v,EL,TL)$$



Upper Bound Estimation

- occ(X') <= F(2,2,EL,TL)

t_{id}	TL	|b|	EL	Step-1	Step-2	Step-3
t_1	4	1	3	√	3/4	√
t_2	3	1	1			
t_3	3	1	1			
t_4	4	1	2	√	3/4	√
t_5	4	1	2	√	3/4	

$$F(u,v,EL,TL) = \frac{1}{u} \times \max_{l_1,\ldots,l_u, EL(l_i)\geq v} \sum_{i=1}^{u} \frac{|X|+v}{TL(l_i)}$$

Upper Bound Estimation

- The upper bound for all frequent supersets of {b}

$$ub(\{b\}) = \max_{0 \le v \le EL\downarrow(2)} F(2, v, EL, TL)$$

$$F(2,0,EL,TL) = 1/2 \times (1/3 + 1/3) \approx 0.33$$

$$F(2,1,EL,TL) = 1/2 \times (2/3 + 2/3) \approx 0.66$$

$$F(2,2,EL,TL) = 1/2 \times (3/4 + 3/4) = 0.75$$

$$ub(\{b\}) = \max_{0 \le v \le 2} F(2, v, EL, TL)$$

$$= \max\{0.33, 0.66, 0.75\} = 0.75$$

Upper Bound Estimation

- The efficient upper bound

t_{id}	TL	|b|	EL	Step-1	Step-2	Step-3
t_1	4	1	3	√	4/4	√
t_2	3	1	1	√	2/3	
t_3	3	1	1	√	2/3	
t_4	4	1	2	√	3/4	√
t_5	4	1	2	√	3/4	

$$F(u, EL, TL) = \frac{1}{u} \times \max_{l1,\ldots,lu} \sum_{i=1}^{u} \frac{|X| + EL(li)}{TL(li)}$$

$$ub(\{b\}) = F(2, EL, TL) = 1/2 \times (4/4 + 3/4) = 0.875$$

Upper Bound Estimation

- The tradeoff between bound tightness and computational efficiency

$$occ(X') \le \max_{0 \le v \le EL\downarrow(fre_{\min})} F(fre_{\min}, v, EL, TL)$$

$$\mathrm{F}(\mathrm{u}, \mathrm{v}, \mathrm{EL}, \mathrm{TL}) = \frac{1}{u} \times \max_{l1,\ldots,lu, EL(li) \ge v} \sum_{i=1}^{u} \frac{|X| + v}{TL(li)}$$

$$occ(X') \le \max_{v1,\ldots,vm} F(fre_{\min}, v_k, v_{k+1}, EL, TL)$$

$$\mathrm{F}(\mathrm{u}, \mathrm{v_k}, \mathrm{v_{k+1}}, \mathrm{EL}, \mathrm{TL}) = \frac{1}{u} \times \max_{l1,\ldots,lu, EL(li) \ge v_k} \sum_{i=1}^{u} \frac{|X| + v_{k+1}}{TL(li)}$$

Solution

- Three different upper bound functions

Upper bound	Bound efficiency	Bound tightness	# of searched node
F	fast	loose	large
F’	slow	tight	small
F^	tradeoff	tradeoff	tradeoff

Experiments

- Does the concept of *occupancy* help to improve the recommendation performance?

- Does our algorithm with the proposed pruning strategy can significantly reduce time complexity?

Experiments

- Evaluation on effectiveness
  - Ground truth
    - 2000 Webpages from 100 printworthy Websites
  - Evaluation method
    - Leave one out cross validation for each webpage
  - Evaluation measure

$$P = \frac{|A_G \cap A_R|}{|A_R|}, R = \frac{|A_G \cap A_R|}{|A_G|}, F1 = 2 \times \frac{P \times R}{P + R},$$

Experiments

- Evaluation on effectiveness

λ	P(%)	R(%)	F1(%)
0.0	90.04	82.15	79.79
0.5	89.67	92.84	88.78
1.0	90.77	94.74	91.3
2.0	91.63	95.96	92.81
4.0	92.65	96.31	93.6
5.0	92.81	96.3	93.64
6.0	93.23	96.19	**93.8**
8.0	93.23	95.95	93.7
10.0	93.34	95.84	93.71
$+\infty$	91.27	91.62	89.82
Average	91.76	93.91	91.12

frequency

increase by 14%

Best

frequency+ occupancy

decrease

The recommendation performance of our method (α=0.05 β=0.1)

Experiments

- Evaluation on effectiveness

α	*P(%)*	*R(%)*	*F1(%)*
0.0	85.85	96.02	88.74
0.05	90.28	92.2	89.81
0.1	90.6	92.84	**90.56**
0.2	90.65	92.12	90.05
0.3	89.5	91.02	88.88
0.4	87.11	87.57	85.48
0.5	82.0	81.75	79.06

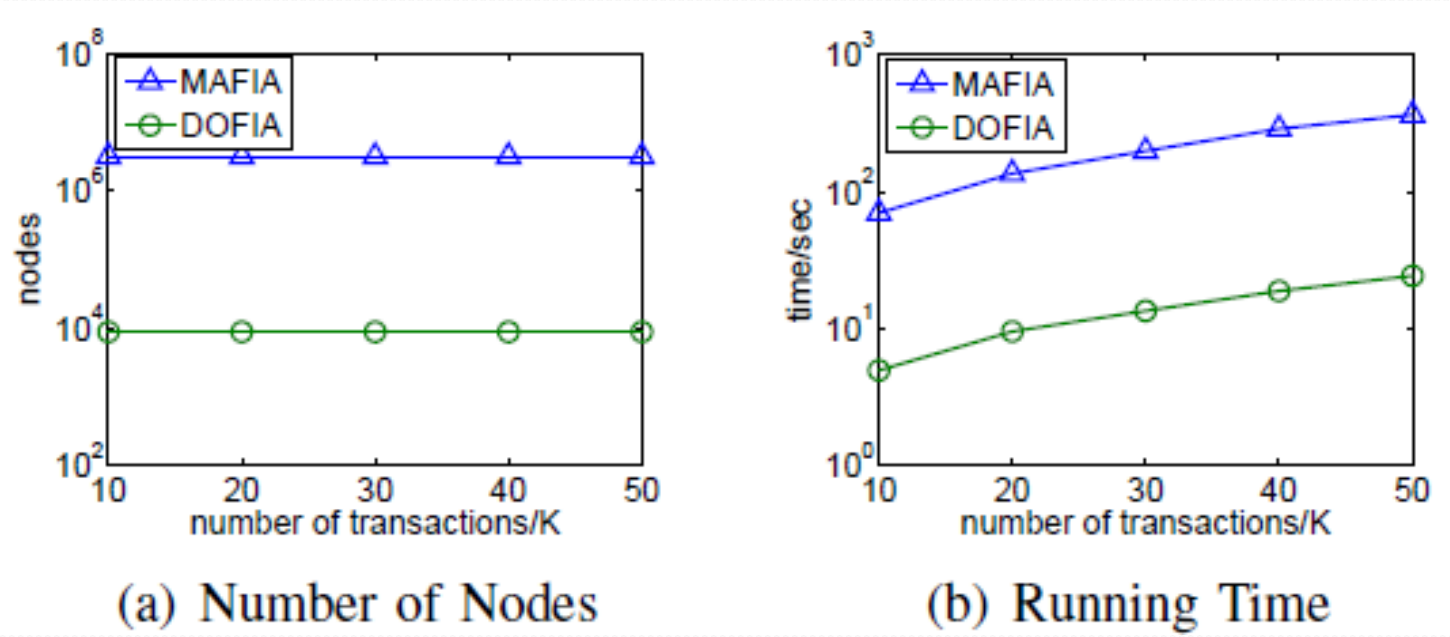
The recommendation performance of the maximal frequent pattern

Experiments

- Evaluation on efficiency
  - DOFIA
    - **Do**minant and **F**requent **I**temset Mining **A**lgorithm
  - Baseline
    - MAFIA
  - Data sets
    - Large synthetic datasets using IBM generator

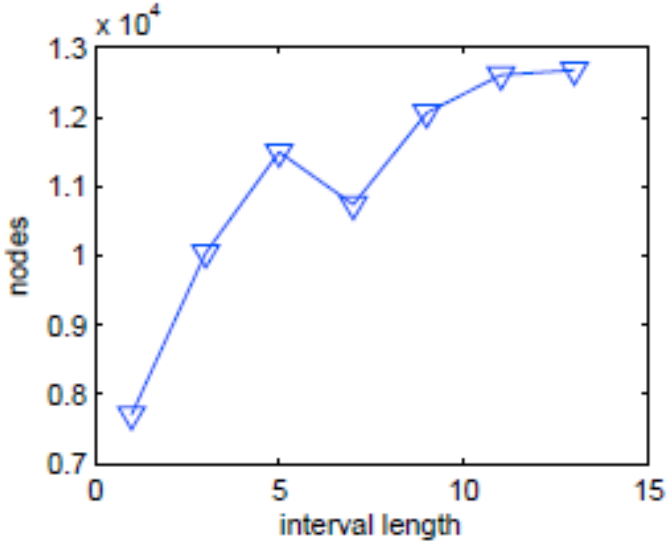
Experiments

- Evaluation on the number of transactions N

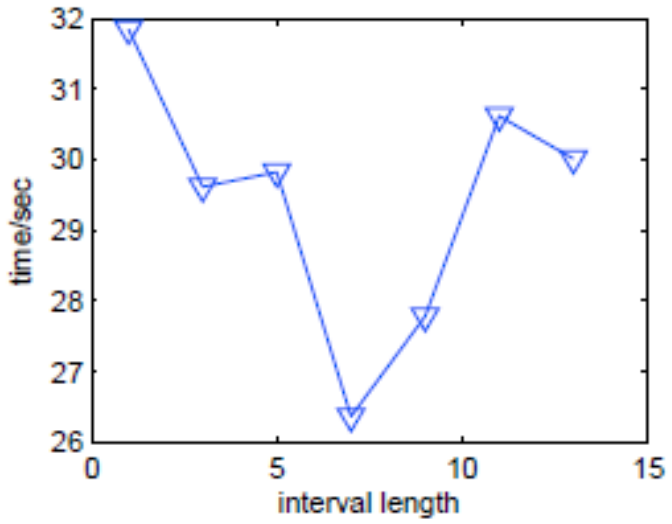


Experiments

- Evaluation on tradeoff between bound tightness and computational efficiency



(a) Number of Nodes



(b) Running Time

Conclusion

- Mining top qualified pattern task
- Motivation application: web print recommendation
  - *Frequency* to measure pattern popularity
  - *Occupancy* to measure pattern completeness
- An efficient algorithm DOFIA for this problem
  - The efficient upper bound
  - The tightest upper bound
  - Tradeoff between bound efficiency and bound tightness
- The extension work
  - We have extended this concept of occupancy in sequential pattern mining
  - We have proposed a general framework for computing the bounds of any pattern measure (e.g. occupancy+utility+block constraints)

Thanks! Q & A