EFFICIENT MINING OF TOP CORRELATED PATTERNS BASED ON NULL-INVARIANT MEASURES

SANGKYUM KIM, MARINA BARSKY, JIAWEI HAN

DEPT OF COMPUTER SCIENCE
UNIV OF ILLINOIS AT URBANA-CHAMPAIGN

ECML-PKDD'11 (Sep 2011, Athens Greece)

PROBLEM DEFINITION

Transaction 1

Transaction 2

⋮

Transaction N

Correlated Pattern Mining

Correlation Analysis

Correlated Patterns in DBLP

Rank	Pattern	Support	Cosine
1	object, orient, database	748	0.19
2	sense, word, disambiguation	26	0.18
3	support, vector, machine	122	0.17
4	enforcement, law, coplink	7	0.16
5	nearest, neighbor, search	74	0.13
6	reverse, nearest, neighbor	23	0.13
7	server, sql, microsoft	25	0.12
8	retrieval, cross, language	187	0.11
9	model, relationship, entity	139	0.11
10	random, field, conditional	13	0.10

* From the paper titles in DB-DM-IR subset of DBLP dataset with minimum support 0.02%

Background (I)

- Many existing interestingness measures
  - Investigated over 20 measures (Tan et al KDD'02)
    - Showed 3 null-invariant measures
      - Confidence(MaxConf), Jaccard(Coherence), Cosine
  - Only a few null-invariant measures
    - AllConf, Coherence, Cosine, Kulc, MaxConf

Background (II)

- Null-(transaction) invariance is crucial for correlation analysis
- Lift and χ^2 are not null-invariant
- 5 null-invariant measures

(copied from Wu et al. DMKD'10)

Measure	Definition	Range	Null-Invariant
$\chi^2(a,b)$	$\sum_{i,j=0,1} \frac{(e(a_i,b_j)-o(a_i,b_j))^2}{e(a_i,b_j)}$	$[0,\infty]$	No
$Lift(a,b)$	$\frac{P(ab)}{P(a)P(b)}$	$[0,\infty]$	No
$AllConf(a,b)$	$\frac{sup(ab)}{max\{sup(a),sup(b)\}}$	$[0,1]$	Yes
$Coherence(a,b)$	$\frac{sup(ab)}{sup(a)+sup(b)-sup(ab)}$	$[0,1]$	Yes
$Cosine(a,b)$	$\frac{sup(ab)}{\sqrt{sup(a)sup(b)}}$	$[0,1]$	Yes
$Kulc(a,b)$	$\frac{sup(ab)}{2}(\frac{1}{sup(a)}+\frac{1}{sup(b)})$	$[0,1]$	Yes
$MaxConf(a,b)$	$max\{\frac{sup(ab)}{sup(a)},\frac{sup(ab)}{sup(b)}\}$	$[0,1]$	Yes

Null-transactions w.r.t. a and b

Null-invariant

Dataset	a, b	~a, b	a, ~b	~a, ~b	χ^2	Lift	AllConf	Coherence	Cosine	Kulc	MaxConf
D_1	10,000	1,000	1,000	100,000	90557	9.26	0.91	0.83	0.91	0.91	0.91
D_2	10,000	1,000	1,000	100	0	1	0.91	0.83	0.91	0.91	0.91
D_3	100	1,000	1,000	100,000	670	8.44	0.09	0.05	0.09	0.09	0.09
D_4	1,000	100	1,000	100,000	8173	9.18	0.09	0.09	0.29	0.5	0.91

Background (III)

- Generalized null-invariant measure
  - All five measures can be generalized
    - Using mathematically generalized mean
    - Wu et al (DMKD'10)

$$\mathbb{M}^k(X) = \mathbb{M}^k\left(P(a_2, \ldots, a_n|a_1), \ldots, P(a_1, \ldots, a_{n-1}|a_n)\right) = \sqrt[k]{\frac{sup(X)^k}{n}\left(\frac{1}{sup(a_1)^k} + \cdots + \frac{1}{sup(a_n)^k}\right)}$$

- Mining null-invariant measures
  - AllConf, Coherence
    - anti-monotonic (CoMine: Lee et al, ICDM'03)
  - MaxConf
    - monotonic
  - Kulc
    - not (anti)-monotonic
    - association upper bound (GAMiner: Wu et al, DMKD'10)
  - Cosine
    - not (anti)-monotonic

Our Contribution

- How to efficiently mine correlated patterns?
  - Challenge: Cosine and Kulc are not (anti)-monotonic
  - Propose two pruning properties that work for all 5 null-invariant measures
  - Develop two different correlated pattern mining algorithms
    - Mine (highly) correlated null-invariant patterns given a (positive) correlation threshold
    - Mine top-k correlated null-invariant patterns
      - Hard even for domain experts to find correct threshold value

Two-Step Correlated Pattern Mining

- <Step 1>
  - Mine frequent patterns
- <Step 2>
  - Select top correlated patterns

Toy Transactional DB

TID	Transaction
T1	1, 3, 4, 5, 6
T2	3, 5, 6
T3	2, 4
T4	1, 4, 5, 6
T5	3, 6
T6	2, 4, 5

1-item	Support
1	2
2	2
3	3
4	4
5	4
6	4

2-itemset	(sup,cos)
[1,4]	(2, 0.71)
[1,5]	(2, 0.71)
[1,6]	(2, 0.71)
[2,4]	(2, 0.71)
[3,5]	(2, 0.58)
[3,6]	(3, 0.87)
[4,5]	(3, 0.75)
[4,6]	(2, 0.5)
[5,6]	(3, 0.75)

3-itemset	(sup,cos)
[1,4,5]	(2, 0.63)
[1,4,6]	(2, 0.63)
[1,5,6]	(2, 0.63)
[3,5,6]	(2, 0.55)
[4,5,6]	(2, 0.5)

4-itemset	(sup,cos)
[1,4,5,6]	(2, 0.59)

* minsup: 2, mincos: 0.75

$$* \cos(i_1, i_2, \cdots, i_n) = \frac{\sup(i_1, i_2, \cdots, i_n)}{\sqrt[n]{\prod_{k=1}^{n} \sup(i_k)}}$$

Two Pruning Principles

$$corr(a_1, \cdots, a_{n+1}) \leq \max(corr(a_1, \cdots, a_n), \cdots, corr(a_2, \cdots, a_{n+1}))$$

- Level-based pruning
  - If all n-subitemsets of $\{a_1,\ldots,a_{n+1}\}$ are not correlated, then $\{a_1,\ldots,a_{n+1}\}$ is not correlated.
  - If all n-itemsets are not correlated, then no (n+1)-itemsets are correlated. (Termination of Pattern Growth, TPG)
- Single-Item Based Pruning (SIBP)
  - Let a has the minimum support between single items existing in DB. If all n-itemsets containing item a are not correlated, then all n'-itemsets containing a can be pruned for $n' \geq n$.
  - Iteratively shrink DB and apply SIBP.

Mining correlated patterns

- Threshold-based correlation mining
  - Apriori-like level-wise computation

Algorithm 1: The threshold-based version of the NICoMiner Algorithm.

input : a transactional database $\mathcal{D} = \{T_1, T_2, ..., T_n\}$, minimum correlation threshold γ, minimum support threshold θ

output: all patterns with correlation at least γ

```
1 scan D and find all frequent 1-itemsets I_1;
2 for n = 2, ··· do
3     generate candidate itemsets I_n from I_{n-1};
4     scan D to compute support and Cosine values of itemsets in I_n;
5     output frequent n-itemsets with Cosine ≥ γ;
6     prune itemsets from I_n based on SIBP and support;
7     if (maxCos(I_n) < γ) OR (no frequent n-itemsets) then break;
8 end
```

Termination of Pattern Growth (TPG)

TID	Transaction
T1	1, 3, 4, 5, 6
T2	3, 5, 6
T3	2, 4
T4	1, 4, 5, 6
T5	3, 6
T6	2, 4, 5

1-item	Support
1	2
2	2
3	3
4	4
5	4
6	4

2-itemset	(sup,cos)
[1,4]	(2, 0.71)
[1,5]	(2, 0.71)
[1,6]	(2, 0.71)
[2,4]	(2, 0.71)
[3,5]	(2, 0.58)
[3,6]	(3, 0.87)
[4,5]	(3, 0.75)
[4,6]	(2, 0.5)
[5,6]	(3, 0.75)

3-itemset	(sup,cos)
[1,4,5]	(2, 0.63)
[1,4,6]	(2, 0.63)
[1,5,6]	(2, 0.63)
[3,5,6]	(2, 0.55)
[4,5,6]	(2, 0.5)

4-itemset	(sup,cos)
[1,4,5,6]	(2, 0.59)

*minsup: 2, mincos: 0.75

Single-Item Based Pruning (SIBP)

TID	Transaction
T1	1, 3, 4, 5, 6
T2	3, 5, 6
T3	2, 4
T4	1, 4, 5, 6
T5	3, 6
T6	2, 4, 5

1-item	Support
1	2
2	2
3	3
4	4
5	4
6	4

2-itemset	(sup,cos)
~~[1,4]~~	~~(2, 0.71)~~
~~[1,5]~~	~~(2, 0.71)~~
~~[1,6]~~	~~(2, 0.71)~~
~~[2,4]~~	~~(2, 0.71)~~
[3,5]	(2, 0.58)
[3,6]	(3, 0.87)
[4,5]	(3, 0.75)
[4,6]	(2, 0.5)
[5,6]	(3, 0.75)

3-itemset	(sup,cos)
~~[1,4,5]~~	~~(2, 0.63)~~
~~[1,4,6]~~	~~(2, 0.63)~~
~~[1,5,6]~~	~~(2, 0.63)~~
[3,5,6]	(2, 0.55)
[4,5,6]	(2, 0.5)

1-item	Max cos
~~1~~	~~0.71~~
~~2~~	~~0.71~~
3	0.87
4	0.75
5	0.75
6	0.87

*minsup: 2, mincos: 0.75

Mining top-k correlated patterns

- Naïve way: iterative method
- Advanced method: one iteration method

Algorithm 2: The top-k version of NICOMINER

input : a transactional database $\mathcal{D} = \{T_1, T_2, ..., T_n\}$, number k, minimum support threshold θ

output: set TOP of top-k correlated patterns

```
1  γ ← 0; TOP ← ∅;
2  scan D and find all frequent 1-itemsets I_1;
3  for n = 2, ··· do
4      generate candidate itemsets I_n from I_{n-1};
5      scan D to compute support and Cosine values of all candidate k-itemsets;
6      TOP ← TOP ∪ {correlated n-itemsets};
7      if |TOP| ≥ k then
8          keep only top-k in TOP;
9          γ ← minimum Cosine value in TOP;
10     end
11     prune itemsets from I_n based on SIBP and support;
12     if (maxCos(I_n) < γ) OR (no frequent n-itemsets) then break;
13 end
```

Naïve way (iterative method)

minsup: 2, top-3 → minsup: 2, mincos: 1 → minsup: 2, mincos: 0.5

TID	Transaction
T1	1, 3, 4, 5, 6
T2	3, 5, 6
T3	2, 4
T4	1, 4, 5, 6
T5	3, 6
T6	2, 4, 5

1-item	Support
1	2
2	2
3	3
4	4
5	4
6	4

2-itemset	(sup,cos)
[1,4]	(2, 0.71)
[1,5]	(2, 0.71)
[1,6]	(2, 0.71)
[2,4]	(2, 0.71)
[3,5]	(2, 0.58)
[3,6]	(3, 0.87)
[4,5]	(3, 0.75)
[4,6]	(2, 0.5)
[5,6]	(3, 0.75)

3-itemset	(sup,cos)
[1,4,5]	(2, 0.63)
[1,4,6]	(2, 0.63)
[1,5,6]	(2, 0.63)
[3,5,6]	(2, 0.55)
[4,5,6]	(2, 0.5)

4-itemset	(sup,cos)
[1,4,5,6]	(2, 0.59)

Advanced Method (one-time method)

minsup: 2, top-3 → minsup: 2, top-3, mincos: 0.75 → minsup: 2, top-3, mincos: 0.75

TID	Transaction
T1	1, 3, 4, 5, 6
T2	3, 5, 6
T3	2, 4
T4	1, 4, 5, 6
T5	3, 6
T6	2, 4, 5

1-item	Support
1	2
2	2
3	3
4	4
5	4
6	4

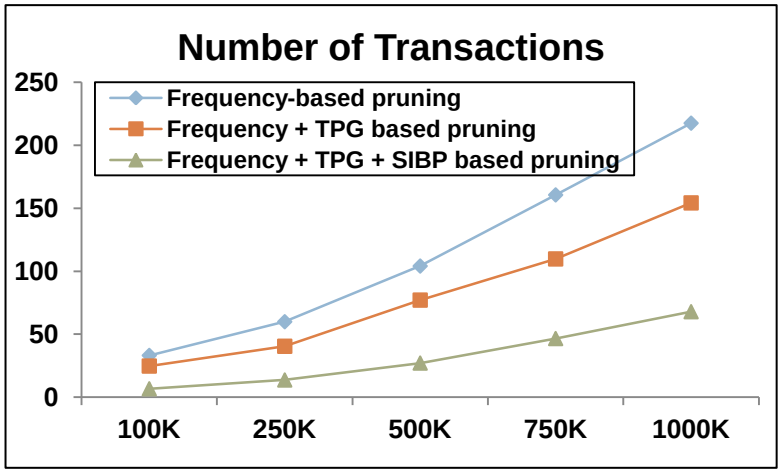
2-itemset	(sup,cos)
~~[1,4]~~	~~(2, 0.71)~~
~~[1,5]~~	~~(2, 0.71)~~
~~[1,6]~~	~~(2, 0.71)~~
~~[2,4]~~	~~(2, 0.71)~~
[3,5]	(2, 0.58)
[3,6]	(3, 0.87)
[4,5]	(3, 0.75)
[4,6]	(2, 0.5)
[5,6]	(3, 0.75)

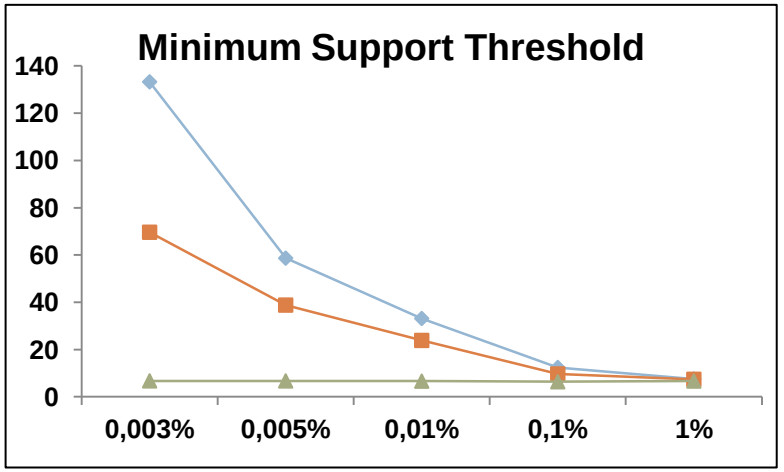
3-itemset	(sup,cos)
[3,5,6]	(2, 0.55)
[4,5,6]	(2, 0.5)

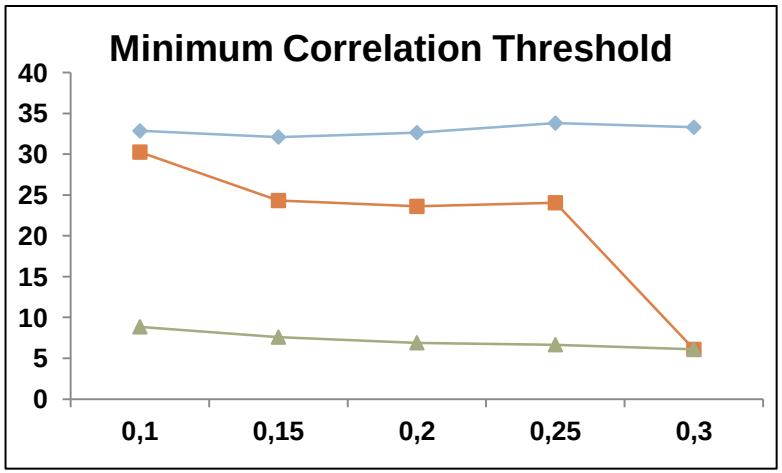
1-item	Max cos
~~1~~	~~0.71~~
~~2~~	~~0.71~~
3	0.87
4	0.75
5	0.75
6	0.87

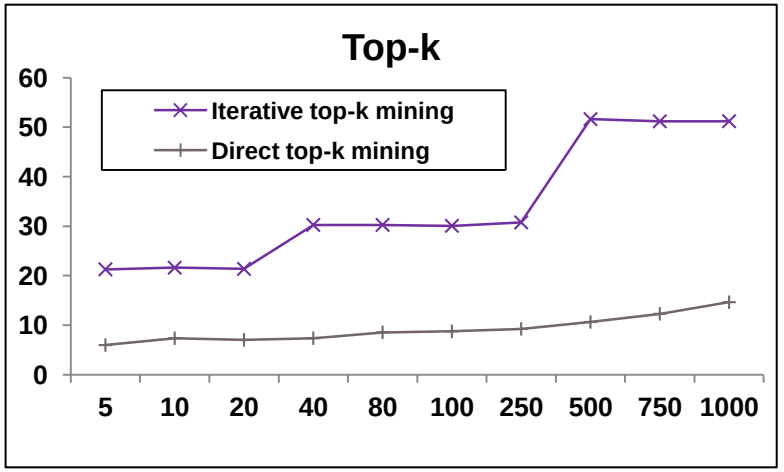
Experiments (I): Synthetic datasets

- Running Time (sec)







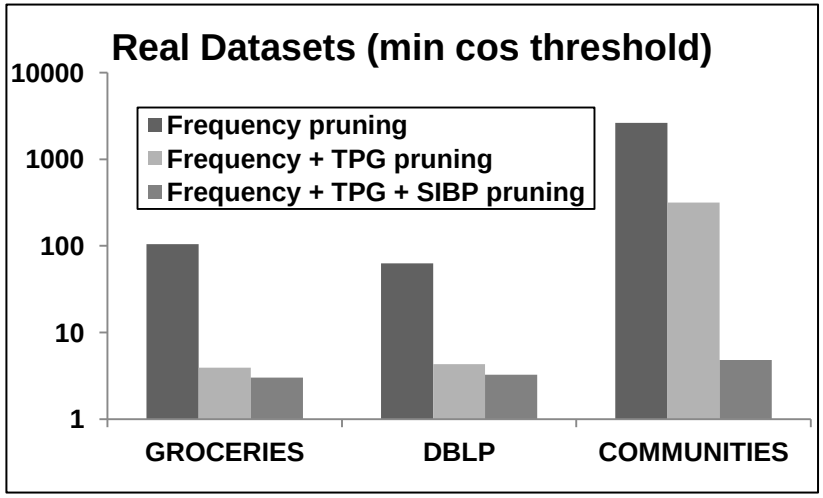


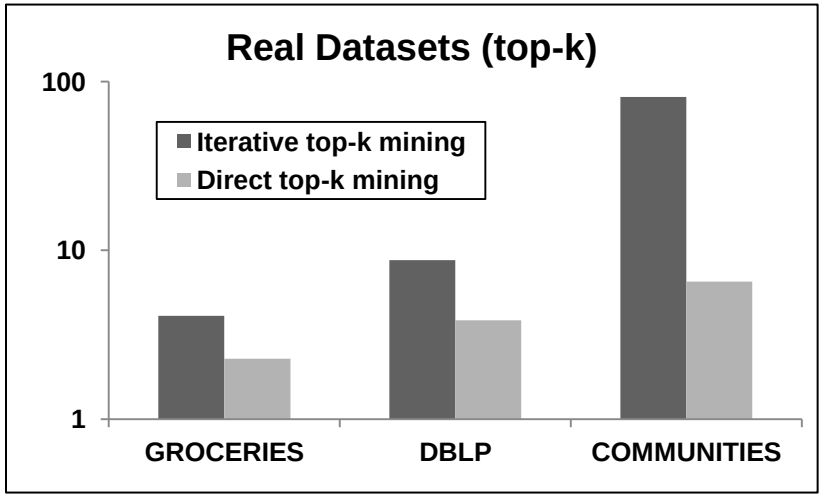
Experiments (II): Real datasets

- Data Sets

	# Transactions	# Items	Avg Trans Length
GROCERIES	10K	169	4.4
DBLP AUTHORS	7.2K	60694	2.7
COMMUNITIES	2K	210	40.3

- Running Time (sec)





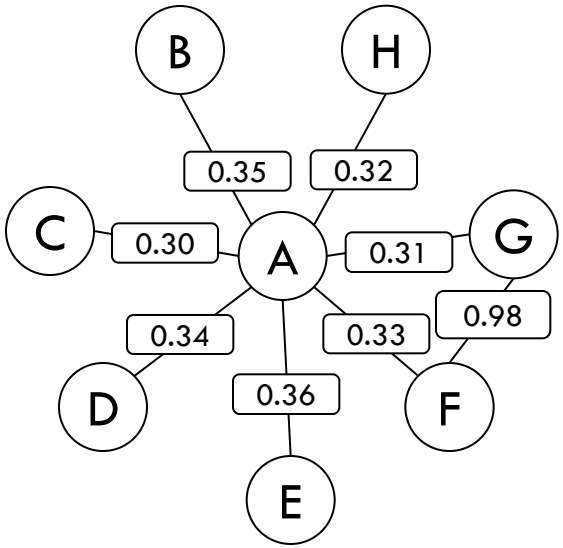
Experiments (III): Real datasets

- Top correlated patterns

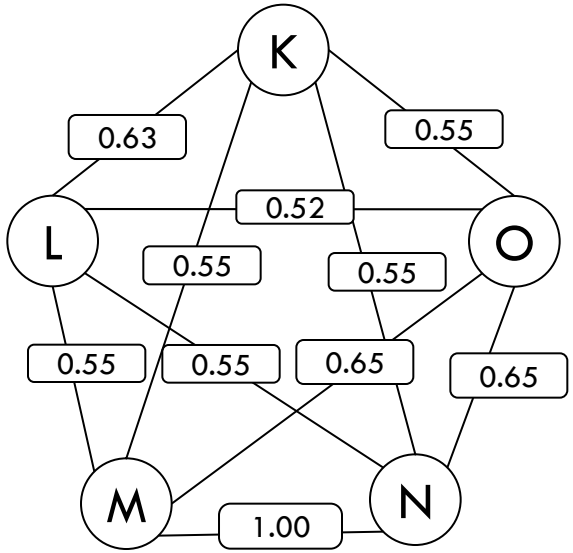
Dataset	Pattern	Support	Cosine
GROCERIES	butter milk, yogurt	84	0.14
	salty snack, popcorn	22	0.14
	chocolate, candy	49	0.13
DBLP AUTHORS	Steven M. Beitzel, Eric C. Jensen	25	1.00
	In-Su Kang, Seung-Hoon Na	20	0.98
	Ana Simonet, Michel Simonet	16	0.94
COMMUNITIES	(People with social security income): >80%, (Age ≥ 65): >80%	47	0.76
	(Large families (≥6)): ≤ 20%, (White): >80%	1017	0.75
	(In dense housing (≥1 per room)): >80%, (Large families (≥6)): >80%, (Hispanic): >80%	53	0.64

Experiments (IV): Real datasets

- Strong pairwise correlations in DBLP AUTHORS



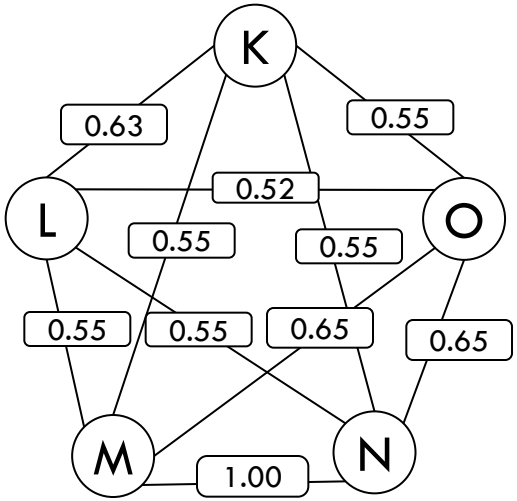
Advisor-advisee co-authorships

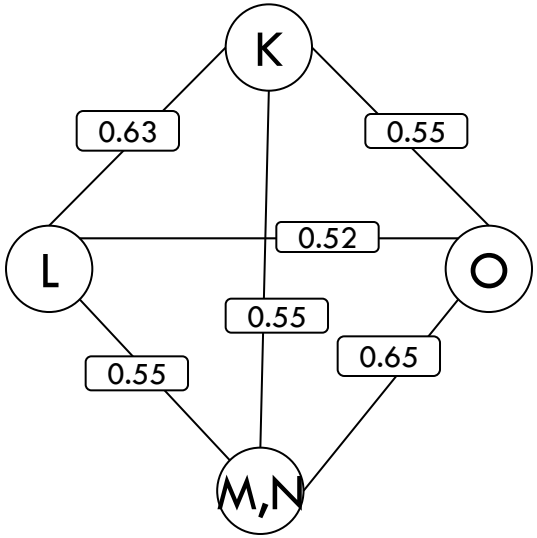


Advanced mutual co-authorships

Discussion

- Redundant patterns





- Can we guess correlation values of multiple items from pairwise correlations?
  - Cosine(K,L,M,N)=0.52, Cosine(K,M,N,O)<0.1
- Minimum support
  - Not main but supplementary factor of mining process

Conclusion

- Mining Correlated Patterns
  - Null-invariant measures are meaningful for real-life large transactional database.
  - Challenge: expensive computational cost
    - Two-step approach: not feasible
    - Direct mining
      - Cosine and Kulc are not (anti)-monotonic
  - Solution
    - Propose two pruning properties
      - Applicable to all 5 null-invariant measures
      - Enable to mine correlated patterns given a correlation threshold
      - Enable to mine top-k correlated patterns

Questions ?

Email to: Sangkyum Kim (kim71@illinois.edu)