Data Mining: A Powerful Tool for Data Cleaning

Jiawei Han

Department of Computer Science

University of Illinois at Urbana-Champaign

Nov. 4, 2003

Outline

- Data mining: A powerful tool for data cleaning
  - How can newer data mining methods help data quality assurance?
- PROM (Profile-based Object Matching): Identifying and merging objects by profile-based data analysis
- CoMine: Comparative correlation measure analysis
- CrossMine: Mining noisy data across multiple relations
- SecureClass: Effective document classification in the presence of substantial amount of noise
- Conclusions

Data Mining: A Tool for Data Cleaning

- Correlation, classification and cluster analysis for data cleaning
  - Discovery of interesting data characteristics, models, outliers, etc.
  - Mining database structures from contaminated, heterogeneous databases
- A comprehensive overview on the theme
  - Dasu & Johnson, Exploratory Data Mining and Data Cleaning, Wiley 2003.
- How can newer data mining methods help data quality assurance?
  - Exploring several newer data mining tasks and their relationships to data cleaning

Where Are the Source of the Materials?

- A. Doan, Y. Lu, Y. Lee and J. Han, Object matching for information integration: A profile-based approach, IEEE Intelligent Systems, 2003.
- Y.-K. Lee, W.-Y. Kim, Y. D. Cai, and J. Han, CoMine: Efficient mining of correlated patterns, Proc. 2003 Int. Conf. on Data Mining (ICDM'03), Melbourne, FL, Nov. 2003.
- X. Yin, J. Han, J. Yang, and P.S. Yu, CrossMine: Efficient classification across multiple database relations, Proc. 2004 Int. Conf. on Data Engineering, Boston, MA, March 2004
- X.Yin, J. Han, A. Mehta, SecureClass: Privacy-Preserving Classification of Text Documents, submitted for publication.

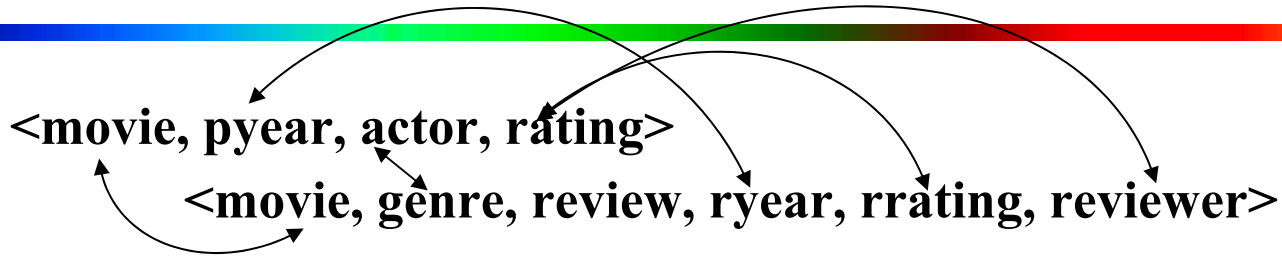
Object Matching for Data Cleaning

- Object matching: Identifying and merging objects by data mining and statistical analysis
- Decide if two objects refer to the same real-world entity
  - (Mike Smith, 235-2143) & (M. Smith, 217 235-2143)
- Purposes: information integration & data cleaning
  - remove duplicates when merging data sources
  - consolidate information about entities
  - information extraction from text
  - join of string attributes in databases

PROM: Profile-based Object Matching

- Key observations
  - disjoint attributes are often correlated
  - such correlations can be exploited to perform "sanity check"
- Example
  - (9, Mike Smith) & (Mike Smith, 200K)
  - Match them?– because both names are "Mike Smith"?
  - Sanity check using profiler:
    - Match? → Mike Smith: 9 years-old with salary 200K
    - Knowledge: the profile of a typical person
    - Conflict with the profile → two are unlikely to match

Example: Matching Movies

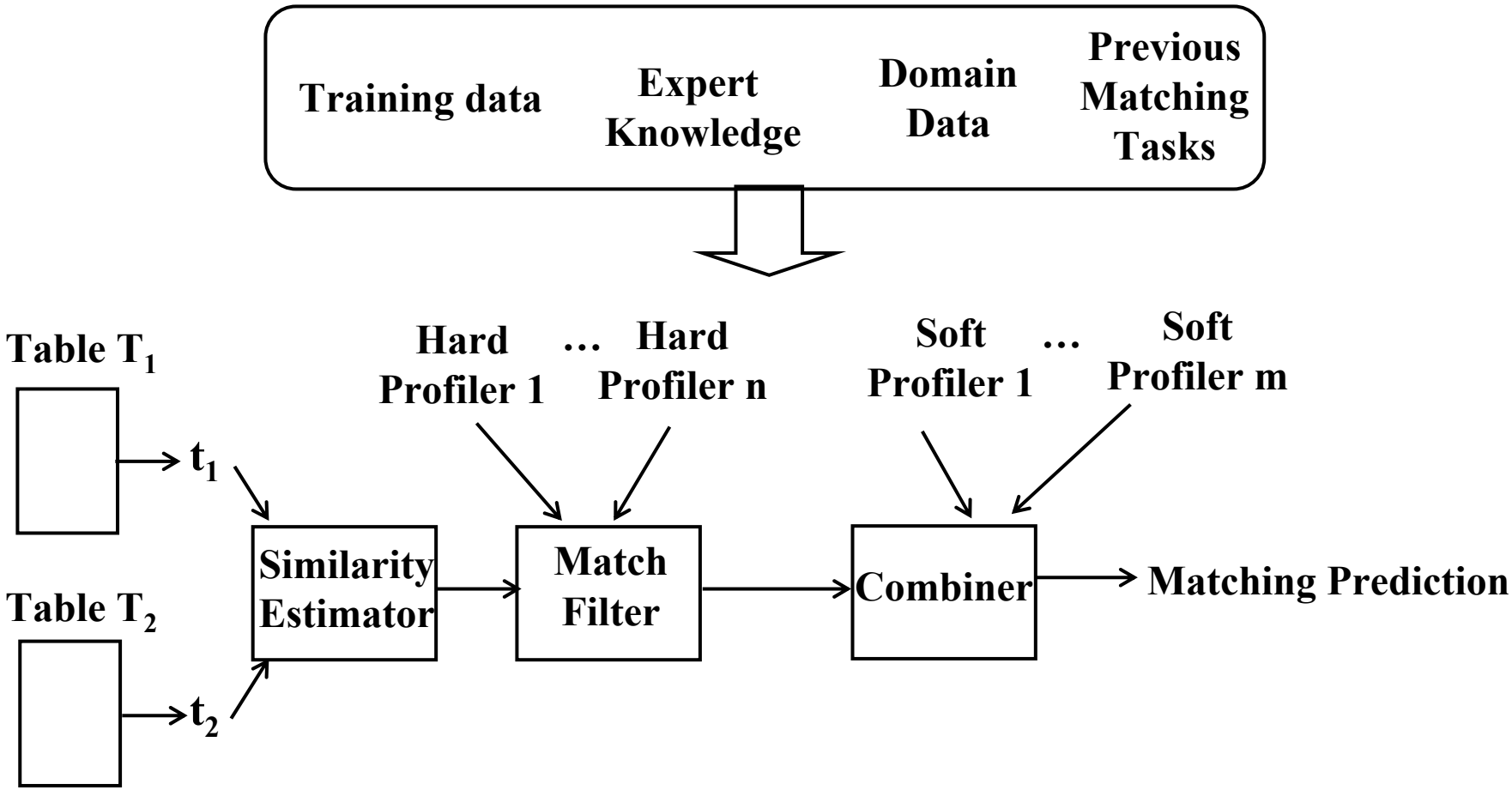


- Step 1: check if two movie names are sufficiently similar
- Step 2: sanity check using multiple profilers
  - review profiler:
    - Production year (pyear) must not be after review year (ryear)
    - Roger Ebert (reviewer) never reviews movies with rating < 5
  - actor profiler:
    - Certain actor has never played in action movies
  - movie profiler:
    - Rating and rrating tend to be strongly correlated
  - PROM combines profiler predictions to reach matching decision

Profilers in Movie Example

- Contain knowledge about domain concepts
  - movies, reviews, actors, studios, etc.
- Constructed once, reused anywhere
  - as long as the new matching task involves same domain concepts
- Can be constructed in many ways
  - manually specified by experts and users
  - learned from data in the domain
    - all movies at Internet Movie Database imdb.com
    - text of reviews from the New York Times
  - learned from training data of a specific matching task
    - then transferred to related matching tasks

Architecture of PROM



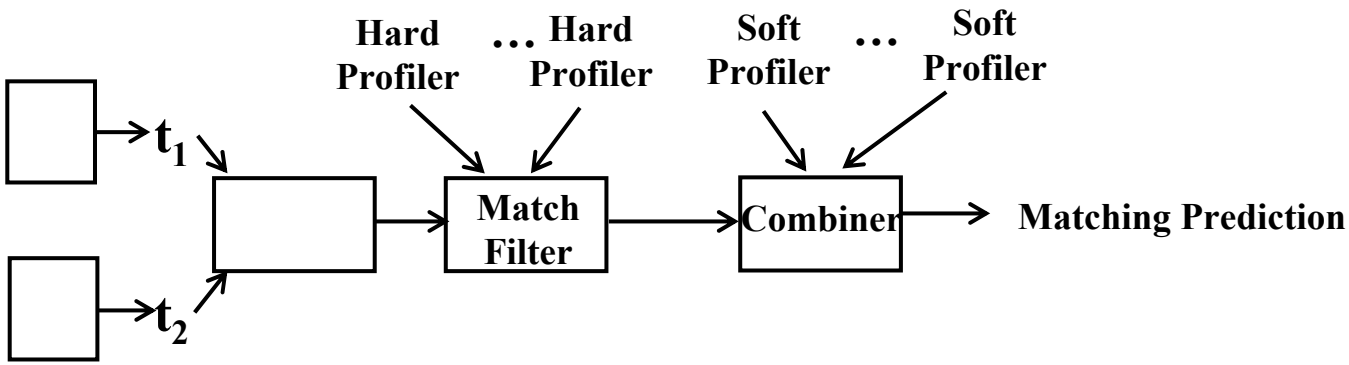
Hard vs. Soft Profilers: Hard Profiler

- Given a tuple pair
  - A profiler issues a confidence score on how well the pair fits the concept (i.e., how well their data mesh together)
- Hard profiler
  - specifies constraints that any concept instance must satisfy
    - review year ≥ production year of movie
    - actor A has only played in action movies
  - can be constructed manually by domain experts and users
  - can be constructed from domain data if data is complete
    - e.g., by examining all movies of actor A

Hard vs. Soft Profilers: Soft Profiler

- Soft profiler
  - Specifies “soft” constraints that most instances satisfy
  - can be constructed manually, from domain data
    (e.g., learning a Bayesian network from imdb.com)
  - from training data of a matching task
    (e.g., learning a classifier from training data)

Combining Profilers



- Step 1: How to combine hard profilers?
  - Any hard profiler says “no match”, declare “no match”
- Step 2: How to combine soft profilers?
  - Each soft profiler examines pair and issue a prediction “match” with a confidence score
  - Combine profilers’ scores
    - currently use weighted sum (with weights set manually)

Empirical Evaluation: CiteSeer Name Match

- CiteSeer: Popularly cited authors but may not match the correct homepages
- Citation list: Highly cited researchers and their homepages
  - The "Jim Gray" citeseer problem: cs.vt.edu/~gray, data.com/~jgray, microsoft.com/~gray
  - Which homepage should be for the real J. Gray?
- Created two data sources
  - source 1: highly cited researchers, 200 tuples
    - (name, highly-cited)
  - source 2: homepages, 254 tuples (manually created from text)
    - (name, title, institute, graduation-year, …)

PROM Improves Matching Accuracy

		Baseline	PROM			
			DT	Man+DT	Man+AR	Man+AR+DT
CiteSeer	Recall	0.99	0.95	0.67	0.96	0.97
	Precision	0.67	0.78	0.87	0.82	0.86
	F-Value	0.80	0.85	0.76	0.88	0.91

- Baseline: exploit only shared attributes
- PROM:
  - Used three soft profilers: DT (decision tree), Man (manual), and AR (association rules)
  - Adding profilers tends to improve accuracy
    - DT < Man+AR < Man+AR+DT

CoMine: Mining Strongly Correlated Patterns

- Why CoMine is closely related to data cleaning?
  - Correlation analysis: A powerful data cleaning tool
  - Current association analysis: generate too many rules!
  - Maybe the correlation rules are what we want
- What should be a good correlation measure to handle large data sets?
  - Find good correlation measure
  - Find an efficient mining method

Why Mining Correlated Patterns?

- Association ≠ correlation
  - high min_support → commonsense knowledge
  - low minimum support → huge number of rules
- Association may not carry the right semantics
  - *"Buy walnuts* $\Rightarrow$ *buy milk* [1%, 80%]" is misleading
    - if 85% of customers buy milk
- What should be a good measure?
  - Support and conf. alone are no good
  - Will lift or χ^2 be better?

A Comparative Analysis of 21 Interesting Measures

symbol	measure	range	formula
ϕ	ϕ-coefficient	-1 . . . 1	$\frac{P(A,B)-P(A)P(B)}{\sqrt{P(A)P(B)(1-P(A))(1-P(B))}}$
Q	Yule's Q	-1 . . . 1	$\frac{P(A,B)P(\overline{A},\overline{B})-P(A,\overline{B})P(\overline{A},B)}{P(A,B)P(\overline{A},\overline{B})+P(A,\overline{B})P(\overline{A},B)}$
Y	Yule's Y	-1 . . . 1	$\frac{\sqrt{P(A,B)P(\overline{A},\overline{B})}-\sqrt{P(A,\overline{B})P(\overline{A},B)}}{\sqrt{P(A,B)P(\overline{A},\overline{B})}+\sqrt{P(A,\overline{B})P(\overline{A},B)}}$
k	Cohen's	-1 . . . 1	$\frac{P(A,B)+P(\overline{A},\overline{B})-P(A)P(B)-P(\overline{A})P(\overline{B})}{1-P(A)P(B)-P(\overline{A})P(\overline{B})}$
PS	Piatetsky-Shapiro's	-0.25 . . . 0.25	$P(A,B)-P(A)P(B)$
F	Certainty factor	-1 . . . 1	$\max(\frac{P(B\|A)-P(B)}{1-P(B)}, \frac{P(A\|B)-P(A)}{1-P(A)})$
AV	added value	-0.5 . . . 1	$\max(P(B\|A)-P(B), P(A\|B)-P(A))$
K	Klosgen's Q	-0.33 . . . 0.38	$\sqrt{P(A,B)}\max(P(B\|A)-P(B), P(A\|B)-P(A))$
g	Goodman-kruskal's	0 . . . 1	$\frac{\Sigma_j \max_k P(A_j,B_k)+\Sigma_k \max_j P(A_j,B_k)-\max_j P(A_j)-\max_k P(B_k)}{2-\max_j P(A_j)-max_k P(B_k)}$
M	Mutual Information	0 . . . 1	$\frac{\Sigma_i\Sigma_j P(A_i,B_j)\log\frac{P(A_i,B_j)}{P(A_i)P(B_J)}}{\min(-\Sigma_i P(A_i)\log P(A_i)\log P(A_i), -\Sigma_i P(B_i)\log P(B_i)\log P(B_i))}$
J	J-Measure	0 . . . 1	$\max(P(A,B)\log(\frac{P(B\|A)}{P(B)}) + P(A\overline{B})\log(\frac{P(\overline{B}\|A)}{P(\overline{B})}))$ $P(A,B)\log(\frac{P(A\|B)}{P(A)}) + P(\overline{A}B)\log(\frac{P(\overline{A}\|B)}{P(\overline{A})})$
G	Gini index	0 . . . 1	$\max(P(A)[P(B\|A)^2 + P(\overline{B}\|A)^2] + P(\overline{A}[P(B\|\overline{A})^2 + P(\overline{B}\|\overline{A})^2] - P(B)^2 - P(\overline{B})^2,$ $P(B)[P(A\|B)^2 + P(\overline{A}\|B)^2] + P(\overline{B}[P(A\|\overline{B})^2 + P(\overline{A}\|\overline{B})^2] - P(A)^2 - P(\overline{A})^2)$
s	support	0 . . . 1	$P(A,B)$
c	confidence	0 . . . 1	$max(P(B\|A), P(A\|B))$
L	Laplace	0 . . . 1	$\max(\frac{NP(A,B)+1}{NP(A)+2}, \frac{NP(A,B)+1}{NP(B)+2})$
IS	Cosine	0 . . . 1	$\frac{P(A,B)}{\sqrt{P(A)P(B)}}$
γ	coherence(Jaccard)	0 . . . 1	$\frac{P(A,B)}{P(A)+P(B)-P(A,B)}$
α	all_confidence	0 . . . 1	$\frac{P(A,B)}{\max(P(A),P(B))}$
o	odds ratio	0 . . . ∞	$\frac{P(A,B)P(\overline{A},\overline{B})}{P(\overline{A},B)P(A,\overline{B})}$
V	Conviction	0.5 . . . ∞	$\max(\frac{P(A)P(\overline{B})}{P(A\overline{B})}, \frac{P(B)P(\overline{A})}{P(B\overline{A})})$
λ	lift	0 . . . ∞	$\frac{P(A,B)}{P(A)P(B)}$
S	Collective strength	0 . . . ∞	$\frac{P(A,B)+P(\overline{AB})}{P(A)P(B)+P(\overline{A})P(\overline{B})} \times \frac{1-P(A)P(B)-P(\overline{A})P(\overline{B})}{1-P(A,B)-P(\overline{AB})}$
χ^2	χ^2	0 . . . ∞	$\Sigma_i \frac{(P(A_i)-E_i)^2}{E_i}$

Let's Look Closely on a few Measures

$$\lambda = lift = \frac{P(A \cup B)}{P(A)P(B)}$$

$$\chi^2 = \sum \frac{(Observed - Expected)^2}{Expected}$$

$$\alpha = all\_conf = \frac{\sup(X)}{\max\_item\_\sup(X)}$$

$$\gamma(Jaccard\_Coeff) = coh = \frac{\sup(X)}{|universe(X)|}$$

Comparison among λ, α, γ, and χ^2

- The contingency table and the behavior of a few measures

DB	mc	¬mc	m¬c	¬(mc)	λ	α	γ	χ^2
A1	1000	100	100	1000	83.64	0.91	0.83	83452
A2	1000	100	100	10000	9.26	0.91	0.83	9055
A3	1000	100	100	100000	1.82	0.91	0.83	1472
A4	100	1000	1000	100000	8.44	0.09	0.05	670
A5	1000	100	10000	100000	9.18	0.09	0.09	8172
A6	1000	1000	1000	1000	1	0.5	0.33	0

	milk	¬milk
coffee	mc	¬mc
¬coffee	m¬c	¬(mc)

What Should Be a Good Correlation Measure?

- Disclose genuine correlation relationship
- Null Invariance Property (Tan, et al. 02)
  - Invariant by adding more null transactions (those not containing these items)
  - Useful in large sparse databases – co-presence is far less than co-absence
- Has the downward closure property
  - for efficient mining (Apriori like algorithms)

Examining a larger set of Measures

ϕ	ϕ-coefficient
Q	Yule's Q
Y	Yule's Y
k	Cohen's
PS	Piatetsky-Shapiro's
F	Certainty factor
AV	Added value
k	Klosgen's Q

range from -1 to 1

g	Goodman-kruskal's
M	Mutual Information
J	J-Measure
G	Gini index
s	support
c	confidence
L	Laplace
IS	Cosine
γ	Coherence(Jaccard)
α	All_confidence

range from 0 to 1

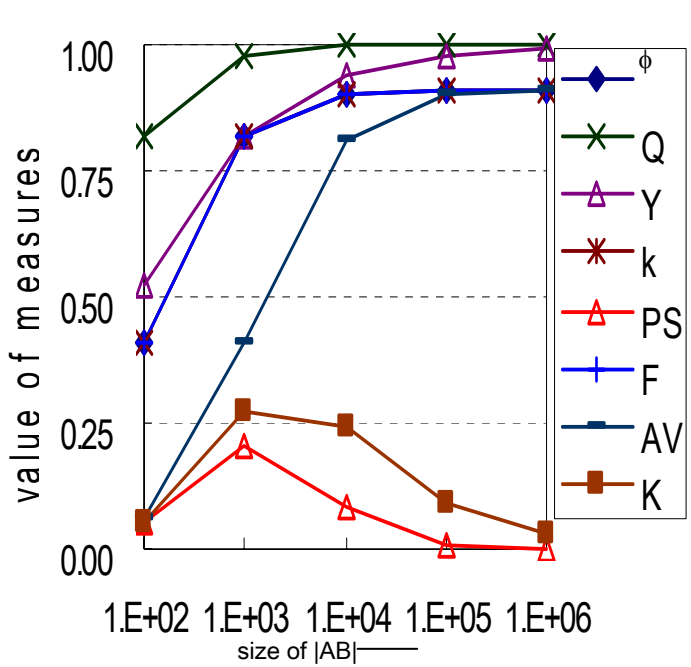
o	odds ratio
V	Conviction
λ	lift
S	Collective Strength
χ^2	χ^2

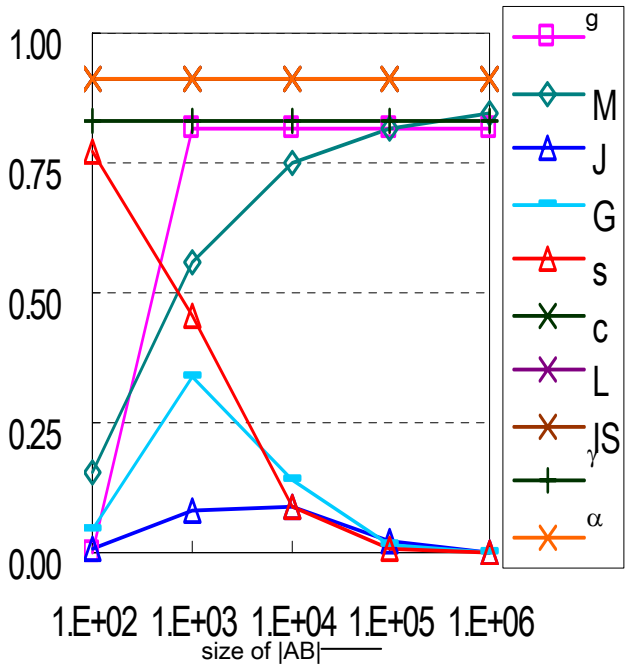
range from 0 to ∞

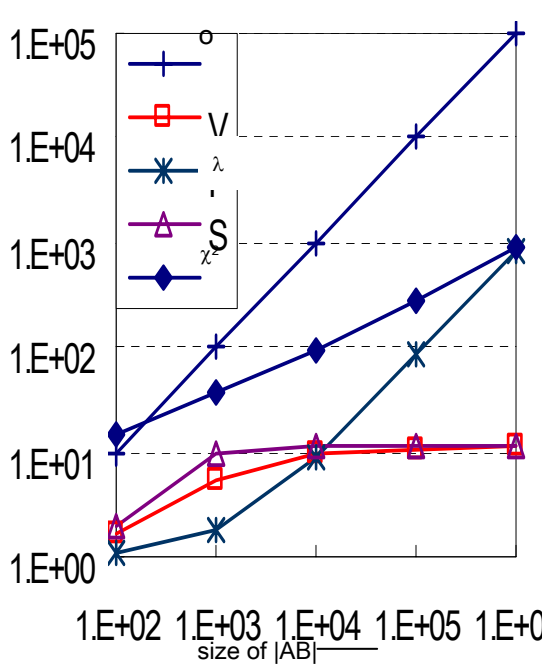
Effect of Null Transactions: Positively Correlated Cases

- Input parameters (symmetric data)
- Results

	B	¬B
A	1000	100
¬ A	100	$\overline{\lvert AB \rvert}$





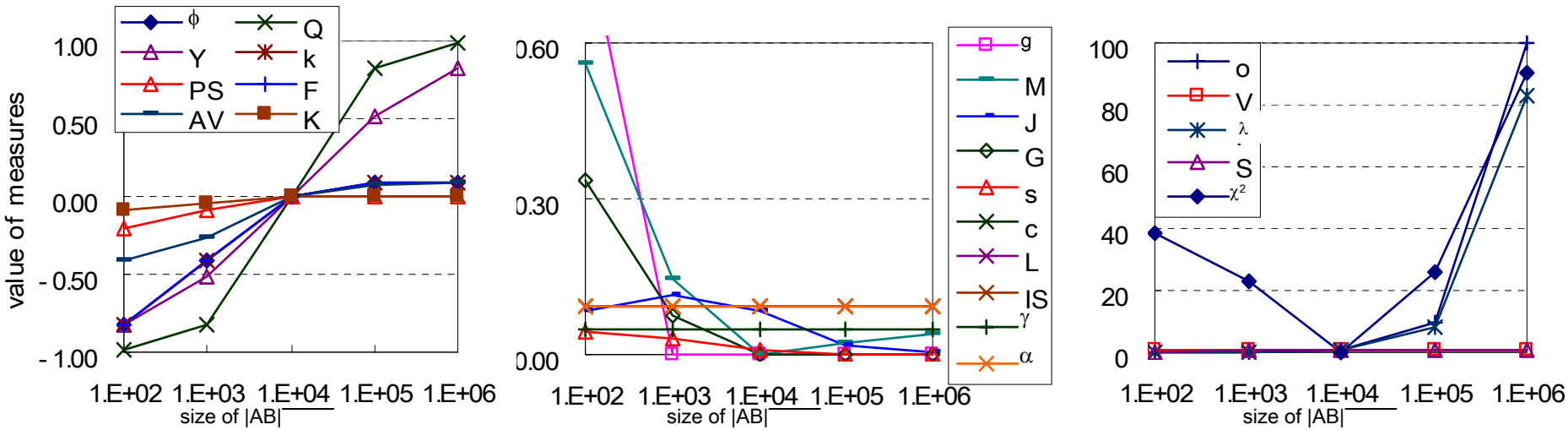


Effect of Null Transactions: Negatively Correlated Cases

- Input parameters

	B	¬B
A	100	1000
¬ A	1000	$\overline{\lvert AB \rvert}$

- Results

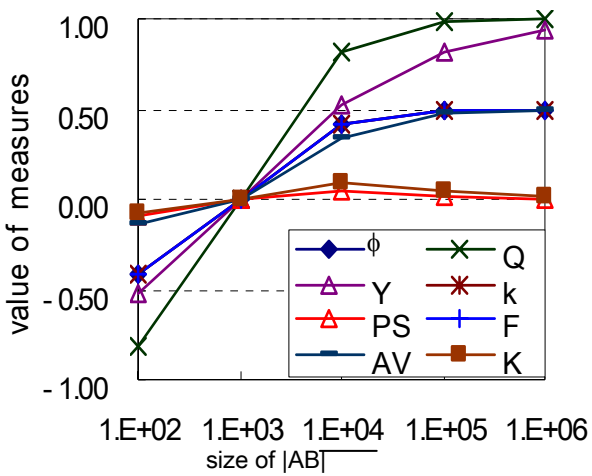


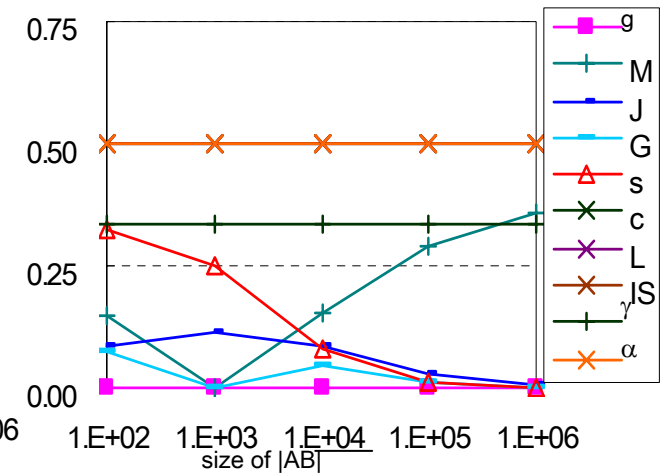
Effect of Null Transactions: Independently Correlated Cases

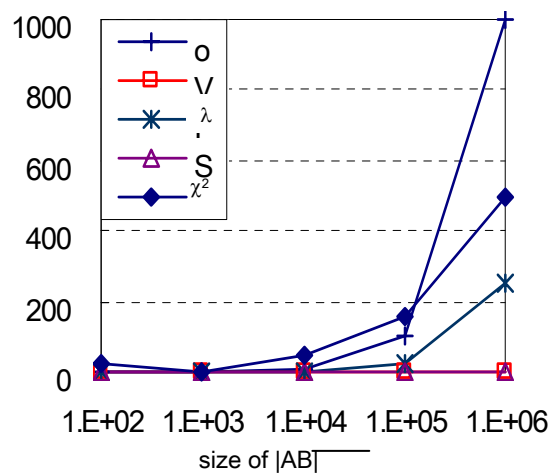
- Input parameters

	B	¬B
A	1000	1000
¬ A	1000	$\overline{\lvert AB \rvert}$

- Results



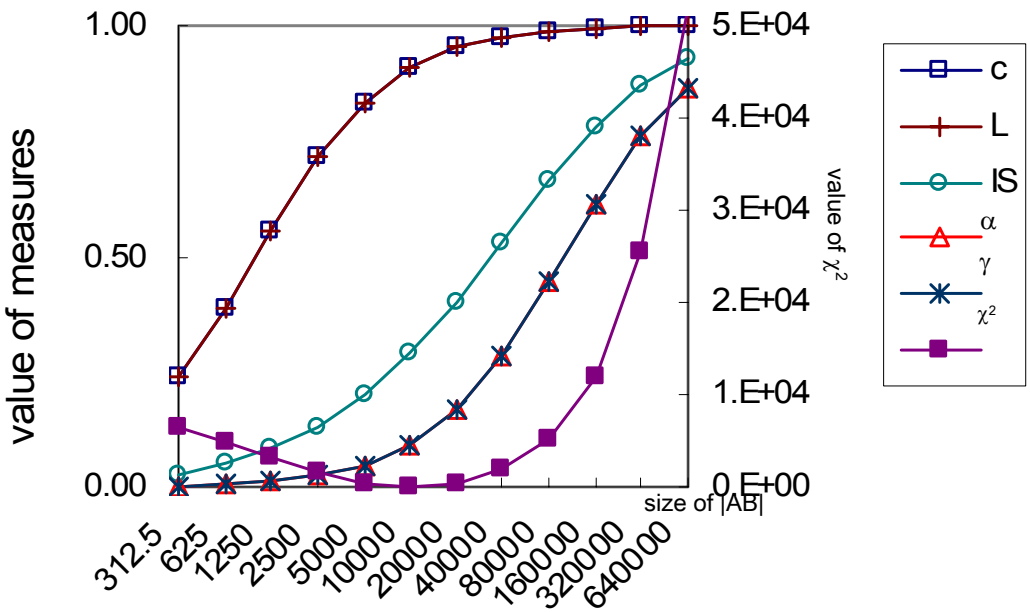




Correlations in Asymmetric Data

- Input parameters (asymmetric data)
- Results

	B	¬B
A	|AB|	10000
¬ A	100	$\overline{10000}$



⇒IS, α, and γ are good. However, IS doesn't have downward close property.

CoMine: Efficient Correlation Mining

- Utilize the downward close property
  - Given a pattern X,
    - if all_conf(X) $\geq$ min_α, then $\forall Y \subseteq X$, all_conf(Y) $\geq$ min_α
    - if coh(X) $\geq$ min_γ, then $\forall Y \subseteq X$, coh(Y) $\geq$ min_ γ.
- Extend the FP-growth: Additional optimization techniques
  - (for both) Counting space pruning
  - (for γ) Efficient computing cardinality of the universe
  - (for γ) Reducing the number of computations of the universe cardinality

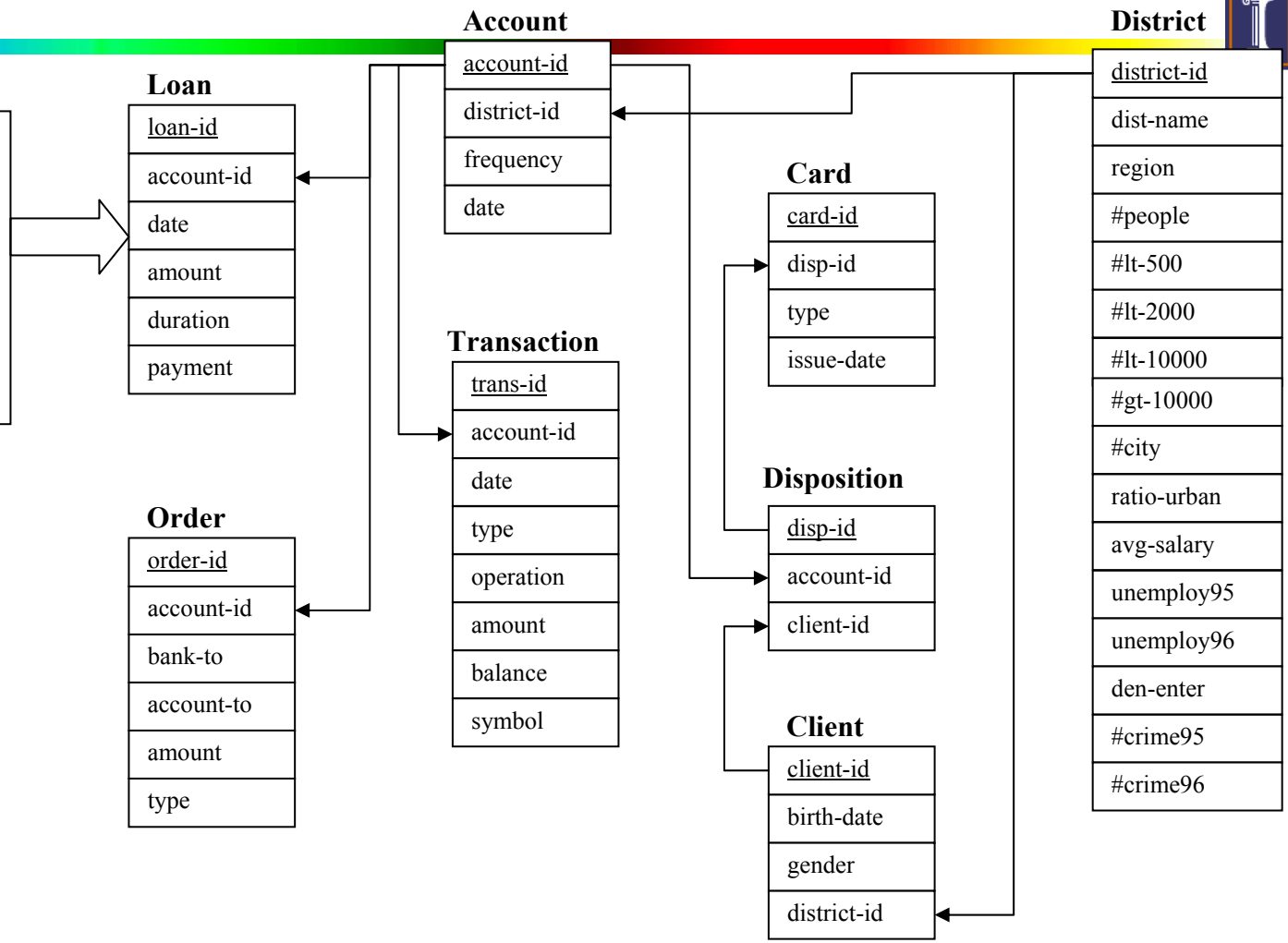
How May CrossMine Help Data Quality?

- CrossMine: Efficient classification across multiple database relations
- Originally designed for efficient multi-relational data mining
- Data quality issue exists across multiple relations
  - Data quality assurance is more challenging in multi-relational environment
- Efficient and effective classification across multi-relations will help data cleans and data quality assurance

Multi-Relational Classification

Target relation:

Each tuple has a class label, indicating whether a loan is paid on time.



Example rules:

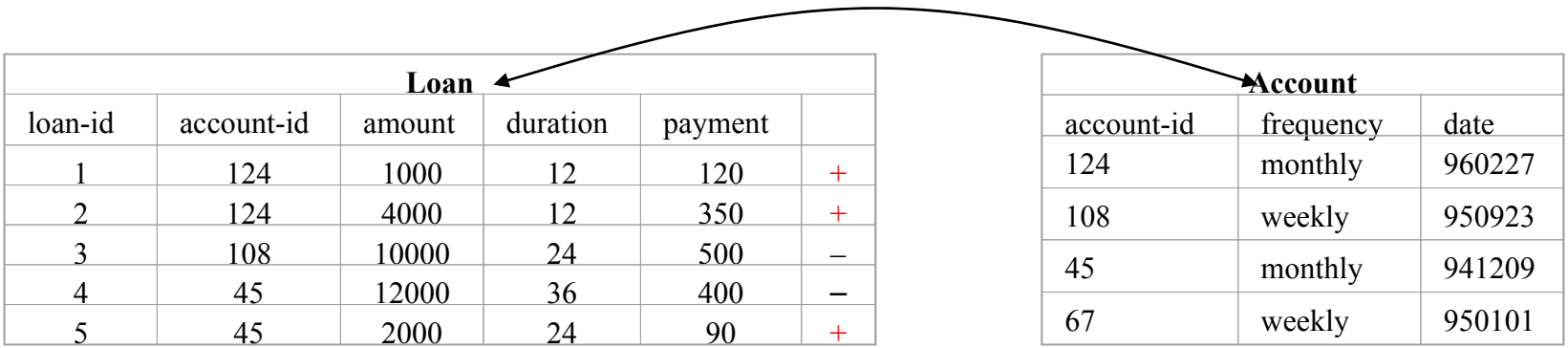
Loan(*L*, +) :- Loan (*L, A,?,?,?,?*), Account(*A*,*?*,'monthly',*?*).

Loan(*L*, +) :- Loan (*L, A*,?,?,?,'<1000'), Account(*A,D,?,?*),

District(*D,?,* region = 'northMoravia'*,?,?,…*).

Existing Approaches

- Inductive Logic Programming (FOIL, Golem, …)
  - Repeatedly find the best predicate.
  - To evaluate a predicate *p* on relation *R*, first join target relation with *R*, which is time consuming.
  - Not scalable w.r.t. size of database schema, because of huge search space.



Loan

loan-id	account-id	amount	duration	payment	
1	124	1000	12	120	+
2	124	4000	12	350	+
3	108	10000	24	500	–
4	45	12000	36	400	–
5	45	2000	24	90	+

Account

account-id	frequency	date
124	monthly	960227
108	weekly	950923
45	monthly	941209
67	weekly	950101

Predicates on Account relation:

Loan (*L, A,?,?,?*), Account(*A*, 'monthly' (or 'weekly'),*?*).

Loan (*L, A,?,?,?*), Account(*A*,*?, date<x* (*date>x*)).

Tuple ID Propagation

- Propagate the tuple IDs of the target relation to non-target relations
- Virtually join the relations, but avoid the high cost of physical joins

Loan

loan-id	account-id	amount	duration	payment	
1	124	1000	12	120	+
2	124	4000	12	350	+
3	108	10000	24	500	–
4	45	12000	36	400	–
5	45	2000	24	90	+

Account

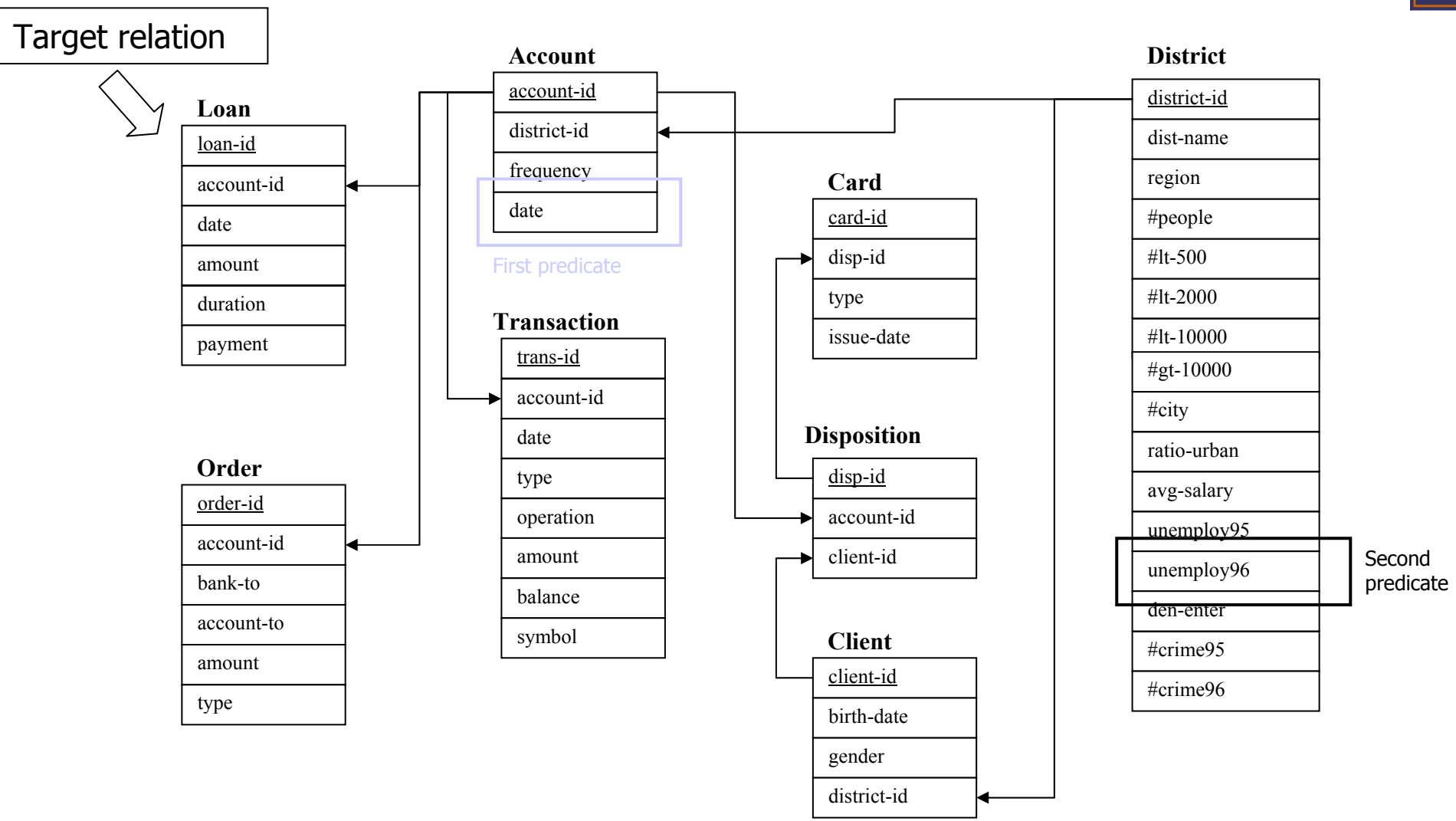
account-id	frequency	date	IDs	Class Labels
124	monthly	960227	1, 2	2+, 0–
108	weekly	950923	3	0+, 1–
45	monthly	941209	4, 5	1+, 1–
67	weekly	950101	--	0+, 0–

- Tuple IDs can be propagated freely among relations
- Search for good predicates in promising directions

Algorithm for Finding the Best Predicate

- Relations used in the current rule are called ***Active Relations***
- To compute foil gain of predicates:
  - Predicates on active relations are computed directly
  - Predicates on relations directly joinable to some active relation: Propagate tuple IDs, then compute
  - Predicates on other relations: Do not compute

Algorithm for Finding the Best Predicate



Performance on Synthetic Datasets:

Scalability w.r.t. number of relations

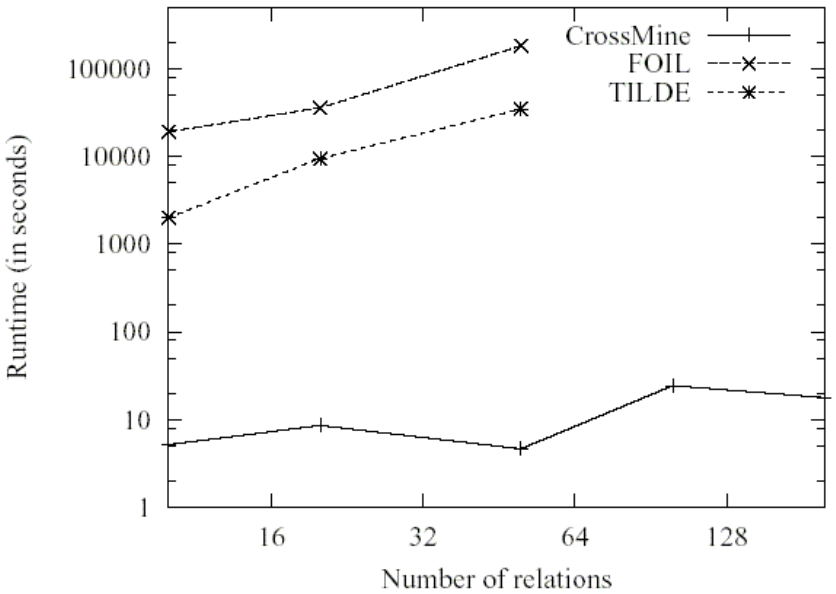


Figure 9. Runtime on R*.T500.F2.

Scalability w.r.t. number of tuples

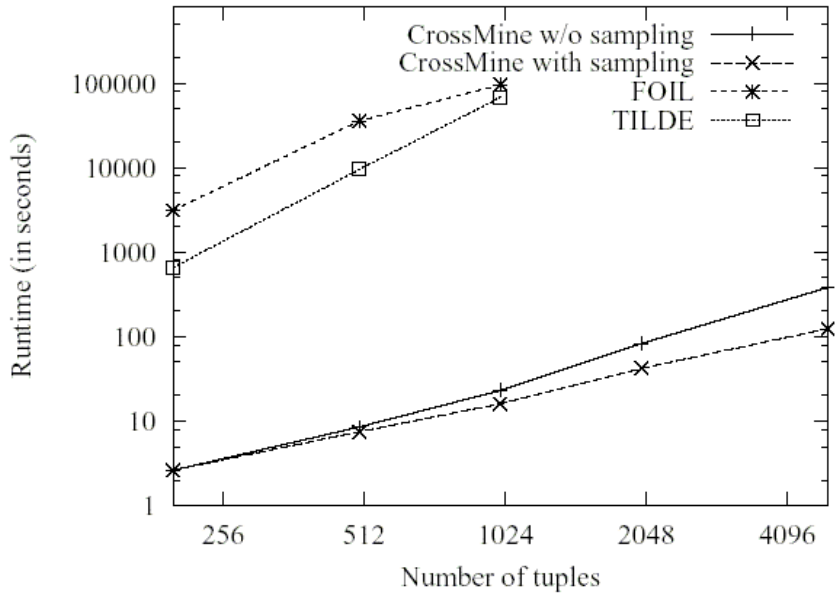
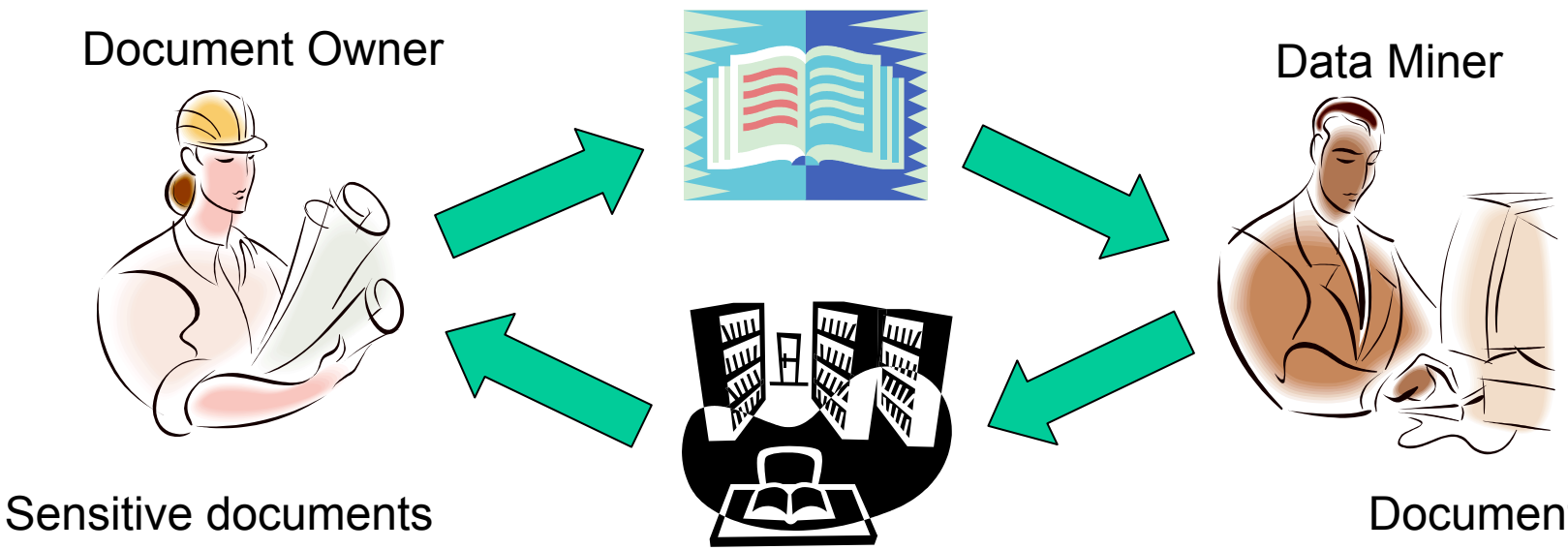


Figure 11. Runtime on R20.T*.F2.

Performance on Real data set: PKDD Cup 99 dataset

	Accuracy	Time
FOIL	74.0%	3338 sec
TILDE	81.3%	2429 sec
CrossMine	90.7%	15.3 sec

Privacy-Preserving Document Classification



SecureClass: Privacy-Preserving Classification of Text Documents, by Xiaoxin Yin, Jiawei Han, Anish Mehta

Why Is SecureClass Related to DQ Issues?

- Philosophy of SecureClass
  - intentionally introduce noises to documents
  - so that documents are not understandable
  - but still preserves classifiable property
- Real data is dirty, but we may still like to do effective classification
- Can we explore privacy-preserving mining methodology for effective classification of documents or other kinds of data?
- Efficient and effective classification despite of noise

Removing Privacy Information

- Randomizing a document
  - Remove sensitive words (names, locations, …), numerical data, dates, etc. Only common words are kept
  - Smash the order of words
  - Remove up to 40% of words and add up to 40% of noises

In regards to fractal compression, I have seen 2 fractal compressed "movies". They were both fairly impressive. The first one was a 64 gray scale "movie" of Casablanca, it was 1.3MB and had 11 minutes of 13 fps video. It was a little grainy but not bad at all. The second one I saw was only 3 minutes but it had 8 bit color with 10fps and measured in at 1.2MB.

I consider the fractal movies a practical thing to explore. But unlike many other formats out there, you do end up losing resolution. I don't know what kind of software/hardware was used for creating the "movies" I saw but the guy that showed them to me said it took 5-15 minutes per frame to generate. But as I said above playback was 10 or more frames per second. And how else could you put 11 minutes on one floppy disk?

davidr@rincon.ema.rockwell.com

My opinions are my own except where they are shared by others in which case I will probably change my mind.

speed, minut, him, assign, regard, complex, took, cheer, reach, idl, send, state, consid, presum, through, divis, resolut, frame, perhap, disclaim, locat, lose, name, qualiti, except, mail, posit, cabl, els, ride, bit, gener, avail, hurt, format, said, sox, littl, own, chang, put, share, upon, softwar, card, mean, impress, util, point, saw, better, consult, file, read, movi, per, drive, mani, unlik, first, realli, occur, imag, practic, floppi, seem, color, thing, system, recent, want, could, apr, sometim, had, them, gui, fine, kind, math, entri, folk, show, seek, gov, second, meet

Document Classification Process

- Build rules that predict for class labels with a sequential covering algorithm.
  - routine, polygon → computer graphics

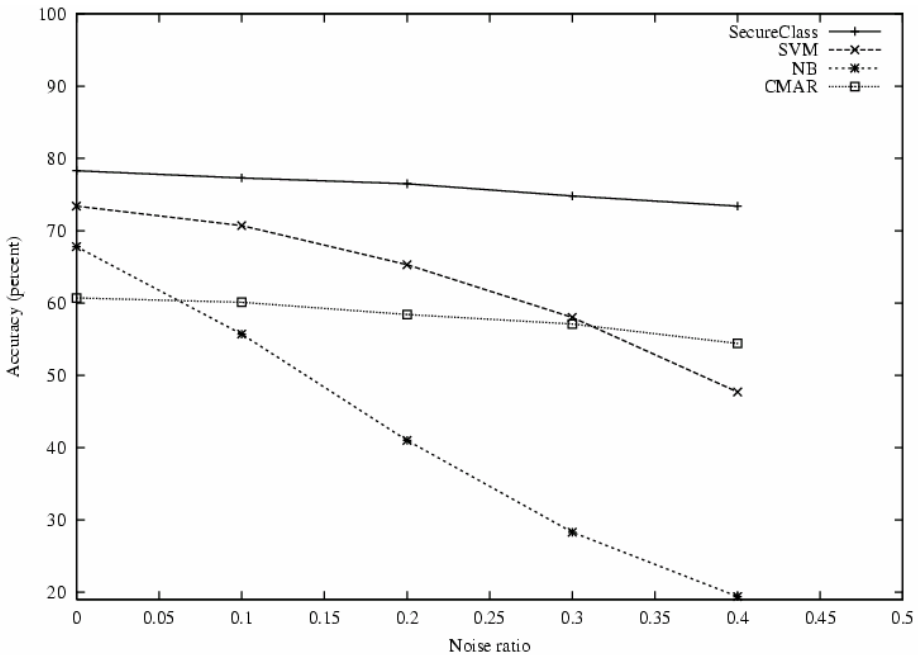
a frequent pattern *a class label*

- Rules may come from noises. Use following constraints to rules:
  - Rules with high support are less likely to come from noises
  - Longer rules are less likely to come from noises
  - For each rule r = "$w_1, \ldots, w_k \rightarrow c$"

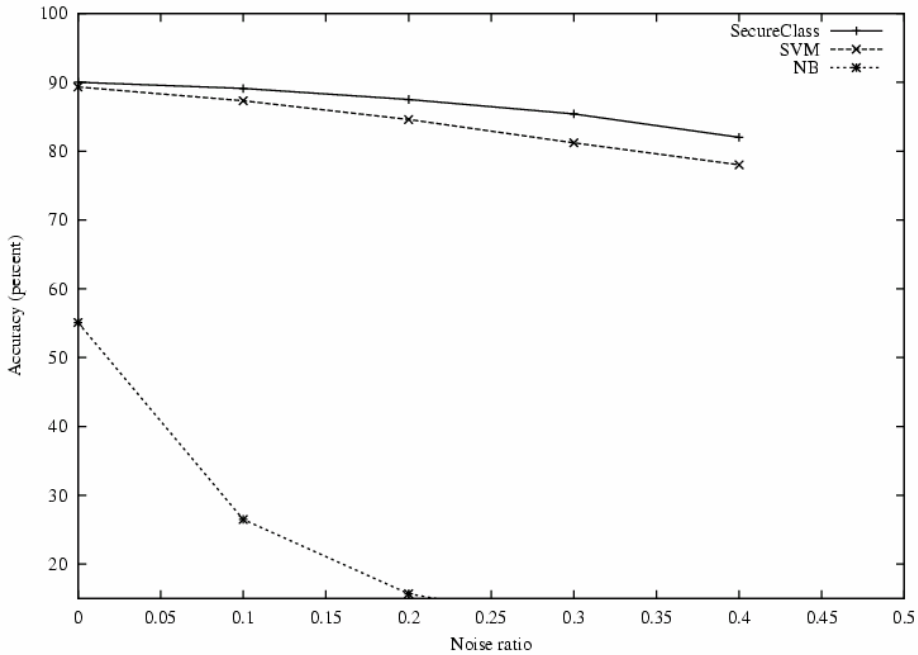
    Make sure that r's confidence is improved at most ε by noises, with probability $(1-\delta)$.

Experimental Results

Accuracy on newsgroup dataset



Accuracy on BankSearch dataset



SecureClass is more accurate than SVM, Naïve Bayes, and CMAR.

The accuracy of SecureClass is less affected than those three approaches.

The efficiency of SecureClass is similar to SVM, and is slower than Naïve Bayes but faster than CMAR.

Conclusions

- Data Mining helps data quality assurance
  - Not only by traditional statistical, machine learning, data mining methods
  - But also potentially with newer techniques
- Explore how to explore new data mining methods for data quality assurance
  - Object matching using profilers, statistical analysis, etc.
  - Correlation mining
  - Cross-relational data mining
  - Privacy-preserving data mining
  - And potentially many others!

www.cs.uiuc.edu/~hanj

Thank you !!!