Parameter-Free Determination of Distance Thresholds for Metric Distance Constraints

Shaoxu Song[†] Lei Chen[‡] Hong Cheng[§]

[†] Tsinghua University
Beijing, China
sxsong@tsinghua.edu.cn

[‡] The Hong Kong University of Science and Technology
leichen@cse.ust.hk

[§] The Chinese University of Hong Kong
hcheng@se.cuhk.edu.hk

ICDE 2012

Data Dependencies

- Recently used for capturing inconsistencies

$$\text{fd}_1 \quad [\text{Address}] \rightarrow [\text{Region}]$$

- t_5 and t_6, with the equal value on Address, but have different values of Region.

Example

ID	Name	Address	Region	
01	West Wood Hotel	Fifth Avenue, 61st Street	Chicago	t_1
01	West Wood	Fifth Avenue, 61st Street	Chicago, IL	t_2
01	West Wood (61)	5th Avenue, 61st St.	Chicago, IL	t_3
22	St. Regis Hotel	No.3, West Lake Road.	Boston, MA	t_4
22	St. Regis Hotel	#3, West Lake Rd.	Boston	t_5
22	St. Regis	#3, West Lake Rd.	Chicago, MA	t_6

Tolerance to Variations

- Real-world information often has various representation formats.
- The strict equality function limits the usage of FDs.

$$\text{fd}_1 \quad [\text{Address}] \rightarrow [\text{Region}]$$

- t_1 and t_2, detected as a "violation" by mistake.
  "Chicago" and "Chicago, IL" denote the same region.
- t_4 and t_6, are true violations.
  Cannot be detected by fd_1, as address values are not equal.

ID	Name	Address	Region	
01	West Wood Hotel	Fifth Avenue, 61st Street	Chicago	t_1
01	West Wood	Fifth Avenue, 61st Street	Chicago, IL	t_2
01	West Wood (61)	5th Avenue, 61st St.	Chicago, IL	t_3
22	St. Regis Hotel	No.3, West Lake Road.	Boston, MA	t_4
22	St. Regis Hotel	#3, West Lake Rd.	Boston	t_5
22	St. Regis	#3, West Lake Rd.	Chicago, MA	t_6

Metric Distance Constraints

In order to be tolerant to small variations

- Differential dependencies (DDs) declare the dependencies with respect to metric distances $(X \rightarrow Y, \varphi)$

$$\text{dd}_1 \quad ([\text{Address}] \rightarrow [\text{Region}], < 8, 3 >)$$

- $< 8, 3 >$ is a pattern φ of distance thresholds on Address and Region respectively.

States a constraint on metric distance:

- Any two tuples have distance on Address less than a threshold (≤ 8),
- then their Region values should be similar as well, i.e., the edit distance on Region is less than the corresponding threshold (≤ 3).

Motivation of This Work

Difficult to determine the proper settings of distance thresholds for metric distance constraints.

- Unlike FDs, already imply the equality function
- a very tight threshold (≈ 0 as FDs)
  too strict to be tolerant to various information formats
- a loose threshold ($\approx d_{\max}$ the maximum distance value)
  meaningless, since any data can satisfy it

In this study,

- employ certain statistical measures to evaluate the utility of distance threshold patterns
  e.g., support, confidence and dependent quality
- target on automatically determining the best settings of distance thresholds, having higher statistical measures.

Applicable to Other Types

Metric functional dependencies (MFDs)

- $X \xrightarrow{\delta} A$
- equality operator in the left-hand-side
- metric distance operator in the right-hand-side
- for violation detection
- e.g., manu $\xrightarrow{2}$ addr

Matching dependencies (MDs)

- $[X \approx X] \rightarrow [A \rightleftharpoons A]$
- similarity operator in the left-hand-side
- matching operator in the right-hand-side
- for record matching
- e.g., [addr $\approx$ addr] $\rightarrow$ [tel $\rightleftharpoons$ tel]

Outline

Statistical Measures

Support of φ:

- the proportion of tuple pairs whose distances satisfy the thresholds in $\varphi[XY]$.
- a φ with *high support* is preferred in order to detect more violations.

Confidence of φ:

- the ratio of tuple pairs satisfying $\varphi[XY]$ to the pairs satisfying $\varphi[X]$.
- a φ with *high confidence* is preferred to detect violations more precisely.

Dependent quality of φ denotes the quality of tolerance on the dependent attributes Y.

- how close the distance threshold $\varphi[Y]$ to the equality is.
- if the dependent quality is low (i.e., $\varphi[Y]$ is far away from equality), the constraint is meaningless and useless.

Interaction of Measures

If the dependent quality is set too high

- e.g., $\varphi[Y] = 0$, equality in FDs
- too strict and may identify violations by mistake
- confidence measure will be low

Contrarily, consider a φ with the lowest dependent quality

- i.e., $\varphi[Y] = d_{\max}$ the maximum distance value
- has the highest confidence 1.0, since any tuple pairs can always have distances $\leq d_{\max}$ on Y
- miss all the violations and is useless

For example, $([\text{Address}] \rightarrow [\text{Region}], < 8, d_{\max} >)$

- any pair of tuples always has distance on Region $\leq d_{\max}$
- the confidence is 1.0
- violations t_4 and t_6 cannot be detected by such a DD

Parameter-free Determination

To determine φ

- applications prefer metric distance constraints with high statistical measures
- difficult to set the parameters of minimum support, confidence and dependent quality, respectively
- setting the requirements of some measures too high will make the others low

A parameter-free style

- automatically returning those best φ
- s.t., not existing any other settings that can be found having higher support, confidence, and dependent quality than the returned results at the same time.

Assuring the Utility

To avoid tuning parameters manually, we are interested in an overall evaluation of utility.

- Let b be the matching distance of any tuple pair.

$$\mathrm{U}(\varphi) = \Pr(b \vDash \varphi[Y], \mathrm{Q}(\varphi) \text{ is high} \mid b \vDash \varphi[X])$$

- the conditional probability of b satisfying $\varphi[Y]$ with high dependent quality given b satisfies $\varphi[X]$.
- to accurately detect the violations with small distance, we expect the above probability of a φ to be high.

This $\mathrm{U}(\varphi)$ can roughly denote the utility of confidence and dependent quality, while support is not investigated.

Expected Utility

Compute an *expected utility* to refine $\mathrm{U}(\varphi)$ w.r.t. confidence and dependent quality by using support,

- $\bar{\mathrm{U}}(\varphi) = E(\mathrm{U}(\varphi) \mid \mathrm{C}(\varphi), \mathrm{D}(\varphi), \mathrm{Q}(\varphi))$,

$\mathrm{C}(\varphi), \mathrm{D}(\varphi)$ and $\mathrm{Q}(\varphi)$ are the statistics observed from data.

- $\mathrm{C}(\varphi)$ is confidence measure
- $\mathrm{D}(\varphi)$ is the proportion of tuple pairs with distance satisfying $\varphi[X]$, support of $\varphi[X]$
- support of φ is $\mathrm{C}(\varphi)\mathrm{D}(\varphi)$
- $\mathrm{Q}(\varphi)$ is dependent quality

Computation of Expected Utility

The computation is derived by applying the Bayesian rule and Binomial distribution.

$$\begin{aligned}\bar{\mathrm{U}}(\varphi) &= E(\mathrm{U} \mid \mathrm{C}, \mathrm{D}, \mathrm{Q}) \\ &= \int u P(\mathrm{U} = u \mid \mathrm{C}, \mathrm{D}, \mathrm{Q}) du \\ &\vdots \\ &= \frac{\int u f(\mathrm{DCQ}; \mathrm{D}, u) \pi(u) du}{\int f(\mathrm{DCQ}; \mathrm{D}, u) \pi(u) du}.\end{aligned}$$

- $f(k; n, p)$ is the probability mass function of Binomial distribution.

Property of Expected Utility

According to the calculation formula of $\bar{\mathrm{U}}(\varphi)$

Theorem

For any φ_1, φ_2, if φ_1 has higher support than φ_2, denoted by $\frac{\mathrm{S}(\varphi_1)}{\mathrm{S}(\varphi_2)} = \rho, \rho \geq 1$, and the confidence and dependent quality of φ_1 are higher than those of φ_2 as follows $\frac{\mathrm{C}(\varphi_1)}{\mathrm{C}(\varphi_2)} \geq \rho, \frac{\mathrm{Q}(\varphi_1)}{\mathrm{Q}(\varphi_2)} \geq \frac{1}{\rho}$, then we have $\bar{\mathrm{U}}(\varphi_1) \geq \bar{\mathrm{U}}(\varphi_2)$.

This conclusion verifies our intuition that

- higher support, confidence and dependent quality
- contribute to a larger expected utility.

Definition

The distance threshold determination problem is to find a distance threshold pattern φ for the DD on $X \rightarrow Y$ with the maximum expected utility $\bar{\mathrm{U}}(\varphi)$.

Outline

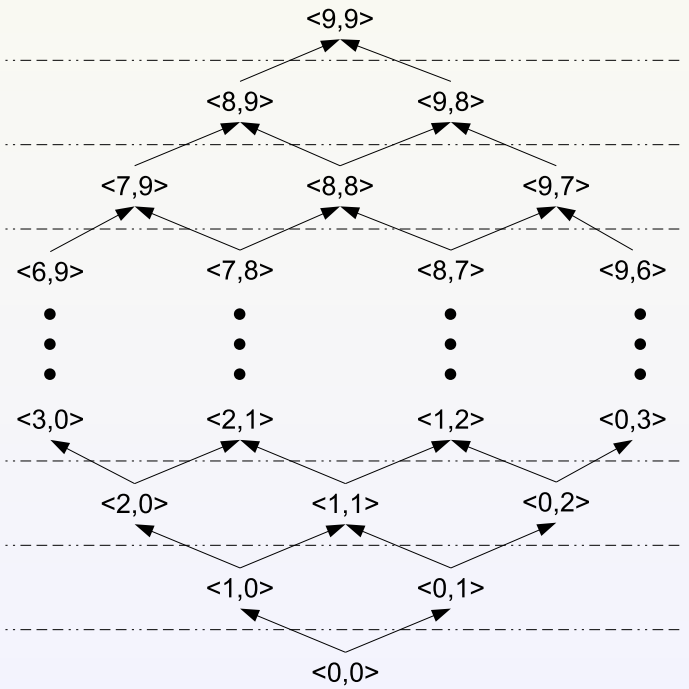
Overview

Determination process for the maximum $\bar{U}(\varphi)$ has two steps:

- **(i)** to find the best $\varphi[Y]$ when given a fixed $\varphi[X]$;
- **(ii)** to find the desired $\varphi[X]$ together with its best $\varphi[Y]$.

Candidate of distance threshold patterns, e.g., $\varphi[Y]$

- for each $A \in Y$, consider the search space of distance threshold $\varphi[A]$ from 0 to $d_{\max}$.
- enumerate all the distance thresholds $\varphi[A]$ for all the dependent attributes $A \in Y$.
- each node, such as $< 1, 1 >$, corresponds to a $\varphi[Y] \in \mathcal{C}_Y$



Determination for Dependent Attributes (PA)

Given a fixed $\varphi[X]$, to find the corresponding best $\varphi[Y]$ on the dependent attributes Y with the maximum $\bar{\mathrm{U}}(\varphi)$.

- $\mathrm{D}(\varphi)$ value is the same for any φ with same $\varphi[X]$.
- study the other two measures $\mathrm{C}(\varphi)$ and $\mathrm{Q}(\varphi)$ in terms of contributions to $\bar{\mathrm{U}}(\varphi)$.

Theorem

Consider any two φ_1, φ_2, *having the same* $\mathrm{D}(\varphi_1) = \mathrm{D}(\varphi_2) = \mathrm{D}$. *If their confidence and dependent quality satisfy* $\mathrm{C}(\varphi_1)\mathrm{Q}(\varphi_1) \geq \mathrm{C}(\varphi_2)\mathrm{Q}(\varphi_2)$, *then we have* $\bar{\mathrm{U}}(\varphi_1) \geq \bar{\mathrm{U}}(\varphi_2)$.

- for a fixed $\varphi[X]$,
- to find a φ with the maximum $\bar{\mathrm{U}}(\varphi)$ is equivalent to find the one with the maximum $\mathrm{C}(\varphi)\mathrm{Q}(\varphi)$.

Dominant Relationship for Pruning

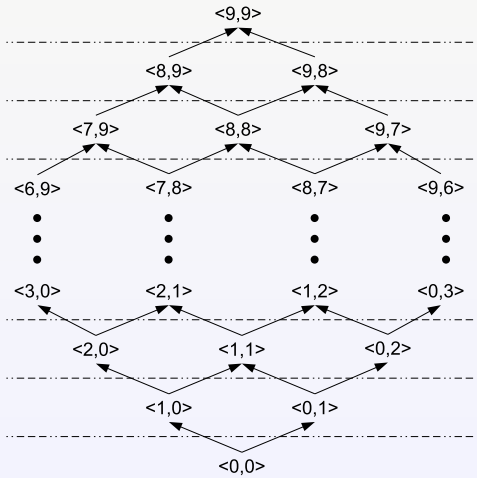
Pruning idea

- $\mathrm{Q}(\varphi)$ directly computed from a given $\varphi[Y]$
- $\mathrm{C}(\varphi)$ is costly to compute by statistics of data
- to avoid evaluate $\mathrm{C}(\varphi)$ for all possible candidates

Definition

For any φ_1, φ_2, if $\varphi_1[A] \geq \varphi_2[A]$, $\forall A \in Z$, then we say that $\varphi_1[Z]$ *dominates* $\varphi_2[Z]$, denoted by $\varphi_1[Z] \preccurlyeq \varphi_2[Z]$.

Any tuple pair satisfying $\varphi_2[Z]$ will always satisfy $\varphi_1[Z]$



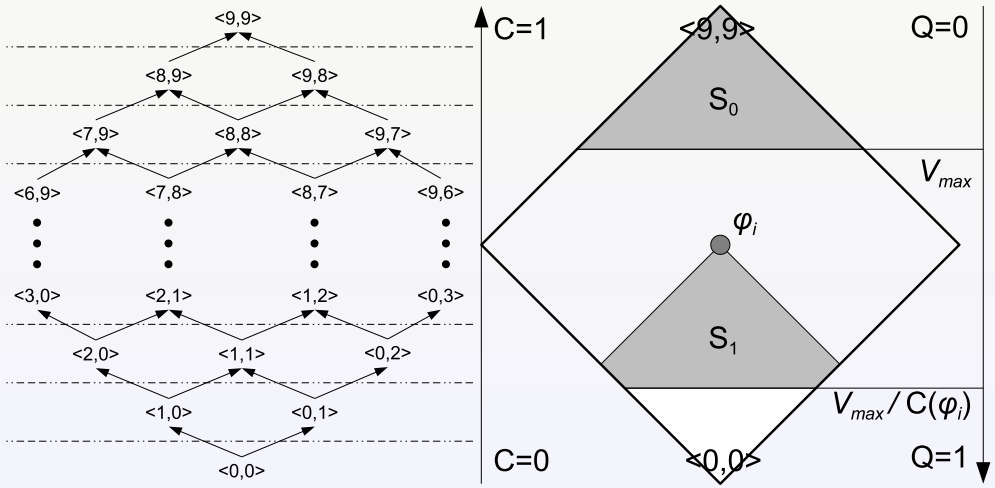
Dominant Relationship for Pruning

Lemma

For any two φ_1, φ_2*, having* $\varphi_1[X] = \varphi_2[X]$ *and* $\varphi_1[Y] \preccurlyeq \varphi_2[Y]$*, then* $\mathrm{C}(\varphi_1) \geq \mathrm{C}(\varphi_2)$ *and* $\mathrm{Q}(\varphi_1) \leq \mathrm{Q}(\varphi_2)$*.*

By a downward traversal of candidates in the dominant graph,

- the dependent quality increases from 0 to 1
- the confidence decreases from 1 to 0



Pruning of Candidate Patterns (PAP)

Consider the current φ_i in traversal of $\mathcal{C}_Y$.

i) *Pruning by* $\varphi_{\max}$.

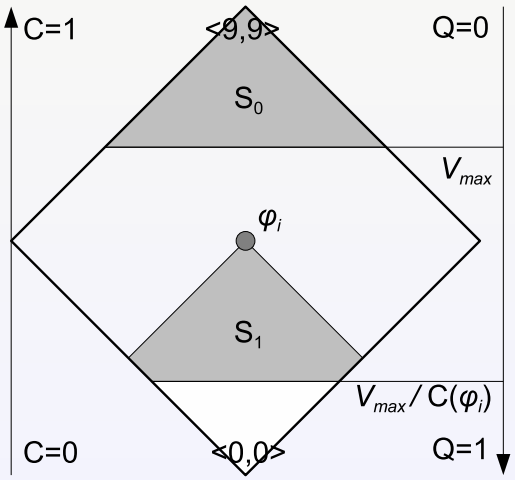
- The first pruning opportunity is introduced by $\varphi_{\max}$ of the previously processed $i-1$ candidates.
- Let $V_{\max}$ denote the maximum value of $\mathrm{C}(\varphi)\mathrm{Q}(\varphi)$ in the first $i-1$ candidates, i.e.,

$$V_{\max} = \max_{j=1}^{i-1} \mathrm{C}(\varphi_j)\mathrm{Q}(\varphi_j)$$

- $S_0 = \{\varphi_k \mid \mathrm{Q}(\varphi_k) \leq V_{\max}, \varphi_k[Y] \in \mathcal{C}_Y\}$ can be pruned

For any $\varphi_k[Y] \in \mathcal{C}_Y$ with $\mathrm{Q}(\varphi_k) \leq V_{\max}$,

- $\mathrm{C}(\varphi_k)\mathrm{Q}(\varphi_k) \leq \mathrm{Q}(\varphi_k) \leq V_{\max}$.
- $\bar{\mathrm{U}}(\varphi_{\max}) \geq \bar{\mathrm{U}}(\varphi_k)$.



Pruning of Candidate Patterns (PAP)

ii) *Pruning by* φ_i.

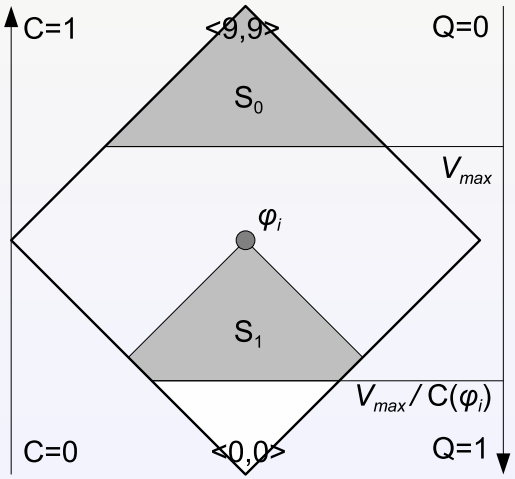
- The second pruning opportunity is developed according to the current φ_i in i-th step.
- $S_1 = \{\varphi_k \mid \varphi_i \lessdot \varphi_k, \mathrm{Q}(\varphi_k) \leq \frac{V_{\max}}{\mathrm{C}(\varphi_i)}, \varphi_k[Y] \in \mathcal{C}_Y\}$ is pruned

For any $\varphi_k[Y] \in \mathcal{C}_Y$ with $\varphi_i[Y] \lessdot \varphi_k[Y]$ and $\mathrm{Q}(\varphi_k) \leq \frac{V_{\max}}{\mathrm{C}(\varphi_i)}$,

- $\varphi_i[Y] \lessdot \varphi_k[Y]$ implies $\mathrm{C}(\varphi_k) \leq \mathrm{C}(\varphi_i)$
- follows $\mathrm{C}(\varphi_k)\mathrm{Q}(\varphi_k) \leq \mathrm{C}(\varphi_i)\mathrm{Q}(\varphi_k) \leq V_{\max}$
- we have $\bar{\mathrm{U}}(\varphi_{\max}) \geq \bar{\mathrm{U}}(\varphi_k)$

φ_k in S_0, S_1 can be safely pruned, without computing $\mathrm{C}(\varphi_k)$

- initialization of $V_{\max} = 0$



Determination for Determinant Attributes (DA)

To find a φ with the maximum $\bar{\mathrm{U}}(\varphi)$

- consider all possible distance threshold patterns of the determinant attributes X, say $\mathcal{C}_X$,
- The straight-forward approach is to compute the best $\varphi[Y]$ for each $\varphi[X] \in \mathcal{C}_X$
- The most costly part is still the computation of $\varphi_i[Y]$, by either PA or PAP.
- In order to improve the pruning power of PAP, we expect to find a larger pruning bound $V_{\max}$.

Pruning of Candidate Patterns

Pruning candidates with different $\varphi[X]$

Theorem

Consider any two φ_1, φ_2*, having* $\mathrm{D}(\varphi_1) \geq \mathrm{D}(\varphi_2)$*. If their confidence and dependent quality satisfy*

$$\mathrm{C}(\varphi_2)\mathrm{Q}(\varphi_2) \leq 1 - \frac{\mathrm{D}(\varphi_1)}{\mathrm{D}(\varphi_2)}\Big(1 - \mathrm{C}(\varphi_1)\mathrm{Q}(\varphi_1)\Big)$$

then we have $\bar{\mathrm{U}}(\varphi_1) \geq \bar{\mathrm{U}}(\varphi_2)$.

- We can prune those φ_2 whose $\mathrm{C}(\varphi_2)\mathrm{Q}(\varphi_2)$ is no higher than $1 - \frac{\mathrm{D}(\varphi_1)}{\mathrm{D}(\varphi_2)}\Big(1 - \mathrm{C}(\varphi_1)\mathrm{Q}(\varphi_1)\Big)$
- To apply this pruning bound, we require a precondition $\mathrm{D}(\varphi_1) \geq \mathrm{D}(\varphi_2)$.

Advanced Pruning Bound (DAP)

We process $\mathcal{C}_X$ in descending order of $\mathrm{D}(\varphi)$ values

- Let $\varphi_{\max}$ be the current result with the maximum expected utility by evaluating the first $i-1$ candidates in $\mathcal{C}_X$.
- for the next φ_i, we have $\mathrm{D}(\varphi_{\max}) \geq \mathrm{D}(\varphi_i)$

An advanced pruning bound for computing $\varphi_i[Y]$

$$V_{\max} = 1 - \frac{\mathrm{D}(\varphi_{\max})}{\mathrm{D}(\varphi_i)}\Big(1 - \mathrm{C}(\varphi_{\max})\mathrm{Q}(\varphi_{\max})\Big)$$

- in the original PAP, initialization of $V_{\max} = 0$
- replace with the above possibly large bound

Analysis of Pruning

Practically, the worst case of DAP is exactly the basic DA, when working together with PAP

- If the calculated bound $V_{\max}$ is less than 0, we can simply assign 0 to it.
- Once the bound is $V_{\max} > 0$, it can achieve a tighter pruning bound.

Theoretically, the theorem for advanced pruning is a generalization of the theorem for basic pruning

- when $\mathrm{D}(\varphi_1) = \mathrm{D}(\varphi_2)$,

$$1 - \frac{\mathrm{D}(\varphi_1)}{\mathrm{D}(\varphi_2)}\Big(1 - \mathrm{C}(\varphi_1)\mathrm{Q}(\varphi_1)\Big) = \mathrm{C}(\varphi_1)\mathrm{Q}(\varphi_1)$$

Our experiments also verify that DAP+PAP is at least no worse than DA+PAP.

Outline

Settings

Preprocessing of three real data sets

- pre-compute edit distance of all tuple pairs
- store the distance results as up to 1,000,000 matching tuples
- proposed techniques are then evaluated on the prepared matching tuples

To determine the distance thresholds for

$Rule1 : cora(\text{author}, \text{title} \rightarrow \text{venue}, \text{year})$

$Rule2 : cora(\text{venue} \rightarrow \text{address}, \text{publisher}, \text{editor})$

$Rule3 : restaurant(\text{name}, \text{address} \rightarrow \text{city}, \text{type})$

$Rule4 : citeseer(\text{address}, \text{affiliation}, \text{description} \rightarrow \text{subject})$

where Rule 2 has a larger Y while Rule 4 has a larger X.

Example Results

Results also verify our property analysis of expected utility

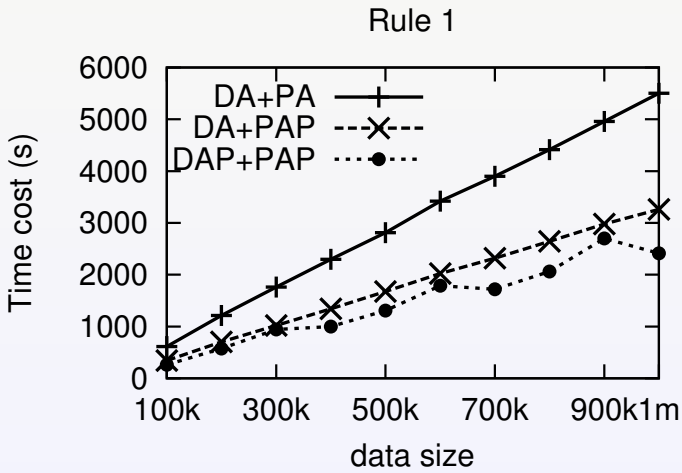
- higher support, confidence and dependent quality yield higher expected utility
- e.g., $\bar{\mathrm{U}}(\varphi_2) \geq \bar{\mathrm{U}}(\varphi_4)$
- there does not exist any φ which has higher support, confidence and dependent quality at the same time than the returned φ_1 with the maximum expected utility
- the expected utility can reflect the usefulness in applications

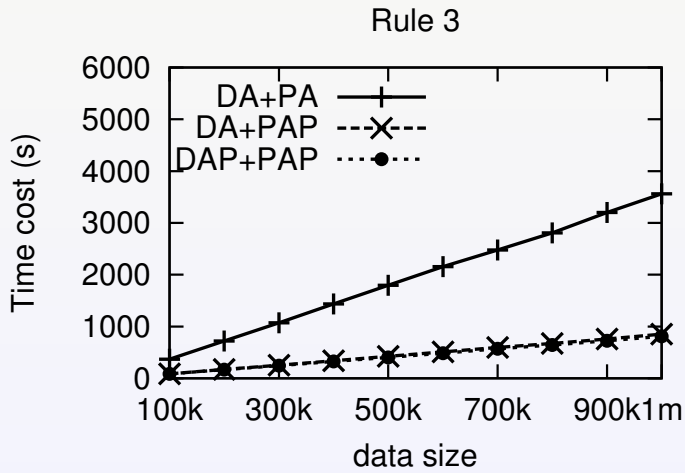
	$\varphi[X]$		$\varphi[Y]$		Measures				Violation Detection		
	author	title	venue	year	$\mathrm{S}(\varphi)$	$\mathrm{C}(\varphi)$	$\mathrm{Q}(\varphi)$	$\mathrm{U}(\varphi)$	Precision	Recall	F-measure
φ_1	4	1	3	1	0.1529	0.3760	0.80	0.2325	0.3725	0.5425	0.4418
φ_2	5	2	3	1	0.1764	0.3667	0.80	0.2296	0.3718	0.6266	0.4667
φ_3	5	1	3	2	0.1632	0.3774	0.75	0.2232	0.3179	0.4492	0.3723
φ_4	4	2	3	2	0.1657	0.3657	0.75	0.2188	0.3073	0.4457	0.3638
φ_5	4	1	4	2	0.1529	0.3852	0.70	0.2108	0.2654	0.3267	0.2928
φ_6	5	2	5	2	0.1764	0.3985	0.65	0.2106	0.2459	0.3337	0.2831
fd	0	0	0	0	0.0064	0.3595	1.00	0.1064	0.4315	0.0735	0.1256

Pruning Evaluation

Performance of DA and DAP for determinant side X, PA and PAP for dependent side Y

- DAP and PAP outperform DA and PA, respectively
  Rule 1 shows best performance when applying DAP+PAP
- DAP+PAP approach can provide a pruning bound that is at least no worse than the DA+PAP one
  Rule 3 verifies that the DAP is at least no worse than the DA

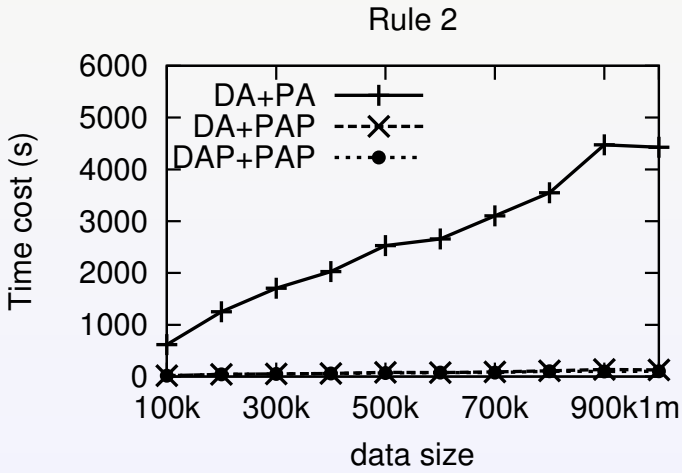


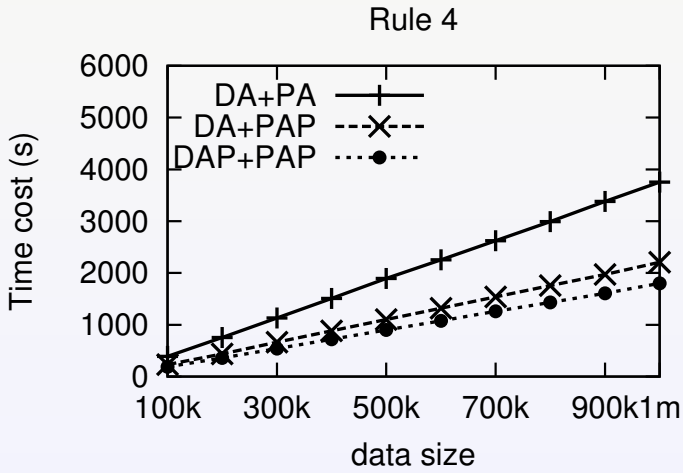


Pruning Evaluation

Rule 2 has a larger Y while Rule 4 has a larger X

- Rule 2, which has more attributes in the dependent side, may have more opportunities of pruning by PAP
  PAP can achieve a significant improvement in Rule 2
- Rule 4, with smaller Y, is not as significant as Rule 2 on the improvement by PAP
  DAP do help in providing an advanced pruning bound for PAP





Conclusion

We study the problem of determining the distance thresholds for metric distance constraints

- difficult to manually specify requirements of various statistical measures
- conduct the determination in a parameter-free style
- i.e., to compute an expected utility of the distance threshold pattern and return the results with the maximum expected utility
- several advanced pruning algorithms are then developed in order to efficiently find the desired distance thresholds