

User Profiling
in
Relational Database Systems

Christina Chung
chungy@cs.ucdavis.edu
Department of Computer Science
University of California, Davis

May 3, 1999

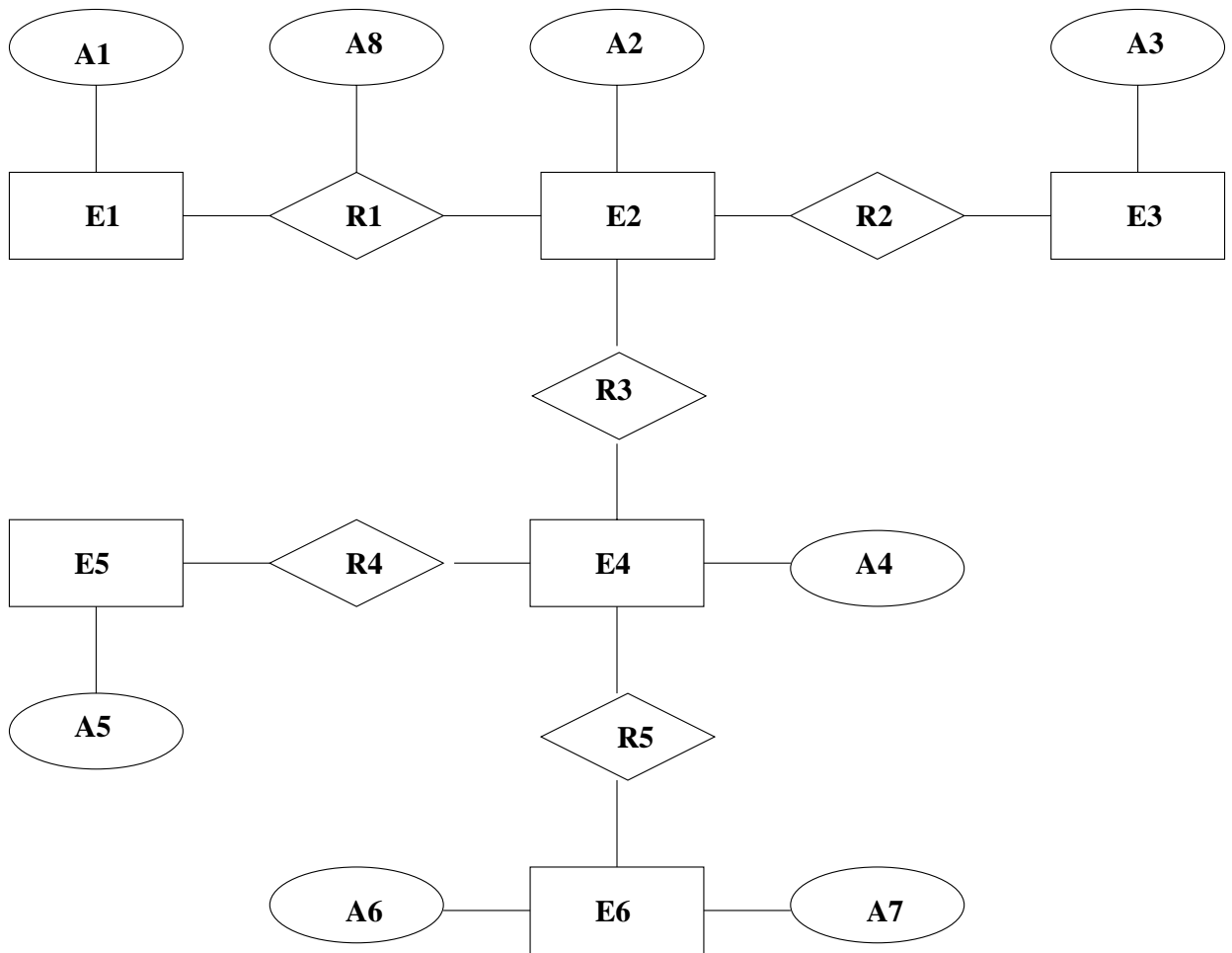
Outline

- What I Am Doing
- ◇ System Environment
- ◇ Goal
- ◇ The Profiler
- ◇ Profiler Based On Clustering
- ◇ Profiler Based On Association Rules
- ◇ Profiler Based On Frequent Itemsets
- ◇ Discussion

What I Am Doing

Discover profiles for users and roles
in a relational database system.

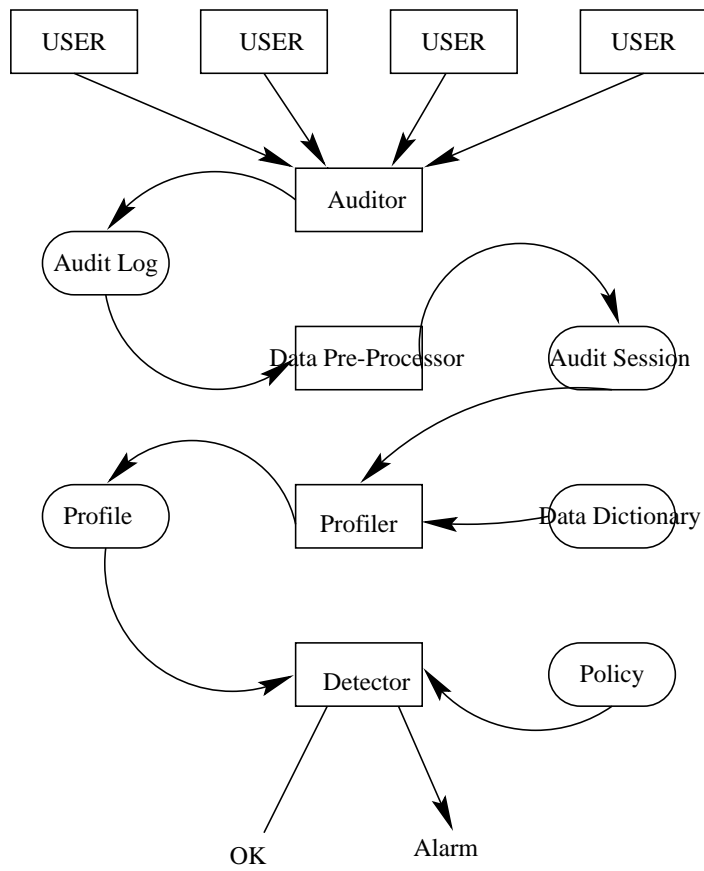
An E-R Diagram For A DB Schema



Outline

- ◇ What I Am Doing
- System Environment
- ◇ Goal
- ◇ The Profiler
- ◇ Profiler Based On Clustering
- ◇ Profiler Based On Association Rules
- ◇ Profiler Based On Frequent Itemsets
- ◇ Discussion

System Environment - Architecture



Architecture

System Environment - The Auditor

Function: User Applications + Free Form Queries
→ Audit Log

What features to audit?

- ◇ user ID, timestamp
- ◇ what query: select, insert, delete, update
- ◇ what attributes are referenced
- ◇ new and old values for the attributes for update query

System Environment - The Auditor

Example of Audit Log

TID	FEATURE	FVALUE
1	user	'tom'
1	timestamp	1000000
1	queryType	'select'
1	E1.A1	1
1	E1.A1.oldValue	NULL
1	E1.A1.newValue	NULL
1	R1.A8	0
1	R1.A8.oldValue	NULL
1	R1.A8.newValue	NULL
...
51	user	'john'
51	timestamp	1000010
51	queryType	'update'
51	E1.A1	1
51	E1.A1.oldValue	'old'
51	E1.A1.newValue	'new'
51	R1.A8	1
51	R1.A8.oldValue	NULL
51	R1.A8.newValue	NULL

System Environment - The Data Preprocessor

Function: Audit Log $\rightarrow_{p_{session}}$ Audit Sessions

Examples of predicates $p_{session}$:

◇ According to user ID

$$P_{session} \equiv \{user='tom', user='john'\}$$

◇ According to timestamp

$$P_{session} \equiv \{Day(timestamp)='Mo' \vee 'Tu' \vee 'We' \vee 'Th' \vee 'Fr', Day(timestamp)='Sa' \vee 'Su'\}$$

System Environment - The Data Preprocessor

$P_{session} \equiv \{user='tom', user='john'\}$

Audit Session under $p_{session_1} \equiv user='tom'$

TID	FEATURE	FVALUE
1	timestamp	1000000
1	queryType	'select'
1	E1.A1	1
1	E1.A1.oldValue	NULL
1	E1.A1.newValue	NULL
1	R1.A8	0
1	R1.A8.oldValue	NULL
1	R1.A8.newValue	NULL
...

System Environment - The Data Preprocessor

$P_{session} \equiv \{user = 'tom', user = 'john'\}$

Audit Session under $p_{session_2} \equiv user = 'john'$

TID	FEATURE	FVALUE
51	timestamp	1000010
51	queryType	'update'
51	E1.A1	1
51	E1.A1.oldValue	'old'
51	E1.A1.newValue	'new'
51	R1.A8	1
51	R1.A8.oldValue	NULL
51	R1.A8.newValue	NULL
...

System Environment - The Profiler

Function: Audit Sessions + Data Dictionary → Profiles

- ◇ A profile is a set of rules
precondition → *consequent*
that give typical values of the features in Audit Sessions
- ◇ Data Dictionary gives the data structure (e.g. the relation schema)

Examples of profiles:

- ◇ $user='tom' \wedge queryType='select' \rightarrow E1.A1=1 \wedge E2.A2=1$
- ◇ $user='john' \wedge queryType='update' \wedge E1.A1=1 \rightarrow E1.A1.oldValue='old' \wedge E1.A1.newValue='new' \wedge Day(timestamp)='Sa'$

System Environment - The Detector

Function: Profiles + Policy / System Parameters
→ Suspicion Level

- ◇ Policy expressed as rules *precondition* → *consequent* can specify the expected user behavior or known malicious user behavior in terms of feature values
- ◇ System parameters are specific to the Profiler
- ◇ Suspicion Level gives a score of how suspicious an audit session is.

Goal

Given Audit Sessions, Data Dictionary, Policy, System Parameters,
how can we discover profiles for users and roles that are complete, minimal and accurate?

Outline

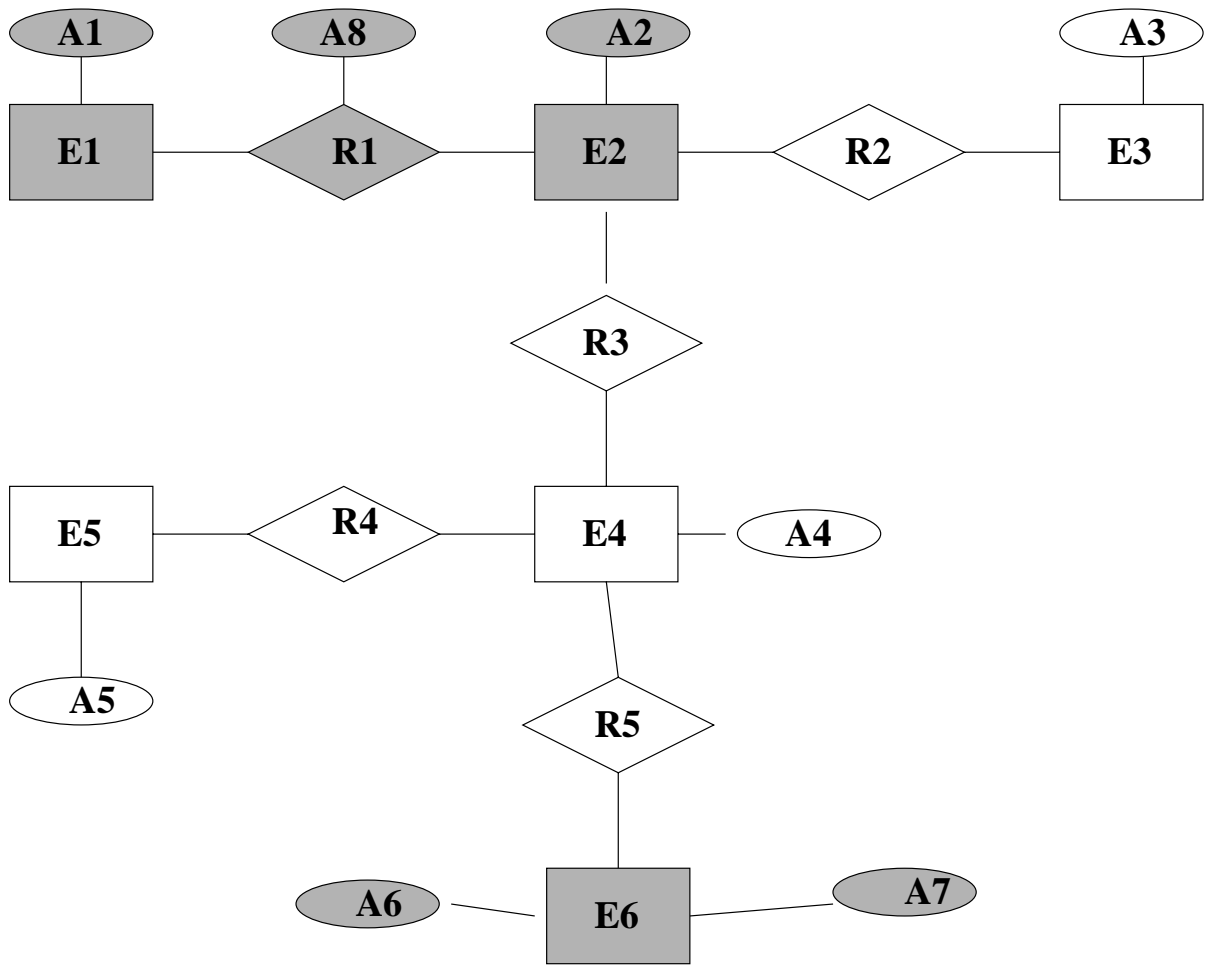
- ◇ What I Am Doing
- ◇ System Environment
- ◇ Goal
- The Profiler
 - ◇ Profiler Based On Clustering
 - ◇ Profiler Based On Association Rules
 - ◇ Profiler Based On Frequent Itemsets
 - ◇ Discussion

The Profiler

Differences between profiling at o/s and DBMS levels

- ◇ security threat
- ◇ command set
- ◇ querying capability
- ◇ domain knowledge

Working Scopes



Distance Measure

- ◇ We define distance measure between a set attributes.
- ◇ More related the attributes in the database schema → smaller the distance measure
- ◇ More often they are referenced together → smaller the distance measure
- ◇ Attributes with smaller distance measure → more likely they are in the same working scope

Distance Measure

◇ *PairwiseSchemaDistance*(a_i, a_j)

$$\frac{\text{ShortestDist}(\text{Node}(a_i), \text{Node}(a_j))}{\text{Max}_{\text{Node}(a_p), \text{Node}(a_q) \in V[G]} \{\text{ShortestDist}(\text{Node}(a_p), \text{Node}(a_q))\}}$$

◇ *SchemaDistance*(a_1, \dots, a_n)

$$\text{Max}_{1 \leq i, j \leq n} \{\text{PairwiseSchemaDistance}(a_i, a_j)\}$$

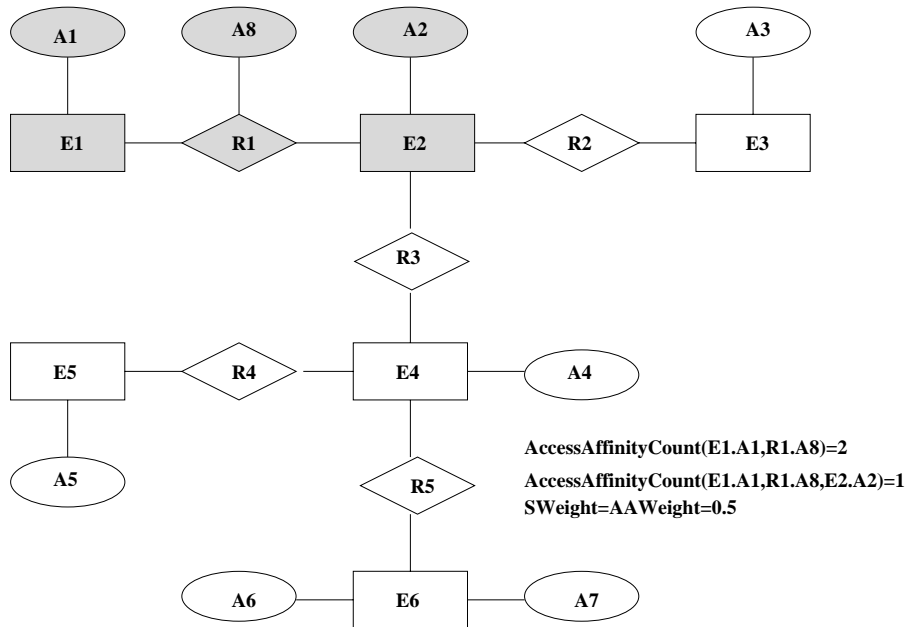
◇ *AccessAffinity*(a_1, \dots, a_n)

$$\frac{\text{AttributeAffinityCount}(a_1, \dots, a_n)}{\text{Max}_{a_{i_1} \dots a_{i_m} \in \text{SCHEMA.R}} \{\text{AttributeAffinityCount}(a_{i_1}, \dots, a_{i_m})\}}$$

◇ *Dist*(a_1, \dots, a_n)

$$\begin{aligned} & \text{SWeight} \times \text{SchemaDistance}(a_1, \dots, a_n) \\ & + \text{AAWeight} \times (1 - \text{AccessAffinity}(a_1, \dots, a_n)) \end{aligned}$$

Distance Measure



$\text{PairwiseSchemaDistance}(E1.A1,R1.A8)=1/6=0.17$
 $\text{PairwiseSchemaDistance}(E1.A1,E2.A2)=2/6=0.33$
 $\text{PairwiseSchemaDistance}(R1.A8,E2.A2)=1/6=0.17$
 $\text{SchemaDistance}(E1.A1,R1.A8,E2.A2)=\max\{0.17,0.33,0.17\}=0.33$
 $\text{PairwiseSchemaDistance}(E1.A1,E6.A6)=6/6=1.00$

$\text{AccessAffinity}(E1.A1,R1.A8)=2/2=1.00$
 $\text{AccessAffinity}(E1.A1,E2.A2)=0/2=0.00$
 $\text{AccessAffinity}(R1.A8,E2.A2)=0/2=0.00$
 $\text{AccessAffinity}(E1.A1,R1.A8,E2.A2)=1/2=0.5$
 $\text{AccessAffinity}(E1.A1,E6.A6)=0/2=0.00$

$\text{Dist}(E1.A1,R1.A8)=0.5*0.17+0.5*(1-1)=0.09$
 $\text{Dist}(E1.A1,E2.A2)=0.5*0.33+0.5*(1-0)=0.67$
 $\text{Dist}(R1.A8,E2.A2)=0.5*0.17+0.5*(1-0)=0.58$
 $\text{Dist}(E1.A1,E6.A6)=0.5*1.00+0.5*(1-0)=1.00$
 $\text{DIst}(E1.A1,R1.A8,E2.A2)=0.5*0.33+0.5*(1-0.5)=0.42$

The Profiler

Profiling at two levels

- ◇ Data level - by looking at the actual values of the features in Audit Sessions
 - ♡ Clustering
 - ♡ Association rules
 - ♡ Frequent itemsets
- ◇ Query formulation level - by looking at the way SQL queries are formed

Outline

- ◇ What I Am Doing
- ◇ System Environment
- ◇ Goal
- ◇ The Profiler
- Profiler Based On Clustering
- ◇ Profiler Based On Association Rules
- ◇ Profiler Based On Frequent Itemsets
- ◇ Discussion

Profiler Based On Clustering

Motivation

- ◇ Working scopes are sets of attributes that are usually referenced together in the queries.
- ◇ A cluster is a collection of objects that are close to each other under some distance measure.
- ◇ So use clusters to approximate the working scopes.

Profiler Based On Clustering

Input

◇ We audit what attributes are referenced by select queries.

◇ Example:

E1.A1	R1.A8	E2.A2	E6.A6	E6.A7
1	1	1	0	0

◇ Profiling parameters specify

♡ which clustering algorithm is used

♡ the threshold for clustering (maximum diameter of each cluster)

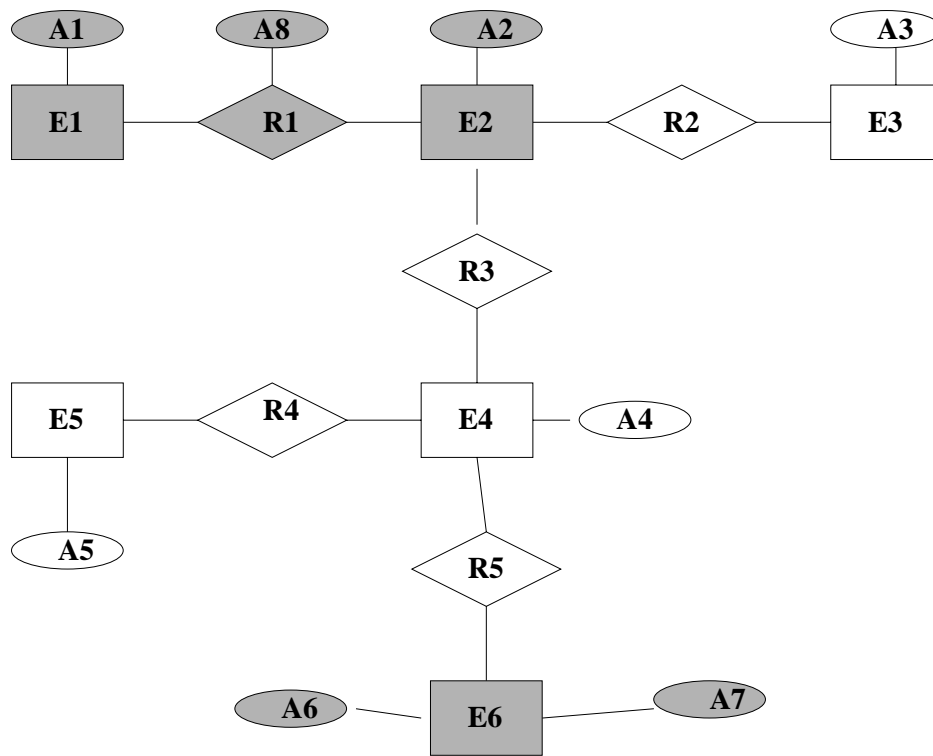
♡ the total number of clusters desired

Profiler Based On Clustering

Example

- ◇ $p_{session} \equiv \text{user}='tom' \wedge \text{queryType}='select'$
- ◇ *clusters*
 - ♡ {E1.A1, R1.A8, E2.A2}
 - ♡ {E6.A6, E6.A7}
- ◇ *profiles*
 - $\text{user}='tom' \wedge \text{queryType}='select' \rightarrow$
 $E1.A1=1 \wedge R1.A8=1 \wedge E2.A2=1$
 - $\text{user}='tom' \wedge \text{queryType}='select' \rightarrow$
 $E6.A6=1 \wedge E6.A7=1$

Profiler Based On Clustering



Clusters

{E1.A1, E2.A2, R1.A8}

{E6.A6, E6.A7}

Outline

- ◇ What I Am Doing
- ◇ System Environment
- ◇ Goal
- ◇ The Profiler
- ◇ Profiler Based On Clustering
- Profiler Based On Association Rules
- ◇ Profiler Based On Frequent Itemsets
- ◇ Discussion

Profiler Based On Association Rules

Motivation

- ◇ Association rules proposed by Agrawal are of the form

$$X_1 = x_1 \wedge \dots \wedge X_n = x_n \rightarrow$$

$$Y_1 = y_1 \wedge \dots \wedge Y_m = y_m,$$

with support sup , confidence conf

- ◇ We extend the association rules by incorporating the concept of distance measure to discover the working scopes

Input

- ◇ Features audited: query type, attributes referenced, their old and new values (if it is an update query)
- ◇ Profiling parameters: support, confidence and distance measure threshold

Profiler Based On Association Rules

Example

◇ $p_{session} \equiv \text{user}='tom' \wedge \text{Session}(\text{timestamp})='morning'$

◇ *rules*

queryType='select' →
E1.A1=1 ∧ R1.A8=1 ∧ E2.A2=1
sup=10, conf=0.9, dist=0.1

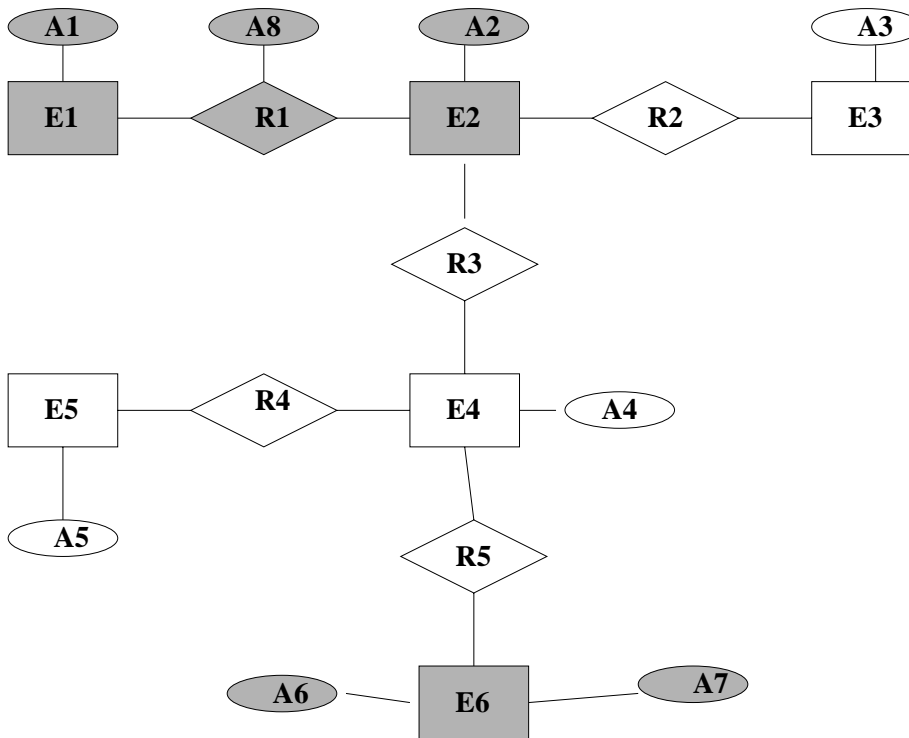
queryType='select' →
E6.A6=1 ∧ E6.A7=1
sup=1000, conf=0.6, dist=0.2

◇ *profiles*

user='tom' ∧ Session(timestamp)='morning'
∧ queryType='select' →
E1.A1=1 ∧ R1.A8=1 ∧ E2.A2=1
sup=10, conf=0.9, dist=0.2

user='tom' ∧ Session(timestamp)='morning'
∧ queryType='select' →
E6.A6=1 ∧ E6.A7=1
sup=1000, conf=0.6, dist=0.1

Profiler Based On Association Rules



Association Rules

queryType='select' -> E1.A1=1 and R1.A8=1 and E2.A2=1

queryType='select' -> E6.A6=1 and E6.A7=1

queryType='select' and E1.A1=1 -> R1.A8=1 and E2.A2=1

Outline

- ◇ What I Am Doing
- ◇ System Environment
- ◇ Goal
- ◇ The Profiler
- ◇ Profiler Based On Clustering
- ◇ Profiler Based On Association Rules
- Profiler Based On Frequent Itemsets
- ◇ Discussion

Profiler Based On Frequent Itemsets

Motivation

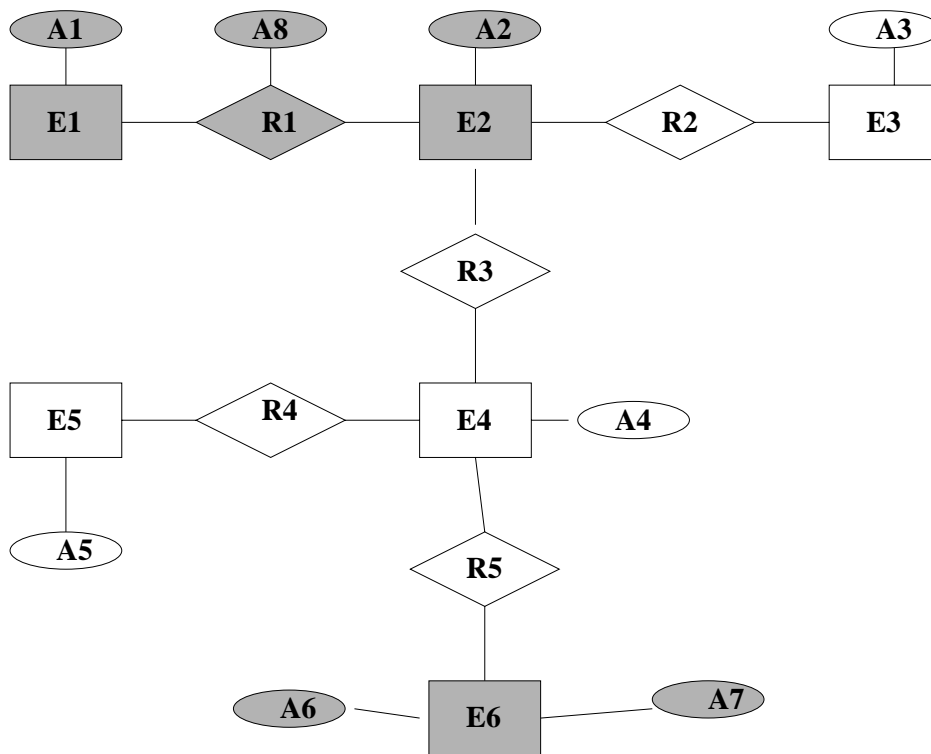
- ◇ clustering
 - ♡ distance measure
 - ♡ non-scalable, do not use DBMS query processing power
 - ♡ clusters are organized into a hierarchy
- ◇ association rules
 - ♡ scalable (implemented by SQL queries)
 - ♡ rules with different combinations of items in the precondition may be redundant?
 - ♡ rules do not have hierarchical structure
- ◇ we want: scalability, customized distance measure, fast algorithms → frequent itemsets
 - ♡ They are sets of features with certain values having certain support, confidence and distance measure
 - ♡ $FI = \{F_1 = F_1Val, \dots, F_m = F_mVal\}$
sup, dist
 - ♡ They are generated by hierarchical clustering algorithms using SQL queries

Profiler Based On Frequent Itemsets

Example

- ◇ $p_{session} \equiv \text{user}='tom \wedge \text{Season}(\text{timestamp})='spring'$
- ◇ FI
{queryType='select', E1.A1=1, R1.A8=1, E2.A2=1},
{queryType='select', E6.A6=1, E6.A7=1}
- ◇ $profile$
user='tom \wedge Season(timestamp)='spring' \rightarrow
queryType='select' \wedge E1.A1=1 \wedge R1.A8=1 \wedge E2.A2=1
user='tom \wedge Season(timestamp)='spring' \rightarrow
queryType='select' \wedge E6.A6=1 \wedge E6.A7=1

Profiler Based On Frequent Itemsets



Frequent Itemsets

{queryType='select',E1.A1=1,R1.A8=1,E2.A3=1}

{queryType='select',E6.A6=1,E6.A7=1}

Discussion

Please Comment!