Answering Queries from Statistics and Probabilistic Views

Nilesh Dalvi and Dan Suciu,
University of Washington.

Background

- **'Query answering using Views'** problem: find answers to a query q over a database schema R using a set of views $V = \{v_1, v_2 \cdots\}$ over R.
- **Example**: R(name,dept,phone)

$v_1(n,d) : R(n,d,p)$

$v_1=$

NAME	DEPT
Larry	Sales
John	Sales

$v_2(d,p): R(n,d,p)$

$v_2=$

DEPT	PHONE
Sales	x1234
Sales	x5678
HR	x2222

$q(p) : R(\text{Larry},d,p)$

Background: *Certain Answers*

Let U be a finite universe of size n. Consider all possible data instances over *U*

D_1 D_2 D_3 D_4 .... D_m

Data instances consistent with the views *V*

~~D_1~~ D_2 D_3 ~~D_4~~ .... D_m

Certain Answers: tuples that occur as answers in all data instances consistent with *V*

Example

$v_1(n,d) : R(n,d,p)$

v_1 =

NAME	DEPT
LARRY	SALES
JOHN	SALES

$v_2(d,p): R(n,d,p)$

v_2=

DEPT	PHONE
SALES	x1234
SALES	x5678
HR	x2222

$q(p) : R(\text{LARRY},d,p)$

Data instances consistent with the views:

D_1=

NAME	DEPT	PHONE
LARRY	SALES	x1234
JOHN	SALES	x5678
SUE	HR	x2222

D_2=

NAME	DEPT	PHONE
FRANK	SALES	x5678
LARRY	SALES	x1111
JOHN	SALES	x1234
SUE	HR	x2222

....

Example (contd.)

v_1 =

NAME	DEPT
LARRY	SALES
JOHN	SALES

v_2=

DEPT	PHONE
SALES	x1234
SALES	x5678
HR	x2222

- No certain answers, but some answers are more likely that others.
- Domain is huge, cannot just guess LARRY's number.
- A data instance is much smaller. If we know average employes per dept = 5, then x1234 and x5678 have 0.2 probability of being answer.

Going beyond certain answers

- *Certain answers* approach assumes complete ignorance about the knowledge of how likely is each possible database
- Often we have additional knowledge about the data in form of various statistics

Can we use such information to find answers to queries that are *statistically meaningful*?

Why Do We Care?

- **Data Privacy**: publishers can analyze the amount of information disclosed by public views about private information in the database
- **Ranked Search**: a ranked list of probable answers can be returned for queries with no certain answers.

Using Common Knowledge

- Suppose we have a priori distribution Pr over all possible databases:

  $Pr: \{D_1, \ldots, D_m\} \rightarrow [0,1]$

- We can compute the probability of a tuple t being an answer to q using $\Pr[(t \in q) \mid V]$

Query Answering using views = Computing conditional probabilities on a distribution

Part I

Query answering using views under some specific distributions

Binomial Distribution

U : a domain of size n

We start from a simple case

- R(name,dept,phone) a relation of arity 3
- Expected size of R is c

Binomial: Choose each of the n^3 possible tuples independently with probability p.

Expected size of R is c $\Rightarrow$ $p = c/n^3$

Let μ_n denote the resulting distribution. For any instance D,

$$\mu_n[D] = p^k(1-p)^{n^3 - k}, \text{ where } k = |D|$$

Binomial: Example I

R(name,dept,phone) $|R| = c$, domain size = n

v : R(Larry, -, -)

q : R(-, -, x1234)

$\mu_n[q \mid v] \approx (c+1)/n$ = negligible if n is large

$\lim_{n \to \infty} \mu_n[q \mid v] = 0$

v gives negligible information about q when domain is large

Binomial: Example II

R(name,dept,phone) |R| = c, domain size = n

v : R(LARRY, -, -), R(-, -, x1234)

q : R(LARRY, -, x1234)

$\lim_{n \to \infty} \mu_n[q \mid v] = 1/(1+c)$

v gives non-negligible information about q, even for large domains

Binomial: Example III

R(name,dept,phone)　　|R| = c, domain size = n

v : R(Larry, Sales, -), R(-, Sales, x1234)

q : R(Larry, Sales, x1234)

$\lim_{n \to \infty} \mu_n[q \mid v] = 1$

Binomial distribution cannot express more interesting statistics.

A Variation on Binomial

- Suppose we have following statistics on R(name,dept,phone):
  - Expected number of distinct R.dept = c_1
  - Expected number of distinct tuples for each R.dept = c_2
- Consider the following distribution μ_n
  - For each $x_d \in U$, choose it as a R.dept value with probability c_1/n
  - For each x_d chosen above, for each $(x_n,x_p) \in U^2$, include the tuple (x_n,x_d,x_p) in R with probability c_2/n^2

Examples

R(name,dept,phone) |dept|=c_1, |dept ⇒ name,phone| = c_2, |R|=$c_1 c_2$

Example 1:

v : R(Larry, -, -), R(-, -, x1234)
q : R(Larry, -, x1234)
$\mu[q \mid v] = 1/(c_1 c_2+1)$

Example 2:

v : R(Larry, sales, -), R(-, sales, x1234)
q : R(Larry, sales, x1234)
$\mu[q \mid v] = 1/(c_2+1)$

Part II : Representing Knowledge as a Probability Distribution

Knowledge about data

- A set of statistics Γ on the database
  - cardinality statistics : $card_R[A] = c$
  - fanout statistics: $fanout_R[A \Rightarrow B] = c$
  -
- A set of integrity constraints Σ
  - functional dependencies: $R.A \rightarrow R.B$
  - inclusion dependencies: $R.A \subseteq R.B$

Representing Knowledge

Statistics and constraints are statements on the probability distribution P

- $card_R[A] = c$ implies the following

  $$\Sigma_i P[D_i] \, card(\Pi_A(R^{D_i})) = c$$

- $fanout_R[A \Rightarrow B]$ implies a similar condition
- A constraint Σ implies that $P[D_i] = 0$ on data instances D_i that violate Σ

Problem: P is not uniquely defined by these statements!

The Maximum Entropy Principle

- Among all the probability distributions that satisfy Σ and Γ, choose the one with maximum entropy.
- Widely used to convert prior information into prior probability distribution
- Gives a distribuion that commits the least to any specific instance while satisfying all the equations.

Examples of Entropy Maximization

- R(name,dept,phone) a relation of arity 3
- Example 1:

  Γ = empty, $\Sigma = \{ \text{card}[R] = c \}$

  Entropy maximizing distribution = Binomial
- Example 2:

  Γ = empty, $\Sigma = \{ \text{card}_R[\text{dept}] = c_1,$
  $\text{fanout}_R[\text{dept} \Rightarrow \text{name,phone}] = c_2 \}$

  Entropy maximizing distribution = variation on Binomial distribution we studies earlier.

Query answering problem

Given a set of statistics Σ and constraints Γ, let $\mu_{\Sigma,\Gamma,n}$ denote the maximum entropy distribution assuming a domain of size n.

Problem: Given statistics Σ, constraints Γ, and boolean conjunctive queries q and v, compute the asymptotic limit of $\mu_{\Sigma,\Gamma,n}[q \mid v]$ as $n \to \infty$

Main Result

- For Boolean conjunctive queries q and v, the quantity $\mu_{\Sigma,\Gamma,n}[q \mid v]$ always has an asymptotic limit and we show how to compute it.

Glimpse into Main Result

- For any conjunctive query Q, we show that $\mu_{\Sigma,\Gamma,n}[Q]$ is a polynomial of the form

$$c_1(1/n)^d + c_2(1/n)^{d+1} + \ldots$$

- $\mu_{\Sigma,\Gamma,n}[q \mid v] = \mu_{\Sigma,\Gamma,n}[qv]/\mu_{\Sigma,\Gamma,n}[v]$ = ratio of two polynomials.
- Only the leading coefficient and exponent matter, and we show how to compute them.

Conclusions

- We show how to use common knowledge about data to find answers to queries that are statistically meaningful
  - Provides a formal framework for studying database privacy breaches using statistical attacks.
- We use the principle of entropy maximization to represent statistics as a prior probability distribution.
- The techniques are also applicable when the contents of views are themselves uncertain.

Questions?