CERIAS

K-anonymity using Clustering*

Elisa Bertino, Ninghui Li, Ji-Won Byun and Ashish Kamra
{bertino, ninghui, byunj, akamra}@cs.purdue.edu

Objective : To achieve k-anonymization of data with minimal loss in data quality

Anonymization Process

Quasi Identifier | Sensitive

Age	Sex	Zip code	Diagnosis
18	M	47803	Flu
29	F	56789	HIV
34	F	56763	Cancer
15	M	47567	Asthma

Clustering

Age	Sex	Zip code	Diagnosis
18	M	47803	Flu
15	M	47567	Asthma
34	F	56763	Cancer
29	F	56789	HIV

Anonymization

Age	Sex	Zip code	Diagnosis
(18,15)	M	(47803,47567)	Flu
(18,15)	M	(47803,47567)	Asthma
(34,29)	F	(56763,56789)	Cancer
(34,29)	F	(56763,56789)	HIV

Related Concepts

Distance Measures

Distance between two records, r_1 and r_2 :

$\Delta(r_1, r_2) = \Sigma_{i=1,\ldots m} \delta_N (r_1[N_i], r_2[N_i]) + \Sigma_{j=1,\ldots,n} \delta_C (r_1[C_j], r_2[C_j])$,

where $r(N_i)$ is the i-th numerical attribute of r, $r(C_j)$ is the j-th categorical attribute, and δ_N and δ_C are a numerical and categorical distance metric, respectively.

Clustering Algorithm

Given a set of *n* records,

1. Randomly pick a record r_c.
2. Make a cluster e_i which contains only r_c.
3. Include to e_i a record r_j that makes $IL(e_i \cup \{r_j\})$ minimal.
4. Repeat Step 2 until $|e_i| = k$.
5. Choose a record that is furthest from r_c and go to Step 2.
6. Repeat Steps 1-4 until there are less than *k* records left.
7. Iterate over the leftover records and insert each record into a cluster with respect to which the increment of the information loss is minimal.

Information Loss

Total information loss of anonymizing table T :

$\text{Total-IL}(T) = \Sigma_{i=1,\ldots,n} IL(e_i)$,

where IL(e) is the information loss (i.e., data distortion) of a cluster e, defined as $IL(e) = |e| * \Sigma_{j=1,\ldots,m} \max\text{-}\delta(at`tr_j)$.

Future Work

- New measures of information loss based on Information Theoretic concepts
- Use of outlier detection algorithms for identifying candidates for tuple suppression
- Better clustering algorithms to partition the data

* Supported by NSF under Grant No. 0430274

PURDUE UNIVERSITY | PURDUE HOMELAND SECURITY INSTITUTE | CERIAS | REGENSTRIEF CENTER FOR HEALTHCARE ENGINEERING | Discovery Park e-Enterprise Center