[PAN]Técnicas de anonimidad (Resumen)

La anonimización de los datos puede ser considerada un mecanismo para sanear la información, de forma que la privacidad de los sijetos referenciados en estos pueda ser garantizada por lo tanto:

- La información personalmente identificable debe ser tratada para prevenir su filtración
- Se minimiza el riesgo de filtrado de información cuando se muestran datos al público general, permitiendo el análisis de datos.
- Muchas regulaciones requieren que el uso de este tipo de mecanismos mantengan la información de ciudadanos a salvo.

Atributos personalmente identificables que deben anonimizarse:

- Identificadores (Atributos que identifican de manera única al individuo): DNI, carnet de conducir, fotos, etc…
- Pseudo-Indentificadores: Atributos que combinados pueden identificar a un individuo.

Atributos sensibles vs no sensibles:

- Cualquier atributo que pueda ser enlazado al individuo debe ser considerado sensible, aunque depende mucho del contexto.
- Cualquier atributo que no es relevante para el contexto se puede considerar no sensible.

Prácticas típicas:

- **Data Masking:** Se ocultan o eliminan valores del dataset de forma que los valores originales no pueden ser recuperados. Estas modificaciones pueden ser realizadas mediante cifrado, mezclado, diccionarios de sustitución o reemplazo de caracteres. Puede ser estático, lo que requiere monitorización de la base de datos y su enmascarado completo o dinámico, cuando los datos se enmascaran cuando se realizan las consultas.
- **Pseudoanonimización**: Remplaza identificadores personales con pseudónimos o identificadores falsos. Normalmente se mantiene un enlace interno entre los datos originales y los que se muestran, por lo que pueden ser recuperados revirtiendo los datos usando la información apropiada (Que debe ser altamente protegida).
- **Generalización**: Se reemplazan valores específicos de rangos amplios o categorías manteniendo los datos relativamente utilizables. Suele requerirse una cantidad de datos muy grande para asegurarse de que los grupos sean los suficientemente ambiguos sin perder utilidad.
- **Data Swapping**: Se permutan los datos o se mezclan los valores de de una fila dentro de una misma columna
- **Data perturbation**: Añade ruido a los datos y realiza redondeo, tratando de mantener los datos utilizables para su análisis.
- **Datos sintéticos**: En vez de publicar datos reales o anonimizados, se crea un dataset sintáctico basado en los datos originales. Se suelen usar técnicas de machine learnign para generar dichos datos sintéticos mediante el uso de modelos generativos.

K-Anonimidad

Un dataset es K-Anonimo cuando hay al menos k registros diferentes que comparten los mismo quasi-

identificadores:

- Para cualquier registro dado hay al menos otros k-1 registros que comparten los mismos atributos que podrían ser usados para identificar cualquiera de ellos como único.
- El valor K es normalmente empleado para calcular la privacidad, cuando más grande es, más difícil es desanonimizar los datos. La utilidad de los datos suele disminuir cuando más alto sea el valor de k ya que los datos se vuelven demasiado genéricos.

A tener en cuenta:

- Los Cuasi-identificadores y atributos sensibles deben ser distinguidos de forma apropiada para que no pueda revelar información de un atributo ya anonimizado.
- Es crucial que la información sensible de un grupo sea diversificada. Si un grupo contiene solo un registro, entonces puede ser trivial solo identificar a un individuo. Si todos los registros dentro de un grupo tienen el mismo valor para los atributos sensibles, entonces todos esos individuos pueden ser identificados.
- La dimensionalidad de los datos tiene un rol importante, cuando los datos están demasiado desperdigados, agruparlos para alcanzar k-anonimidad significa que se debe añadir mucho rudio. Cuando los datos están demasiado juntos, agruparlos puede no mejorar la privacidad

L-Diversidad

Establece que cada uno de los grupos K-anonimos debe tener al menos L registros sensibles que los distinga, de forma que así será más robusto contra filtraciones. Cuando mayor sea el valor de L, mayor será la dificultad para inferir información de los registros en cada uno de los grupos. Puede distorsionar la verdadera distribución de los datos.

Problemas:

- Si el valor de L es demasiado pequeño, puede filtrar datos importantes sobre los registros.
- Es vulnerable a ataques de asimetría debido a su distribución desbalanceada

T-Cercanía

Busca mantener la distribución de los valores sensibles de cada grupo lo más cerca posible a la distribución original:

- La distancia entre las distribuciones debe ser menor o igual a T.
- Para calcular la distancia entre dos distribuciones se suele usar EMC (Earth Movers Distance)

Algoritmo de Mondrian

Es uno de los métodos más populares para implementar la K-Anonimidad. La idea principal es la de realizar una partición multidimensional de los cuasi identificadores para generar varias regiones. Se realiza una grabación por cada región de forma que los cuasi-identificadores son anonimizados a través de ciertas estadísticas resumen:

- Los valores numéricos suelen ser codificados usando rangos mínimos y máximos
- Para atributos categóricos se suele definir un set represente todos esos elementos

Privacidad de geolocalización

From:
http://www.knoppia.net/ - **Knoppia**

Permanent link:
http://www.knoppia.net/doku.php?id=pan:res_tecnicas_anonimidad&rev=1736290476

Last update: **2025/01/07 22:54**