Gefördert durch:

Bundesministerium
für Wirtschaft
und Energie

aufgrund eines Beschlusses
des Deutschen Bundestages

Smart Data

Leitfaden: Anonymisierungstechniken

Technische Herausforderungen

Im Kontext von Big bzw. Smart Data wird eine klare Abgrenzung zwischen Daten mit Personenbezug und reinen Sachdaten zunehmend schwieriger. So können selbst bereits anonymisierte Datensätze durch die Kombination mit zusätzlichen Informationen aus Quellen Dritter eine Re-Identifizierung vormals anonymisierter Personen möglich machen. Ebenso kann die Anreicherung „unproblematisch" eingestufter Datensätze mit zusätzlichem Kontextwissen dazu führen, dass ein Personenbezug herstellbar wird. Diese Problematik wird insbesondere in offenen Datenverarbeitungsstrukturen begünstigt, in denen die Weitergabe von Daten an Dritte erfolgt.

Es ist daher notwendig, das Maß an Hintergrundwissen von Dritten a priori festzumachen. Sind Dritten zusätzlich andere anonymisierte Datensätze bekannt, welche eine oder mehrere Personen beinhalten, die auch im vorgelegten anonymisierten Datensatz enthalten sind, kann dies zu einer De-Anonymisierung führen: Fehlende Informationen können aus Daten verschiedener Quellen geschlussfolgert werden. Auf diese Situation hat man selten Einfluss, daher sind absolute Gegenmaßnahmen nicht möglich. Ein weiteres Problem kann auftreten, wenn die technische Umsetzung der Anonymisierung bekannt ist, beispielsweise über den Quell-Code der Anonymisierung. Durch die Hinzunahme des Faktors Zufall in den Anonymisierungsprozess reicht das Wissen über den Quell-Code für eine De-Anonymisierung nicht mehr aus. Das Ergebnis eines Anonymisierungsprozesses ist dann nicht mehr eindeutig, sondern teilweise durch einen Zufallsprozess gesteuert.

Interaktive und nicht-interaktive Anonymisierungsverfahren

Bei nicht-interaktiven Verfahren anonymisiert die Herausgeberin bzw. der Herausgeber die Datenbank und veröffentlicht sie anschließend. Interessenten greifen einmalig auf die Datenbank zu und können sie dann auswerten. Für die Herausgeberin bzw. den Herausgeber selbst besteht kein weiterer Handlungsbedarf. Bei interaktiven Verfahren hingehen stellt die Herausgeberin bzw. der Herausgeber den Empfangenden eine Schnittstelle zur Verfügung, über die sie Anfragen an die Datenbank schicken können. Die Ergebnisse der Anfragen werden durch Verrauschen verändert, ehe sie an die Empfangenden zurückgesendet werden. Dabei muss darauf geachtet werden, dass die Anzahl der Anfragen, die von einer empfangenden Person gestellt werden dürfen, limitiert ist.

Zusätzlich besteht die Möglichkeit, aus einem interaktiven Verfahren ein nicht-interaktives Verfahren zu erzeugen: Dazu formuliert die Herausgeberin bzw. der Herausgeber selbst Anfragen und führt diese aus. Anschließend werden die Ergebnisse veröffentlicht, wobei diese wie vorgesehen verrauscht werden. In der Regel schützt ein interaktives Verfahren die Privatsphäre besser als ein nicht-interaktives Verfahren. Dies ist jedoch auch aufwändiger in der Konstruktion und im Betrieb. Daher sollte gründlich abgewogen werden, wie schützenswert die zu veröffentlichenden Daten sind, und ob eine interaktive Lösung notwendig ist. Im Folgenden werden diesbezüglich einige Begriffe und Methoden der beiden Verfahren vorgestellt.

Anonymitätsbegriffe

Die Anonymitätsbegriffe K-Anonymität, L-Vielfalt und T-Geschlossenheit beziehen sich auf das Ergebnis eines nicht-interaktiven Anonymisierungsprozesses, Differential Privacy ist ein Anonymitätsbegriff für interaktive Anonymisierungsprozesse.

Gefördert durch:

Bundesministerium
für Wirtschaft
und Energie

aufgrund eines Beschlusses
des Deutschen Bundestages

Smart Data

K-Anonymität (k-anonymity)

K-Anonymität soll verhindern, dass Attribute einem einzelnen Datenbankeintrag eindeutig zugeordnet werden können. In obenstehender Datenbank kann man einem bekannten Geburtsdatum eindeutig einen Datensatz zuordnen. Um dies zu verhindern, werden die in der Datenbank gespeicherten Einträge in Gruppen mit gleichen Inhalten und mindestens Größe k zusammengefasst. Die vorhandenen Informationen werden dazu entsprechend verallgemeinert. Beispielsweise können Geburtsdaten auf Geburtsjahr oder eine Adresse auf den Wohnort reduziert werden, um einen größeren Personenkreis abzudecken.

Geburtsdatum	Krankenkasse	PLZ	Diagnose (sensibel)
08.01.1953	ABC	76131	Schlaganfall
13.01.1953	ABC	76135	Herzinfarkt
21.02.1949	ABC	76131	Herzinfarkt
03.03.1949	ABC	76149	Herzinfarkt

Tabelle 1: Datenbank ohne K-Anonymität

L-Vielfalt (l-diversity)

Eine Schwäche von K-Anonymität ist, dass nicht ausgeschlossen werden kann, dass alle sensiblen Daten einer Gruppe denselben Wert annehmen. Die nebenstehenden Datenbankeinträge sind 2-anonym.
Ist einem Dritten bekannt, dass eine Person in der Datenbank den Jahrgang 1949 hat, so weiß er auch sicher, dass diese Person einen Herzinfarkt hatte. L-Vielfalt erweitert k-Anonymität dahingehend, dass zusätzlich in jeder der Gruppe mindestens l verschiedene Werte angenommen werden.

Geburtsdatum	Krankenkasse	PLZ	Diagnose (sensibel)
1953	ABC	7613*	Schlaganfall
1953	ABC	7613*	Herzinfarkt
1949	ABC	761*	Herzinfarkt
1949	ABC	761*	Herzinfarkt

Tabelle 2: Datenbank mit 2-Anonymität ohne L-Vielfalt

T-Nachbarschaft (t-closeness)

L-Vielfalt garantiert zwar eine gewisse Unsicherheit bezüglich sensibler Werte. Kommt einer der l verschiedenen Werte in einer Gruppe jedoch signifikant häufiger vor als in der Verteilung der gesamten Datenbank, sind so wiederum Rückschlüsse möglich. T-Nachbarschaft fordert deshalb, dass der statistische Abstand zwischen der Verteilung innerhalb einer beliebigen Gruppe und der Verteilung auf der gesamten Datenbank maximal t ist.

Differential Privacy

Bei einem Differential-Privacy-konformen, interaktiven Verfahren lernt ein Angreifer nur geringfügig mehr über eine Person, die in der Datenbank enthalten ist, als über sie erfahrbar wäre, wenn sie nicht in der Datenbank enthalten wäre. Diese geringfügige Informationspreisgabe wird über einen Parameter $\varepsilon \geq 0$ festgelegt. Gleichzeitig ist ε ein Kompromiss: Je kleiner ε gewählt wird, desto weniger Informationen werden preisgegeben, aber desto schwieriger ist es auch, die Daten anschließend sinnvoll zu analysieren. Der Extremfall $\varepsilon = 0$ steht für perfekte Geheimhaltung, jedoch müssen die Daten dazu so sehr verrauscht werden, dass sie absolut unbrauchbar sind. Um Differential Privacy korrekt einsetzen zu können, müssen drei Punkte beachtet werden:
(1) Die Wahl von ε sollte gut überlegt sein, auch wenn diese Entscheidung in der Regel schwerfällt.
(2) Alle Einträge in der Datenbank müssen unabhängig voneinander sein.
(3) Bei der Implementierung dürfen keine Seitenkanalangriffe ermöglicht werden.

Weiterführende Links:

- http://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2014/wp216_en.pdf
- http://ceur-ws.org/Vol-1917/paper22.pdf
- https://www.infoq.com/articles/differential-privacy-intro

Ansprechpartner im Smart-Data-Programm

Prof. Dr. Jörn Müller-Quade und Dr. Dirk Achenbach

Begleitforschung Smart Data
www.smart-data-programm.de

c/o FZI Forschungszentrum Informatik
Außenstelle Berlin
Friedrichstr. 60, 10117 Berlin
Mail: kontakt@smart-data-programm.de