Contribution ID: **36** Type: **Presentation**

Privacy-preserving data integration for Big Data

Wednesday, October 11, 2017 11:30 AM (45 minutes)

Eine der großen Herausforderungen im Umgang mit Big Data liegt in der Auswertung (personenbezogener) Daten bei gleichzeitiger Wahrung des Datenschutzes und der Datensouveränität. Im ScaDS Dresden/Leipzig und am Institut für Informatik der Universität Leipzig werden dazu skalierbare Verfahren entwickelt und anhand von in der Praxis auftretenden Problemstellungen evaluiert.

Ein wichtiger Fokus liegt auf Verfahren zum sog. Record Linkage oder Entity Matching. Dabei werden Entitäten (Personen, Produkte oder ähnliches) aus mehreren Quellen verknüpft, die dasselbe Realwelt-Objekt darstellen. Solche Prozesse basieren auf dem Vergleich von bestimmten Attributen (Quasi-Identifier) der zu verknüpfenden Datensätze, z.B. für das Matching von Publikationen werden Titel, Autorennamen und akademische Zugehörigkeit der jeweilige Publikationen miteinander verglichen.
Personen-bezogenen Daten unterliegen in Deutschland und Europa jedoch strengen Datenschutzbestimmungen und dürfen gar nicht oder nur in verschlüsselter Form an eine dritte Partei weitergegeben werden. Daher ist das Matching von personen-bezogenen Daten problematisch –d.h. wenn sensitive Personendaten aus verschiedenen Organsation analysiert werden, muss garantiert werden, dass die Identität (Privacy) der zugehörigen Personen geschützt ist.

Im Vortrag werden Technologien und Methoden zum Privacy Preserving Record Linkage (PPRL) vorgestellt, die beim Linking von sensitiven Daten alle wesentlichen Anforderungen (Qualität, Skalierbarkeit und Privacy) erfüllen. Eine PPRL-Methode muss dabei die folgenden Fragen beantworten: (1) Wie kann man sensitive Personendaten so anonymisieren, dass ein Rückschluss auf die ursprüngliche Daten unmöglich wird? (2) Wie kann man Daten anonymisieren und trotzdem ihre Ähnlichkeit für das Matching beibehalten? (3)Wie kann man große anonymisierte Datenmengen effizient vergleichen und dies ohne Qualitätsverlust?
Die vorgestellten Ansätze werden anhand praxisnaher Anwendungsfälle im Bereich Gesundheit beschrieben z.B. das Verlinken von Patientendaten aus verschiedenen Krankenhäusern zur Identifikation von Zusammenhängen zwischen Krankheiten. Ein weiterer Anwendungsfall widmet sich Zensusdaten und den dabei auftretenden Datenschutzproblemen.

Track

BDAHM

Authors: Prof. RAHM, Erhard (Universität Leipzig); Dr PEUKERT, Eric (Universität Leipzig)

Presenters: Prof. RAHM, Erhard (Universität Leipzig); Dr PEUKERT, Eric (Universität Leipzig)