Samenvatting

Het koppelen van informatiebronnen wordt in de huidige maatschappij steeds belangrijker. Door koppelen ontstaan nieuwe inzichten, omdat meer gegevens van gemeenschappelijke objecten met elkaar in verband kunnen worden gebracht. Dit onderzoek richt zich op het koppelen van bronnen op microniveau. Hierbij worden entiteiten, die naar hetzelfde object verwijzen, aan elkaar gekoppeld: *entiteitreconciliatie* (bijvoorbeeld persoonsentiteiten die naar één persoon verwijzen). Verschillende bronnen hebben vaak geen gemeenschappelijke identificatie, waardoor deze manier van koppelen afvalt. Bronnen die interessant zijn om te koppelen, bevatten vaak weinig gemeenschappelijke informatie. Vanwege de beperkte overlap is de winst van het koppelen het grootst; er kunnen meer nieuwe gegevens met elkaar in verband worden gebracht. Overlap is echter, zonder gemeenschappelijke identificatie, wel de enige troef in de poging om te koppelen.

Om ondanks beperkte overlap toch entiteiten te kunnen reconciliëren, is een theorie ontwikkeld om alle aanwezige overlap van twee bronnen te gebruiken. Overlap bestaat uit eigenschappen die beide bronnen gemeen hebben. Als een gemeenschappelijke eigenschap overeenkomt, dan is er sprake van gelijkenis (Eng. *similarity*). De mate van gelijkenis wordt bepaald door de onderlinge positionering van twee attributen die de eigenschap beschrijven. Met behulp van expertkennis wordt deze positie via een positieverdeling (een trendlijn over het histogram van de verwachte onderlinge positionering van de eigenschap) omgezet in een gelijkeniswaarde. De attributen, die een gemeenschappelijke eigenschap beschrijven, worden geplaatst onder een gemeenschappelijk entiteittype (knoop genoemd). Elke knoop draagt bij aan de beschrijving van de centrumknoop waarin de reconciliatie gewenst is. Zodoende wordt de entiteitgelijkenis per knoop bepaald en wordt ook de objectgelijkenis bepaald, waarin tevens de gelijkenis in andere knopen wordt meegenomen. Hierbij wordt de gelijkenis effectief gedistribueerd naar de centrumknoop. Door de knopen te berekenen in een hiërarchische structuur ontstaat clustering, waardoor het aantal vergelijkingen wordt verlaagd. Voor de entiteitreconciliatie is een methode bedacht, waarmee entiteiten van één knoop efficiënt worden gereconcilieerd.

Om de theorie te toetsen is een prototype (EROS, '*Entity Reconciliation using Object Similarity*') ontwikkeld, waarin een casus is geïmplementeerd. Van deze casus zijn de correcte reconciliaties bekend; deze zijn gebruikt in de analyse van de resultaten. Er is persoonsreconciliatie toegepast op 10.000 personen in de ene bron tegen 8.705 personen in de andere bron. Hierbij zijn 5 gemeenschappelijke eigenschappen gebruikt, waaronder 3 persoonseigenschappen (geboorteland, geslacht en geboortedatum). Voor 5% is de correcte reconciliatie niet gevonden als gevolg van te weinig overlap. Als de correcte reconciliatie is gevonden, dan wordt deze in 98% van de gevallen ook daadwerkelijk gekozen.

Uit dit onderzoek blijkt dat, ondanks beperkte overlap, reconciliatie op microniveau door middel van objectgelijkenis goed mogelijk is. Uiteraard moet de aanwezige overlap discriminerend genoeg zijn. In dit kader moet worden opgemerkt dat door de kleine set van gegevens de persoonseigenschappen voor

sommige combinaties al sterk discriminerend zijn. Meer onderzoek is nodig om te bepalen wanneer overlap voldoende discriminerend is, met name voor grotere datasets waarin de correcte reconciliaties onbekend zijn. De theorie biedt een uitgangspunt voor meer onderzoek in de richting van data mining en privacygerelateerde toepassingen.