On semi-automated matching and integration of database schemas

Ünal Karakaş, Ö.

Publication date
2010

Samenvatting

Semiautomatische Vergelijking en Integratie van Database Schema's[1]

Tegenwoordig zien steeds meer organisaties het belang in van samenwerking om hun gezamenlijke doelstellingen beter te kunnen realiseren. Als een gevolg van deze sterker wordende impuls tot samenwerken zijn er onderzoeks- en ontwikkelingsgebieden ontstaan die zich richten op de studie naar het delen van gegevens en de coöperatie tussen organisaties, en het ontwikkelen van ICT gereedschap en systemen ter ondersteuning hiervan. Er bestaan echter nog vele open uitdagingen op dit gebied. Een focus op het delen en integreren van informative tussen onafhankelijke knooppunten binnen coöperatieve netwerken, en het faciliteren van transparante toegang tot de informatie in hun databases, vormen de basis voor een efficient samenwerking. Het is echter duidelijk dat onafhankelijke knooppunten hun eigen informative heterogeen modelleren in hun database schema's. Voordat het uitwisselen van enige informative tussen deze databases kan plaatsvinden zal daarom een aantal uitdagingen aangegaan moeten worden, waaronder: (i) *identificatie en vaststelling van overeenkomsten tussen concepten gedefinieerd in onafhankelijke database schema's*, (ii) *oplossing van bestaande heterogeniteiten tussen database schema's* van de betrokken knooppunten, en (iii) *integratie van deze database schema's*. Over het algemeen zijn dit gecompliceerde taken, zelfs om met de hand uit te voeren, en deze moeilijkheidsgraad neemt toe met het aantal knooppunten in het federatieve netwerk, alsook met de grootte van de database schema's. Een aantal onderzoeksrichtingen en -prototypes heeft zich dan ook bezig gehouden met het automatiseren van het vergelijken van schema's, terwijl anderen zich onafhankelijk hiervan gericht hebben op het automatiseren van het integreren van schema's. Ondanks deze eerdere pogingen in zowel academische als commerciële omgevingen, vraagt het vergelijken en integreren van schema's vandaag de dag nog steeds een grote hoeveelheid handwerk en zijn er vele open onderzoeksvragen op deze gebieden.

Dit proefschrift beschrijft een *geautomatiseerde maar onder supervisie opererende gecombineerde benadering waarbij de concepten van het vergelijken en integreren van relationele database schema's aangepakt en samengevoegd worden*. De belangrijkste

[1] Vertalling door Leo Breebaart

contributie van dit proefschrift is derhalve een nieuwe gecombineerde schemavergelijkings/integratie benadering. Een *proof of concept* voor deze aanpak wordt gegeven door de implementatie van een prototype systeem genaamd SASMINT – Semi Automatic Schema Matching and INTegration. Het SASMINT systeem identificeert automatisch een groot aantal syntactische, semantische en structurele heterogeniteiten tussen relationele database schema's. Vervolgens probeert het deze heterogeniteiten op te lossen door de gebruiker ter validatie een lijst van mogelijke correspondenties tussen de vergeleken systemen voor te leggen. Het systeem genereert dan automatisch uit deze lijst een geïntegreerd schema.

Hoofdstuk 1 van dit proefschrift presenteert de motivatie voor dit werk, de primaire onderzoeksvragen, de doelstellingen, en de belangrijkste contributies van dit onderzoek.

Hoofdstuk 2 gaat dieper in op de basisdefinities en de classificatie van concepten uit de huidige *state of the art*, in relatie tot architecturen, methodes, en systemen die het delen en uitwisselen van gedistribueerde en heterogene informatie mogelijk maken.

Daar heterogeniteit centraal staat in de schemavergelijkings en -integratie processen, zijn identificatie en oplossing van verschillende types heterogeniteit cruciaal voor het succes van deze processen. Hoofdstuk 3 presenteert verschillende taxonomieën voor heterogeniteiten, en reduceert deze vervolgens tot database schema heterogeniteiten, het hoofdonderwerp van dit proefschrift.

Na in deze eerste drie hoofdstukken de basis gelegd te hebben voor ons onderzoek, behandelt de rest van dit proefschrift het ontwerp en de uitwerking van onze voorgestelde benadering. Als centraal hoofdstuk van dit proefschrift bevat Hoofdstuk 4 een literatuurstudie die zich richt op (i) database-integratie en -interoperabiliteit, (ii) database schemavergelijking, (iii) database schema-integratie en (iv) ontologievergelijking en -samenvoeging. Vervolgens introduceren wij onze voorgestelde oplossing, SASMINT, om een aantal benoemde open problemen aan te pakken. De SASMINT benadering verhoogt de nauwkeurigheid van de schemavergelijking door het gebruik van een gewogen combinative van een aantal schemavergelijkingsalgoritmes, waar elk algoritme een specifieke categorie syntactische, semantische of structurele conflicten aanpakt. Een andere bijdrage van SASMINT die in Hoofdstuk 4 wordt behandeld is de introductie van onze zogenaamde SAMPLER techniek, die voor elk doeldomein semiautomatisch de juiste gewichten bepaalt voor de algoritmes die toegepast zullen worden in het linguïstische vergelijkingsproces. Hoofdstuk 4 introduceert tevens een overzicht van de verzameling regels die automatische generatie mogelijk maakt van zowel de geïntegreerde schema's als van de afgeleide constructies die representatief zijn voor de geschiedenis van het integratieproces in het coöperatieve netwerk. Ook onze ontwikkeling van de SASMINT Derivation Markup Language (SDML), waarmee het creëren van zowel persistente schemavergelijkingsresultaten als persistente schema-integratieresultaten beschreven en ondersteund wordt, wordt behandeld in dit hoofdstuk.

Net als bij ieder ander onderzoek is het belangrijk om de door ons voorgestelde benadering te verifiëren en te valideren. Om dit te bewerkstelligen hebben wij een implementatie van onze aanpak ontwikkeld in de vorm van het SASMINT systeem, waar in Hoofdstuk 5 het focus op gericht is. In dit hoofdstuk worden zowel de belangrijkste onderdelen als de werking van SASMINT beschreven. Alhoewel naar onze mening het ondersteunen van gebruikersinteractie door middel van een GUI fundamenteel is voor het proces van semiautomatische schemavergelijking en -integratie, is dit typisch een component die ontbreekt in het onderzoeks- en ontwikkelwerk op dit gebied. Een GUI is nodig om de gebruiker te ondersteunen in het verifiëren van zowel de automatisch geïdentificeerde schemaconflicten en -correspondenties als van het voorgestelde geïntegreerde schema. Via SASMINTs GUI kunnen

gebruikers met het systeem interacteren, de juiste procesweegfactoren aangeven, de automatisch gegenereerde resultaten goedkeuren/veranderen/afkeuren, en de resultaten bewaren van zowel schemavergelijking als schema-integratie.

Om de resultaten van dit onderzoek te demonstreren en te evalueren en om de kwaliteit van het SASMINT systeem te meten, hebben wij een aantal experimenten uitgevoerd. In het bijzonder is de schemavergelijkende component van SASMINT vergeleken met, en geëvalueerd tegen een ander state of the art vergelijkingssysteem. De benadering van SASMINT is echter uniek in het feit dat het een fusie is van zowel schemavergelijkings- als schema-integratieprocessen. Het was daarom niet mogelijk een tegenhangen te vinden waarmee de schema-integratiecomponent van SASMINT vergeleken kon worden. In onze evaluatie-experimenten hebben we derhalve alleen maar gemeten hoe succesvol er correcte resultaten geproduceerd worden. Deze experimenten worden uitvoerig behandeld in Hoofdstuk 6 van dit proefschrift.

Dit proefschrift eindigt met een uitleg over hoe de primaire onderzoeksvragen zijn beantwoord.