Semantic Support for Quantitative Research

Rijgersberg, H.

2013

document version
Publisher's PDF, also known as Version of record

Samenvatting

De titel van dit proefschrift luidt: “Semantische ondersteuning voor kwantitatief onderzoek.” We definiëren kwantitatief onderzoek als de wetenschappelijke bestudering van fenomenen en hun eigenschappen en relaties met gebruikmaking van kwantitatieve concepten zoals getallen, meetschalen, eenheden, mathematische operaties, tabellen, grafieken, etc. Semantische ondersteuning impliceert het ondersteunen van wetenschappers door middel van acties die gedaan kunnen worden op basis van formele contextuele betekenis die is toegekend aan de kwantitatieve data en modellen. In dit proefschrift laten we zien hoe het formeel beschrijven van data en modellen en hun ontstaan – in het bijzonder door middel van computationele methoden – hergebruik en reproductie van wetenschappelijke resultaten kan bevorderen. Dit past in een visie over het verbeteren van wetenschappelijke samenwerking en kwaliteit en de academische uitdaging om computersemantiek te ontwikkelen, te evalueren, en toe te passen om data te verrijken.

Formele representaties kunnen gebaseerd worden op vocabulaires, in het bijzonder *ontologieën*. Ontologieën zijn systemen van concepten en relaties tussen deze concepten. Ontologieën vervullen een centrale rol in wat het Semantisch Web wordt genoemd, het Internet gebouwd op (geformaliseerde) betekenis. Het Internet speelt hier de rol van medium voor het communiceren van het vocabulaire en data uitgedrukt in het vocabulaire, een belangrijke technische conditie voor het werkelijk *delen* van vocabulaire en data.

In dit proefschrift onderzoeken we hoe we kwantitatief onderzoek kunnen ondersteunen met behulp van ontologieën. Daarom construeren we een ontologie van kwantitatief onderzoek (OQR), laten zien hoe de ontologie gebruikt kan worden om kwantitatieve kennis en zijn verkrijging uit te drukken, passen we de ontologie toe in computerapplicaties en evalueren deze met gebruikers. We construeren de ontologie stapsgewijs en baseren het op algemeen aanvaarde principes van de wetenschapsfilosofie en officiële standaarden voor grootheden en eenheden. We passen de voorgestelde ontologie toe in een onderzoekscase uit het voedseldomein. Het blijkt dat de argumentaties, metingen en geanalyseerde resultaten die verkregen zijn in deze case op adequate wijze kunnen worden uitgedrukt door het voorgestelde vocabulaire. Vervolgens passen we het model voor deze case toe in een prototypecomputersysteem en evalueren het met gebruikers, op deze wijze de bruikbaarheid van het model in de praktijk aantonend.

Om een vocabulaire voor kwantitatief onderzoek te creëren hebben we eerst enig begrip nodig van de fundamentele mechanismen van wetenschappelijk onderzoek, naast een model van de onderzoeksworkflow. Deze workflow bevat stappen zoals “ontwerp experiment”, “voer meting uit”, en “analyseer data”. We zetten een stap in de richting van het construeren van een (initiële) epistemologische ontologie, gebaseerd op modellen van bekende wetenschapsfilosofen zoals Karl Popper en Mario Bunge. De ontologie kan gebruikt worden om acties op basis waarvan wetenschappelijke kennis wordt verkregen uit te drukken (zoals het uitvoeren van een meting of het stellen van een hypothese) en deze te relateren aan de data. Dit stelt onderzoekers in staat de herkomst van hun data vast te leggen en anderen om hun werk te traceren en reproduceren. Een belangrijke conclusie van ons werk is om concepten zoals “hypothese’, “theory”, etc. als eigenschappen van acties in de wetenschappelijke workflow te definiëren in plaats van als onafhankelijke concepten. Dit is belangrijk omdat wetenschappelijke statements altijd binnen de scope van een specifieke wetenschappelijke redenatie of studie worden gesteld. Iets dat een geaccepteerde theorie is in de ene wetenschappelijke school kan een (vooralsnog ongedragen) hypothese zijn in de andere.

Een belangrijk deel van OQR is de Ontologie van Eenheden en gerelateerde concepten (OM). Om te bepalen welke concepten en relaties dit domein representeren hebben we een semiformele beschrijving van het domein opgesteld op basis van tekstuele beschrijvingen van standaarden in het veld. Vervolgens hebben we bestaande ontologieën van eenheden vergeleken met deze beschrijving, wat duidelijk maakte dat de bestaande ontologieën slechts subsets van de vereiste concepten en relaties definiëren. Daarom stellen we een nieuwe ontologie voor, OM. Deze ontologie is gebaseerd op de semiformele beschrijving van tekstuele standaarden en definieert daarom de meest veelomvattende set van relevante concepten in het domein. OM breidt de overeenkomstige delen van de geanalyseerde ontologieën uit. Daardoor kan de ontologie een grotere verscheidenheid aan competentievragen beantwoorden dan de bestaande aanpakken. Het aanhouden van een tussenfase in de vorm van een semiformele beschrijving van het domein is een levensvatbare benadering omdat de fasen van het samensmelten van de verschillende standaarden en het opstellen van het uiteindelijke formele vocabulaire onderscheiden zijn en transparant gemaakt. OM is ook vergeleken met QUDT, een ander actueel OWL-model in het domein van grootheden en eenheden. Het vergelijk is gebaseerd op use cases uit onze eigen projecten en algemene ervaring in het veld. Het samensmelten van QUDT en OM is een aanbeveling voor de toekomst.

De tweede kwestie die we aanpakken is hoe dataverwerkingsstappen te representeren en hoe geaggregeerde data die traditioneel in (wetenschappelijke) tabellen staan weer te geven. We definiëren computationele methoden die geïnstantieerd kunnen worden en verbonden met input- en outputdata en -modellen.

Generieke methoden worden onderscheiden van hun implementaties in externe softwarepakketten, zoals Matlab, R, en SPSS. Deze methoden (generiek en implementatie-) zijn aan elkaar gerelateerd; de gebruiker kan beslissen welk externe pakket zijn berekening zal uitvoeren. Interfacing tussen deze methoden gebeurt op basis van eigenschappen die variabelen representeren. Deze variabelen (eigenschappen van deze methoden) komen als onafhankelijke concepten voor in vertalingsregels van de generieke methode naar een implementatie van de methode in een extern pakket. Mechanismes voor het strippen en verrijken van kwantitatieve informatie, vereist om tussen het conceptuele en het numerieke perspectief te migreren, worden geëxploreerd. De modelleerstappen worden genomen door het verder analyseren van de onderzoekscase uit het food-engineering-domein. Een van de belangrijkste voordelen van het modelleren van wetenschappelijke tabellen is dat de informatie die zich in headers en cellen bevindt netjes geïdentificeerd en met elkaar verbonden is. Dit opent de poort voor het vinden van gerelateerde kwantitatieve gegevens uit verschillende bronnen. De data kan worden geselecteerd, gecombineerd (geïntegreerd), en indien nodig automatisch geconverteerd. Het toevoegen van semantiek aan headers en cellen gaat verder dan huidige databases en spreadsheets die alleen elementaire datatypes bevatten. Een uitdaging is het ontwikkelen van geautomatiseerde methoden voor het converteren van bestaande computationele methoden en tabulaire data naar OQR. Een stap in de richting van het laatste (tabulaire data) wordt gemaakt in dit proefschrift, zie verder.

Na het definiëren van het vereiste vocabulaire onderzoeken we welke tools kunnen worden ontwikkeld om het kwantitatieve onderzoeksproces te ondersteunen. Teneinde OM beschikbaar te maken voor willekeurige softwaresystemen voorzien we in een groot aantal web services die een gestandaardiseerde interface bieden. Drie applicaties demonstreren de bruikbaarheid van OM en zijn services. Ten eerste checkt een webapplicatie dimensie- en eenheidconsistentie van formules. Ten tweede berekent een engineering-applicatie voor agriculturele distributieketens productrespiratiegrootheden en -maten. Ten derde assisteert een Microsoft Excel add-in in data-annotatie en eenheidconversie, en een extensie in dataintegratie. Gebruikersevaluaties geven aan dat OM en de aan OM gerelateerde services een bruikbare component voor softwareapplicaties in de wetenschap en engineering bieden. We laten zien hoe OQR kan worden toegepast in Quest, een computertool die we ontwikkelen voor het verbinden van data en modellen aan computationele methoden, en het uitbesteden van berekeningen aan externe software. OQR/Quest ondersteunen geautomatiseerde reproductie van berekende resultaten, wat we hebben getest met gebruikers. Onze testpersonen achtten Quest van grote importantie en gemak. In huidige computerondersteuning belemmeren de vele handmatige acties zoals het linken van inputgegevens aan computationele methoden, het in de juiste format gieten van deze gegevens en na evaluatie het interpreteren van de numerieke waarden (het toekennen van betekenis)

het experimenteren met berekeningen. Als dit automatisch gebeurt wordt de onderzoeker in staat gesteld en zelfs aangemoedigd om te proberen te experimenteren met verschillende methoden "on the fly". Verwacht wordt dat dit onderzoek een boost zal geven. OQR/Quest stellen in staat om computationele (numerieke) methoden automatisch aan te roepen vanaf een conceptueel niveau. Deze benadering vult het gat tussen de mens die textuele informatie interpreteert en de computer die de onderliggende data en modellen verwerkt. Computationele software kan deze methoden uitvoeren waarbij de vereiste input- en outputgegevens automatisch gelinkt worden. Op dit moment bevat OQR een beperkt aantal computationele methoden teneinde het principe te illustreren. Toekomstig onderzoek moet uitwijzen welke kant de ontwikkeling van tools op moet gaan, in technische zin dan wel leidend tot nieuwe onderzoeksvragen.

Tenslotte bestuderen we hoe relatief ongestructureerde "legacy data" opgeslagen in tabellen geconverteerd en geannoteerd kan worden tot een semantische representatie in RDF(S). We introduceren nieuwe disamiguatiestrategieën gebaseerd op OM, die assisteren in het verbeteren van de kwaliteit van de annotaties zoals nog niet door bestaande systemen bereikt. We laten verschillende manieren zien hoe OM kan helpen in het oplossen van amiguïteitsproblemen gebaseerd op detectie van samengestelde eenheden, dimensionele analyse, identificatie van toepassingsgebieden en identificatie van grootheid-eenheidkoppels. Een voorbeeld van zo'n heuristieke regel is "Symbolen die naar aan elkaar gerelateerde grootheden en eenheden refereren zijn waarschijnlijker dan ongerelateerde grootheden en eenheden." Bijvoorbeeld, "T (C)" refereert waarschijnlijker naar temperatuur en graad Celsius dan naar tijd en coulomb. Echter, de performance is nog niet perfect. Meer heuristische regels moeten worden geformuleerd en, bijvoorbeeld, meer toepassingsgebieden moeten opgesteld worden om kennis aan te kunnen bieden over grootheden en eenheden zoals ze voorkomen in de praktijk.

We kunnen concluderen dat de relevantie van het ontwikkelen en gebruiken van ontologieën in de wetenschap en engineering bevestigd is voor de beschouwde cases. We hebben laten zien dat het de moeite waard is deze weg te bewandelen bij het streven naar geavanceerde computerondersteuning van kwantitatief onderzoek. De wetenschappelijke gemeenschap is altijd een drijvende kracht geweest achter innovatie in communicatietechnologieën, waarbij het (Semantisch) Web een treffend voorbeeld is. Echter, nu pas krijgt het omgekeerde effect van het gebruiken van het web voor het uitvoeren van wetenschap aandacht in wat e-science wordt genoemd. Door een aantal ontwikkelingen verwachten we dat e-science de wetenschappelijke en engineering-praktijk in de nabije toekomst flink gaat veranderen. Ten eerste omdat wetenschappers migreren van vrije-tekstdocumenten naar gedigitaliseerde, gestructureerde informatie die door geautomatiseerde systemen kan worden verwerkt. Ten tweede omdat de interactie tussen

wetenschappers veel intensiever is geworden, waarbij disciplinaire grenzen overschreden worden, in een vroeg stadium van het onderzoek. Dit zal de dynamiek van wetenschappelijk onderzoek significant veranderen. Het zal een uitdaging voor e-science zijn om andere hindernissen zoals politieke, sociologische en juridische te overwinnen. Dit proefschrift beoogt te laten zien dat vocabulaires het wetenschappelijke proces in technische zin kunnen ondersteunen.

We staan slechts aan het begin van het ontwerpen, implementeren en gebruiken van wetenschappelijke ontologieën in e-science. Als meer ontwikkelaars beseffen wat het nut is van collectief en onafhankelijk vocabulaire en het gebruik daarvan in onderzoeksondersteunende systemen voorspellen we een grote toename in geavanceerde ondersteuning van onderzoeksprocessen.