End-user support for access to heterogeneous linked data

Hildebrand, M.

Samenvatting

Ondersteuning van eindgebruikers in het ontsluiten van heterogene gelinkte data

Organisaties maken op het Web hun data en services beschikbaar voor hergebruik. Door dit openbaar beschikbaar maken kan informatie afkomstig van verschillende bronnen worden gecombineerd. Dit maakt nieuwe manieren mogelijk om de informatie te ontsluiten, die de oorspronkelijke dataleveranciers niet noodzakelijk voorzien hadden. Webapplicaties die verschillende databronnen en services hergebruiken zijn bekend als 'Mash-ups', en worden een gangbare oplossing om specifieke gebruikerstaken te ondersteunen. Het ontwerpen van dit soort applicaties vereist echter een ontwikkelaar die moet aangeven hoe de data van verschillende bronnen moeten worden gebruikt en geïntegreerd.

Technologieën voor het Semantische Web beloven het hergebruik van verschillende databronnen makkelijker te maken. De representatietalen voor het Semantische Web bieden een gestandaardiseerde manier om kennis te modelleren en te delen. Bovendien kan met deze talen de betekenis van verschillende aspecten in de data expliciet gemaakt worden, zodat machines automatisch kunnen bepalen hoe de data moeten worden gebruikt en geïntegreerd.

Steeds meer databronnen worden op het Semantische Web gepubliceerd en deze databronnen worden onderling met elkaar verbonden. Het resultaat is een grote graaf met gelinkte data, die een grote verscheidenheid aan objecten beschrijft. In een eerste studie analyseren wij hoe bestaande applicaties gebruikers toegang verschaffen tot deze heterogenene gelinkte data. We concluderen uit deze studie dat applicaties voor het Semantische Web een ruime verscheidenheid aan taken ondersteunen en toegang verschaffen tot verschillende soorten databronnen. Het blijft echter onduidelijk hoe goed de gebruikte semantische technologieën eindgebruikers ondersteunen, omdat algemeen geaccepteerde methoden om deze applicaties te evalueren ontbreken. Gegeven een specifiek domein en taak is het dus moeilijk

om te bepalen hoe de semantiek in de data moet worden gebruikt om de beste ondersteuning te bieden aan eindgebruikers.

In dit proefschrift is ons culturele erfgoed gekozen als het applicatiedomein om de toegang tot heterogenene en gelinkte data te bestuderen. Het semantisch-rijke cultuurweb, dat is geconstrueerd in het MultimediaN E-Culture project, wordt gebruikt als experimentele data. De praktische resultaten van dit onderzoek zijn beschikbaar als applicaties en web services in Cliopatria, de open source software architectuur van dit project.

Dit proefschrift zet een eerste stap om voor een specifiek domein en aantal taken de eisen te formuleren om eindgebruikers toegang te verlenen tot semantisch-rijke en heterogene gelinkte data. Verschillende aspecten van het zoekprobleem worden geëxploreerd in drie casussen: (i) annotatie van kunstwerken om te bestuderen hoe gebruikers efficiënt naar termen kunnen zoeken in meerdere vocabulaires, (ii) facet gebaseerd navigeren om het formuleren van gestructureerde zoekvragen in meerdere collecties van geannoteerde kunstwerken te bestuderen, en (iii) het semantisch zoeken naar kunstwerken om te bestuderen hoe de relaties in meerdere gelinkte vocabulaires kunnen worden gebruikt om objecten te vinden die semantisch gerelateerd zijn aan een query.

Annotatie In professionele annotatie is de taak van de gebruiker om vocabulaire termen te vinden om een kunstwerk te beschrijven. In de huidige systemen voor collectiemanagement geeft elk annotatieveld toegang tot de termen van één thesaurus. De dekking van de vocabulaires, die intern in musea worden gebruikt, zijn niet altijd toereikend. In het bijzonder voor de beschrijving van het onderwerp afgebeeld op een kunstwerk moeten de catalogiseerders daarom hun kostbare tijd ook besteden aan het toevoegen van nieuwe vocabulaire termen. Externe bronnen op het Web kunnen de dekking van beschikbare annotatietermen verhogen. Dit vereist ondersteuning van gebruikers met het zoeken in meerdere bronnen die verschillende karakteristieken en dataschema's kunnen hebben. Hoe kan tekstgebaseerd zoeken gebruikt worden om termen te vinden in meerdere vocabulaires?

In een studie met professionele catalogiseerders van het prentenkabinet van het Rijksmuseum Amsterdam onderzoeken we hoe meerdere vocabulaires in een annotatie tool kunnen worden geïntegreerd. De initiële eisen op de data, algoritmes en interface worden geformuleerd op basis van een analyse van het huidige annotatie proces. Door in meerdere iteraties verschillende prototypes te ontwikkelen worden de eisen bijgesteld en verschillende oplossingen uitgeprobeerd. De oplossingen in het laatste prototype worden kwalitatief geëvolueerd met de catalogiseerders. We stellen vast dat eindgebruikers effectief kunnen worden ondersteund met bestaande technologieën, zoals interface componenten voor 'autocompletion' en algoritmes voor het zoeken naar termen en het organiseren en presenteren van de geselecteerde termen. De gebruikersinterface en de algoritmes moeten echter wel zorgvuldig worden geconfigureerd voor verschillende annotatievelden en vocabulaires. We

identificeren de vereiste parameters van de algoritmes en de gebruikers-interface en demonstreren hoe deze kunnen worden geconfigureerd voor een specifieke taak en domein.

Facet gebaseerd navigeren Facet gebaseerd navigeren is een populaire strategie voor het formuleren van gestructureerde zoekvragen om (kunst) collecties te exploreren. Hoe kan het facet gebaseerde navigeren worden gebruikt om het formuleren van gestructureerde zoekvragen voor gelinkte data te ondersteunen? Traditionele zoekapplicaties die facet gebaseerd navigeren ondersteunen vereisen een homogene collectie met een vast dataschema. Heterogene databronnen op het Semantische Web kunnen verschillende soorten objecten bevatten, elk met hun eigen facetten. Bovendien kunnen gestructureerde zoekvragen voor gelinkte data indirecte restricties bevatten, die niet door traditioneel facet gebaseerd navigeren worden ondersteund.

Op basis van een 'use case' formuleren we de eisen voor facet gebaseerd navigeren op heterogene databronnen op het Semantische Web. Oplossingen voor de vereiste zoekfunctionaliteit en presentatie methodes worden geëxploreerd door het implementeren van een prototype. We stellen vast dat de vereiste functionaliteit voor kleine en middelgrote RDFS databronnen kan worden ondersteund met een volledige datagestuurde oplossing. De semantische relaties in de data kunnen worden gebruikt om de facetten te organiseren en daarmee de ondersteuning voor de eindgebruikers te verbeteren. Het formuleren van indirecte restricties wordt in het prototype ondersteund door de gebruiker meerdere soorten objecten te laten navigeren in dezelfde interface waarbij de restricties op objecten van één type kunnen worden gebruikt voor objecten van een semantisch gerelateerd type object. We moeten ook concluderen dat de relaties in de data voor specifieke taken niet altijd overeenkomen met het perspectief van de gebruiker. In dit geval moet worden bepaald welke facetten geschikt zijn voor eindgebruikers en deze moeten handmatig worden geconfigureerd.

Semantisch zoeken In veel professionele zoektaken willen gebruikers kunstwerken vinden die op verschillende manieren gerelateerd zijn aan een onderwerp. Om hun niet triviale informatiebehoeftes te bevredigen moeten domeinexperts vaak meerdere zoekvragen formuleren en handmatig de verschillende resultaten integreren in een coherente verzameling. Onze hypothese is dat eindgebruikers ondersteund kunnen worden in het zoekproces door op de juiste manier gebruik te maken van de annotaties van kunstwerken met termen uit gestructureerde en gelinkte vocabulaires en de relaties tussen deze termen. Om dit semantische zoeken mogelijk te maken moeten we eerst beter begrijpen welke en hoe verschillende soorten relaties in de semantisch-rijke databronnen gebruikt kunnen worden.

De ondersteuning van semantisch zoeken is bestudeerd in twee experimenten. Ten eerste is het nut van verschillende paden van relaties onderzocht in een ge-

bruikersstudie met een klein aantal domeinexperts. Een aantal typen paden is geïdentifieerd en hun nuttigheid voor de experts is kwalitatief geëvalueerd. In het tweede experiment is het gebruik van de verschillende typen paden in een interactieve zoekapplicatie onderzocht. Op basis van deze experimenten worden de implicaties voor het ontwerp van een interactieve semantische zoek applicatie voor cultureel erfgoed besproken. We observeerden dat het proces van zoeken naar kunstwerken uit verschillende fases bestaat. Verschillende soorten relaties in de data zijn nuttig in die fases en verschillende soorten interactie zijn nodig om de eindgebruikers te ondersteunen in de fases. De initiële zoekresultaten moeten, bijvoorbeeld, kunstwerken bevatten die gerelateerd zijn aan de zoekvraag op basis van tekstuele eigenschappen en met annotaties van gecontroleerde termen. Ook moeten de equivalentie relaties tussen de vocabulaire termen meegenomen worden. Vervolgens moet er interactie mogelijk zijn om de zoekvraag te disambigueren door vocabulaire termen te selecteren. Ook moet het mogelijk zijn om de zoekvraag te herformuleren door gebruik te maken van verschillende hiërarchische en associatieve relaties tussen termen. Om deze zoekstrategieën te ondersteunen zijn verschillende interface componenten nodig die gebruik maken van graaf zoek algoritmes die op verschillende manieren zijn geconfigureerd.

De casussen brengen verscheidene eisen naar boven op de software architectuur die nodig is voor de zoekfunctionaliteit en de presentatiemethoden. Om annotatie en semantisch zoeken te ondersteunen zijn configureerbare zoek algoritmes noodzakelijk. Bovendien moet de organisatie en de presentatie van de resultaten configureerbaar zijn. Er zijn verschillende interactieve oplossingen nodig om de gebruiker te ondersteunen met het uitproberen van verschillende zoekvragen en het exploreren van verschillende zoekstrategieën. Voor de presentatie van de zoekresultaten en de navigatiepaden zijn er abstracties in data nodig waarmee de grote verscheidenheid aan termen en relaties behapbaar gemaakt kan worden.

De functionaliteit voor tekst gebaseerd zoeken is geïmplementeerd met generieke algoritmes voor RDF data en kan geconfigureerd worden op een aantal dimensies. De algoritmes zijn beschikbaar in ClioPatria als geparameteriseerde web services. Deze services zijn uitgebreid met algoritmes om de resultaten te organiseren en presenteren. De vereiste interactie wordt ondersteund door bestaande Web interface componenten te hergebruiken en uit te breiden. Om deze componenten en de web services waar ze gebruik van maken te configureren hebben we een methode voorgesteld waarin de functionaliteit in een model wordt omschreven. Hierdoor wordt de configuratie een taak waarin de domeinspecifieke data gelinkt worden aan dit model. Dit kan uitgevoerd worden door een domeinexpert waardoor er geen ontwikkelaar meer nodig is.

In de geëxploreerde oplossingen worden vier soorten semantische relaties gebruikt om de gebruikersondersteuning te verbeteren. Ten eerste, de thesaurus specifieke relaties in de oorspronkelijk data, zoals gedefinieerd in SKOS, bieden nuttige

abstracties over vocabulaires waarop specifieke functionaliteit en presentatiemethoden kunnen worden gebaseerd. Ten tweede de equivalentierelaties tussen termen van verschillende vocabulaires maken het mogelijk om informatie van externe bronnen te integreren. Deze relaties maken het ook mogelijk om dubbele resultaten te verwijderen. Ten derde, de simpele relaties om dataschema's te linken maken het mogelijk om geïntegreerde toegang te verschaffen tot heterogene data, terwijl de rijkheid van de individuele collecties en vocabulaires behouden blijft. Deze relaties maken het ook mogelijk om de resultaten op verschillende abstractie niveaus te presenteren. Ten vierde, de ontologische beschrijvingen van eigenschappen verbeteren de geïntegreerde zoekfunctionaliteit in heterogene data.

In dit proefschrift hebben we laten zien dat een Web van culturele data gebruikt kan worden om domeinexperts met een aantal taken te ondersteunen. Om de resultaten toepasbaar te maken op andere domeinen en gebruikersgroepen moeten de relevante abstracties voor dat domein en de specifieke gebruikerseigenschappen geïndentificeerd worden. Het vinden van de juiste configuraties voor de zoekfunctionaliteit en de presentatiemethodes vereist verder onderzoek. We verwachten dat in de toekomst de ondersteuning van semantische zoeksystemen het best kunnen worden geëvalueerd voor specifieke functionaliteit en een specifieke fase in het zoekproces.