SAMENVATTING

Het Wereldwijde Web (WWW) is een grote collectie van informatie, waarbij onderling verbonden is. Voor eindgebruikers is deze informatie toegankelijk in verschillende bestandstypen. Zo zijn bijvoorbeeld nieuwsartikelen beschikbaar als text, en video's beschikbaar via diensten als YouTube. Hoewel deze verschillende bestandsformaten goed bruikbaar zijn voor mensen, is dit voor software agents daarentegen moeilijker te verwerken.

Naast het WWW bestaat Linked Data. Met Linked Data is het mogelijk om informatie te representeren zodat software agents dit kunnen 'begrijpen'. WWW en Linked Data gebruiken dezelfde fundering om toegang te krijgen tot de informatie, namelijk HTTP. Het verschil is dat Linked Data URLs en kleine stukken tekst gebruikt om 'dingen' te duiden, de zogenaamde 'resources'. Dit kan van alles zijn, variërend van geografische locaties en documenten tot abstracte concepten zoals democratie. De Linked Data architectuur maakt datagebruik mogelijk zonder a priori kennis van de inhoud te hebben. Bovendien is het mogelijk om data te publiceren zonder dat vooraf het publicatiedoel bekend hoeft te zijn. De onbekendheid over (her)gebruik van Linked Data heeft positieve en negatieve kanten: het is goed voor de kwaliteit van de data, maar stelt zowel gebruikers als publicisten voor problemen. Onderstaand presenteren we vijf problemen in deze context, en hoe dit proefschrift bijdraagt aan oplossingen voor deze problemen.

ONPRAKTISCH HERGEBRUIK VAN LINKED DATA

Het publiceren van Linked Data als statische bestanden lijkt eenvoudig, maar zelfs deze methode blijkt in de praktijk lastig: veel Linked Datasets voldoen niet aan de geldende standaarden. Wij hebben een gecentraliseerde dienst ontwikkeld die de LOD Laundromat heet, welke 'schone' versies van zoveel mogelijk Linked Datasets herpubliceert. Dit voorziet in een overvloed van uniforme schone datasets die met weinig moeite gebruikt kunnen worden door software agents.

BEVRAAGBARE LINKED DATA IS DUUR OM TE HOSTEN

De de facto standaard voor het hosten van bevraagbare Linked Data is de 'SPARQL endpoint'. SPARQL is een flexibele en rijke query taal die veel voordelen biedt, maar er zijn ook kosten aan verbonden: deze endpoints zijn duur om te hosten. Wij presenteren twee oplossingen voor dit probleem.

Als eerste presenteren we een 'sampling' methode genaamd SampLD welke de grootte vermindert –en daarmee ook de hardware kosten– voor het hosten van een SPARQL endpoint.

Een tweede orthogonale aanpak is om de complexiteit van de query-taal te verminderen, in plaats van het verminderen van de volledigheid van de query antwoorden. Deze aanpak bestaat uit het combineren van de LOD Laundromat met een 'Triple Pattern Fragments API' die alleen simpele triple patronen ondersteunt. Dit leidt tot een goedkope Linked Data API die een fractie van het geheugen en processor-kracht gebruikt vergeleken SPARQL endpoints.

FORMULEREN VAN SPARQL QUERIES IS MOEILIJK

De complexiteit en expressiviteit van SPARQL, maakt het een moeilijke query-taal. Veel van de meest recente query-bewerkers bieden niet de mogelijkheden waaraan web ontwikkelaars gewend zijn. Daarom hebben wij YASGUI ontwikkeld, een SPARQL query bewerker die toegankelijk is vanuit de browser en veel functies bevat die web ontwikkelaars bekend voorkomen, zoals het markeren en controleren van de syntax, en het aanbieden van suggesties tijden het schrijven van de query.

PROBLEMATISCHE TOEGANG NAAR LINKED (META-)DATA

Het gedistribueerde karakter van Linked Data en de afwezigheid van structurele dataset-omschrijvingen maakt het moeilijk om Linked Data te vinden en te benaderen. Een gecentraliseerde oplossing zoals de LOD Laundromat lost dit probleem niet direct op, omdat het vinden van datasets aan de hand van structurele eigenschappen nog steeds handmatige stappen vereist. Onze aanpak voor dit probleem is om de LOD Laundromat uit te breiden met een structurele Meta-Dataset, welke structurele eigenschappen van de verzamelde datasets bevat, een index om deze datasets te vinden, en herkomst informatie over hoe deze datasets verzameld en geanalyseerd zijn.

VARIËTEIT VAN LINKED DATA ONDERZOEK

Linked Data onderzoek heeft te lijden onder onbeschikbaarheid van documenten en methodes om Linked Data als groot geheel te analyseren. Voor dit probleem biedt het werk in dit proefschrift twee oplossingen.

De eerste aanpak die we presenteren richt zich op Linked Data *gebruik*: dit onderzoeksgebied is sterk beperkt door weinig beschikbare query logs. Wij maken het mogelijk om het gebruik van Linked Data te volgen vanaf de *gebruikers*-kant, door middel van de YASGUI query-bewerker. Wij laten zien hoe de verzamelde queries van YAS-

GUI het mogelijk maken om op grote schaal gebruikers patronen te analyseren.

De tweede aanpak richt zich op het vergroten van de variëteit vanuit een *data*-perspectief. We laten zien dat bestaand Linked Data onderzoek vaak wordt uitgevoerd op enkele datasets, en we presenteren een alternatieve aanpak voor het draaien van experimenten op veel grotere schaal, middels de LOD Laundromat, de bijbehorende Meta-Dataset, en de gepubliceerde LOD Laundromat Triple Pattern Fragments API. Door gedeeltes van drie recente publicaties te herevalueren, laten we zien dat dit nieuwe evaluatie paradigma interessante onderzoeksvragen oproept zoals hoe de prestaties van algoritmes in verband staan met (structurele) eigenschappen van de data.

Dit proefschrift biedt een aantal bijdragen aan voor het bouwen van Linked Data gebaseerde diensten. De oplossingen die we aandragen zijn zowel op de gebruiker als publicist gericht, en zijn een stap in de richting van een meer toegankelijk en technisch schaalbaarder Linked Data web.