Collaborative provenance for workflow-driven science and engineering

Altıntaş, İ.

Publication date
2011

Nederlandse Samenvatting

Samenwerkende Oorsprong Informatie voor Workflows in Wetenschap en Ontwerp

İlkay Altıntaş, December 2010

Met de groei van wetenschappelijke kennis en de stijging van het aantal studies die gebruik maken van informatie van verschillende wetenschappelijke disciplines wordt de complexiteit van systematisch wetenschappelijk onderzoek aanzienlijk vergroot. Om de "grand challenge" wetenschappelijke vragen te beantwoorden, gebruiken wetenschappers computer gebaseerde methodieken die bijna dagelijks aangepast worden. Echter, de standaard "scientific method" blijft hetzelfde, maar wordt telkens getransformeerd door de verbeteringen in de computer wetenschap en technologie van de laatste decennia. Deze veranderingen in de computer wetenschap en technologie hebben geleid tot een verzameling van gereedschappen specifiek ontworpen om het wetenschappelijke proces meer efficiënt en sneller te maken. Deze gereedschappen zijn in staat om de creatie en uitvoering van opeenvolgende en coherente computer gestuurde taken te vereenvoudigen en zijn bekend onder de naam van "wetenschappelijke workflows".

Behoorlijke ontwikkelingen in computer technologische hebben plaats gevonden vanaf de initiatie van de wetenschappelijke workflows aan het eind van de jaren negentig en de eerste toepassingen van wetenschappelijke workflows voor het oplossen van uitdagingen in visualisatie. Wetenschappelijk workflows zijn geëvolueerd om te voldoen aan verschillende wetenschappelijke eisen, computer gebaseerde technologieën en wetenschappelijke aanpakken die de "scientific method" hebben getransformeerd in een zwaar computer getuurd proces. Bovendien, terwijl wetenschappers meer leren over het efficiënt ontwerpen en uitvoeren van de wetenschappelijke workflows, wordt het belangrijk om in de gaten te houden hoe en wanneer specifieke wetenschappelijke informatie wordt verkregen, d.w.z. het vastleggen van de oorsprong van de informatie (provenance tracking).

Het vastleggen van de oorsprong (provenance tracking) is een belangrijke eigenschap van wetenschappelijke workflow systemen omdat het hulp biedt in het bijhouden van de origine van wetenschappelijke eind produkten en het valideren en herhalen van computer gebaseerde methodieken die gebruikt waren voor de afleiding van deze wetenschappelijke produkten. Het vastleggen van de oorsprong in wetenschappelijke workflows begint met het ontwerp en executie van de wetenschappelijke workflow en de vergaarde informatie moet de mogelijk

bieden om de associaties tussen workflow ingang variabelen, workflow uitgang variabelen, workflow definities en tussenliggende gegevens te creëren en te onderhouden. De oorsprong van een produkt bestaande uit gegevens bevat informatie hoe het produkt was verkregen en is cruciaal voor wetenschappers om eenvoudig de wetenschappelijke resultaten te begrijpen, herproduceren en verifiëren.

De meeste modellen voor het vastleggen van de oorsprong worden tegenwoordig ontworpen om alleen de oorsprong van een enkele executie en veelal door een enkelvoudige gebruiker vast te leggen. Wetenschappelijke ontdekkingen zijn echter vaak het resultaat van een systematische executie van verscheidene wetenschappelijke workflows met meerdere verzamelingen van gegevens geproduceerd op verschillende tijdstippen bij één of meer gebruikers. Om de uitwisseling van informatie te bevorderen en mogelijk te maken tussen meerdere workflow systemen die oorsprong informatie toelaten, de Open Provenance Model (OPM) is door de wetenschappelijke workflow gemeenschap voorgesteld. Standaards zoals OPM bieden de mogelijkheid oorsprong informatie van verschillende wetenschappelijke workflows executies uitgevoerd op verschillende systemen te verbinden. Dit leidt tot een impliciete samenwerking tussen de gebruikers die de wetenschappelijke workflows ontwerpen en uitvoeren.

Dit proefschrift presenteert vier belangrijke bijdragen op het gebied van het vastleggen van oorsprong informatie in samenwerkende workflows, genaamd “collaborative provenance”. Ten eerste wordt een overzicht gegeven van het effect van wetenschappelijke workflows op hoe wetenschappelijk onderzoek plaats vindt met een concentratie op het vastleggen van oorsprong informatie als een specifieke voordeel van wetenschappelijke workflows. Ten tweede wordt een definitie gegeven voor oorsprong informatie in impliciete samenwerking tussen gebruikers dat gegenereerd kan worden met oorsprong informatie verkregen van wetenschappelijke workflows van een on-line gemeenschappelijke verzameling van gelijke gebruikers. Ten derde wordt een nieuw zoekopdracht (query) model beschreven dat de impliciete samenwerking tussen de gebruikers kan omvatten, laat zien hoe dit model omgezet kan worden naar de OPM en de mogelijkheid biedt om een antwoord te geven op vragen met betrekking samenwerking, bijv. de identificatie van van gecombineerde workflows en de bijdragen van gebruikers die samenwerken in een project gebaseerd op de resultaten van voorafgaande workflow executies. De aanpassing en uitbreiding van het hoge-niveau Query Language for Provenance (QLP) met additionele concepten stelt gebruikers die geen experts zijn in staat om eenvoudig en precies samenwerkende oorsprong vraagstellingen te formuleren. Tenslotte wordt in dit proefschrift een data model geformuleerd dat effectief is in het vastleggen van scenario’s in samenwerkende oorsprong informatie. Het wordt aangetoond hoe dit data model gebruikt kan worden om antwoorden te formuleren op computer gestuurde vraagstellingen met betrekking tot samenwerkende oorsprong informatie, bijv. identificatie van gecombineerde gegevens, workflow executie, en bijdragen van gebruikers die samenwerken in een project gebaseerd op de resultaten van voorafgaande workflow executies.

De voornaamste bijdragen in het vastleggen van samenwerkende oorsprong informatie in dit proefschrift leiden tot de ontwikkeling van computer systemen die interoperatief samenwerken en de herbruik van workflow resultaten vergroten. Hierdoor zal de effectiviteit en

produktiviteit van moderne wetenschappelijke samenwerking toenemen. Dit effect is ook gedemonstreerd in dit proefschrift via de wetenschappelijke test scenario's voor het oprichten en begrijpen samenwerkende studies via interoperatieve workflow oorsprong informatie. Specifiek, de Virolab scenario, Provenance Challenge 1 en 3 workflow, en meerdere on-line gemeenschappelijke gelijke gebruikers van het CAMERA project zijn aangepast als samenwerkende en interoperatieve test scenario's, waarbij verschillende gedeeltes van de workflow uitgevoerd worden als verschillende workflows en mogelijkerwijs ook in verschillende workflow omgevingen.