Feature grammar systems. Incremental maintenance of indexes to digital media warehouses

Windhouwer, M.A.

Samenvatting

Met de opmars van personal computers en allerlei vormen van randapparatuur om traditionele media te digitaliseren, met name onder de invloed van het als maar dalen van de aanschafkosten en het stijgen van de opslagcapaciteit, zijn grote collecties van digitale media (*digital media warehouses*) gemeengoed geworden. Maar het opslaan van digitale objecten is slechts één kant van de zaak, de gebruikers willen deze objecten ook weer terugvinden. Het terugvinden van deze objecten is echter geen sinecure en een omvangrijke en multidisciplinaire onderzoeksgemeenschap houdt zich daar dan ook mee bezig.

De focus van dit proefschrift wordt gevormd door één stap in het zoekproces. De media objecten in hun digitale vorm zijn namelijk niet gemakkelijk te vinden, daarvoor moeten ze geannoteerd worden. Deze annotaties kunnen zowel handmatig als automatisch gecreëerd worden. In het laatste geval worden de annotaties geproduceerd door uitgeprogrammeerde extractie-algoritmes, die op de objecten worden losgelaten.

De extractie-algoritmes zijn van elkaar en van elkaars annotaties afhankelijk. Allereerst kan er sprake zijn van een uit/invoer afhankelijkheid: de uitvoer van het ene algoritme is de invoer van een volgend algoritme. Een voorbeeld hiervan is het bepalen van het type van een afbeelding, een tekening of een foto. Dit gebeurt op basis van eerder geproduceerde annotaties, zoals het aantal en de gemiddelde verzadiging van de kleuren in de afbeelding. Daarnaast is er de mogelijkheid van een context afhankelijkheid. Hierbij wordt een extractie-algoritme alleen uitgevoerd als een eerder algoritme geslaagd is. Dit wordt geïllustreerd door de volgende afhankelijkheid: het bepalen of een afbeelding een of meerdere gezichten bevat is alleen nodig als eerst bepaald is dat de afbeelding een foto is.

Een complicatie is de semantiek van de annotaties. Eenvoudige annotaties, bijvoorbeeld de kleur geel komt in deze afbeelding voor, zijn eenduidig. Meer abstracte annotaties, zoals dit is een grimmige foto, zijn ambigu. Hun validiteit is afhankelijk van de context waarin het object zich bevindt, of de (cultureel bepaalde) context van de gebruiker. Een annotatie systeem moet dan ook de productie en het gebruik van alternatieve interpretaties ondersteunen.

In het proefschrift wordt een formele taal, kenmerk grammatica systemen (*feature grammar systems*), die voor het gelijktijdig beschrijven van de afhankelijkheden en de (alternatieve) contexten is ontwikkeld. In een natuurlijke taal worden valide zinnen beschreven door een grammatica. De grammatica bepaalt welke woorden samen, in een specifieke context, mogen voorkomen. Een kenmerk grammatica systeem doet hetzelfde voor annotaties en extractie-algoritmes. Daartoe bestaat een kenmerk grammatica systeem uit één of meerdere grammatica componenten. Elk component beschrijft het resultaat van een algoritme en de afhankelijkheid van andere componenten. Deze beschrijving kan alternatieve interpretaties bevatten.

Het extractie-proces kan nu gestuurd worden door een kenmerk grammatica systeem te interpreteren. Dit proces komt overeen met het parseren van zinnen in een natuurlijke of artificiële taal. Hiervoor zijn door de jaren heen veel efficiënte algoritmes ontwikkeld. Echter slechts enkele hiervan zijn geschikt voor kenmerk grammatica systemen. Een probleem is het dynamisch groeien van de annotatie zin: het activeren van een grammatica component leidt tot het uitvoeren van een extractie-algoritme en dus tot de productie van annotaties. Een ander probleem wordt gevormd door de afhankelijkheden: een extractie-algoritme kan pas worden uitgevoerd als zijn invoer, reeds eerder geproduceerde annotaties, beschikbaar is. Hierdoor komen alleen parseer algoritmes in aanmerking die van boven naar beneden werken. Een specifiek algoritme, die aan deze kenmerken voldoet, is geimplementeerd en produceert de annotaties, beschreven door een of meerdere parseerbomen. Deze bomen worden opgeslagen in een database management systeem.

Verschillende factoren, zoals wijzigingen in de algoritmes, kunnen er echter toe leiden dat de opgeslagen bomen, en dus ook de annotaties, niet meer de werkelijkheid weerspiegelen. Om de invloed van deze wijzigingen te lokaliseren wordt er van het kenmerk grammatica systeem een afhankelijkheidsgraaf afgeleid. Een planningsproces kan dan het extractie-proces gedeeltelijk herstarten om daarmee de database te modificeren.

Het aldus ontworpen systeem, Acoi, is de afgelopen jaren aan het CWI ontwikkeld en ingezet bij verschillende praktijkstudies. Analyse van deze studies toont aan dat een kenmerk grammatica systeem een praktisch inzetbaar hulpmiddel is om (alternatieve) annotaties te produceren en te onderhouden.