Eenvoudige modellen en Big Data troeven slimme modellen af*

Djoerd Hiemstra

Universiteit Twente
http://www.cs.utwente.nl/~hiemstra

Inleiding

Big Data – of het beter allitererende “Grote Gegevens” – is een term die sinds het begin van deze eeuw wordt gebruikt om gegevensverzamelingen aan te duiden die moeilijk verwerkt konden worden met behulp van de software van die tijd, verzamelingen van vele terabytes of petabytes in grootte. Technieken om zulke enorme verzamelingen gegevens te kunnen verwerken en analyseren werden met name ontwikkeld door Google. Het uitgangspunt van Google: Zet heel veel goedkope machines bij elkaar in grote datacentra, en gebruik slimme gereedschappen zodat applicatieontwikkelaars en gegevensanalisten het hele datacentrum kunnen gebruiken voor hun gegevensanalyses. Het datacentrum is de nieuwe computer! De slimme gereedschappen van Google raken veel kernelementen van de Informatica: bestandssystemen (Google File System), nieuwe programmeerparadigma's (MapReduce), nieuwe programmeertalen (bijvoorbeeld Sawzall) en nieuwe aanpakken voor het beheren van gegevens (BigTable), allemaal ontwikkeld om grote gegevensverzamelingen gemakkelijk toegankelijk te maken. Deze technieken zijn inmiddels ook beschikbaar in *open source* varianten. De bekendste, Hadoop, werd voor een belangrijk deel ontwikkeld bij Googles concurrent Yahoo. Aan de Universiteit Twente worden de technieken sinds 2009 onderwezen in het masterprogramma *Computer Science*.

De kracht van grote gegevens

Grote gegevens introduceren nieuwe uitdagingen en nieuwe kansen voor machinaal leren, met name toegepast op natuurlijke taalverwerking, zoals het zoeken op het web, automatische vertaling en spraakherkenning. In het artikel “The unreasonable effectiveness of data” beschrijven Google onderzoekers Alon Halevy, Peter Norvig en Fernando Pereira (2009) de volgende belangrijke uitdaging: *Maak gebruik van grootschalige gegevens die direct beschikbaar zijn, in plaats van te hopen op geannoteerde gegevens die (nog) niet beschikbaar zijn*. Zoekmachines, automatische vertalers en spraakherkenners zijn de laatste jaren enorm verbeterd dankzij de enorme hoeveelheid gegevens die beschikbaar zijn voor het trainen van statistische modellen. Veel gegevens worden routinematig gemaakt, en zijn dus in overvloed aanwezig. Voorbeelden zijn de hyperlinks tussen webpagina's, vertalingen van websites die in meerdere talen beschikbaar zijn; ondertitels voor doven en slechthorenden, de geografische positie bij berichten en foto's als gevolg van GPS in smartphones en camera's, enz., enz.

* Geschreven voor STAtOR 14(3-4), Tijdschrift van de Vereniging voor Statistiek en Operationele Research.

De kracht van simpele modellen

De statistische technieken waarmee systemen getraind worden zijn de afgelopen jaren minder veranderd dan hun succes doet vermoeden. Nu we in staat zijn om te trainen op grootschalige gegevensverzamelingen doet zich het volgende fenomeen voor: *Eenvoudige modellen getraind met grote gegevens troeven complexe modellen op basis van minder gegevens af*. Michele Banko en Eric Brill (2001), beiden onderzoekers bij Microsoft, waren een van de eersten die dit aantoonden. Ze trainden verschillende methoden voor het desambigueren van woorden met behulp van gegevensverzamelingen van verschillende groottes. Geen van de methoden leek nog asymptotisch gedrag te vertonen bij het trainen op een miljoen woorden, een redelijke hoeveelheid gegevens – zeker voor die tijd. Interessant is dat simpele methoden die het relatief slecht doen op een miljoen woorden, niet onder doen voor complexe modellen als er een miljard woorden beschikbaar is. Eenvoudige methoden die grote gegevens kunnen benutten hebben de voorkeur boven complexer methoden die grote gegevens niet gemakkelijk aankunnen.

Torsten Brants en collega's bij Google (2007) deden een soortgelijk onderzoek naar methoden voor het trainen van statistische taalmodellen voor automatisch vertalen, waarbij ze in staat waren om maarliefst een biljoen woorden te gebruiken. Ze introduceerden daarvoor een nieuwe, eenvoudige methode voor *smoothing* van statistsche taalmodellen, *Stupid Backoff* genoemd. Deze methode is goedkoop om te trainen op grote gegevens en benadert de kwaliteit van het complexere en krachtigere *Kneser-Ney smoothing* als de hoeveelheid trainingsgegevens toeneemt. Het werk van Brants en collega's laat zien dat een complex model gebaseerd op grote gegevens, niet beter zal presteren dan een simpel model gebaseerd op dezelfde hoeveelheid gegevens. De ware kennis zit blijkbaar in de gegevens, niet in de generalisaties die een slim model kan doen.

De kracht van grote gegevens wordt letterlijk geïllustreerd door het werk van James Hays en Alexei Efros (2007) van Carnegie Mellon University, die *afbeeldingsvoltooiing* onderzoeken: dat wil zeggen, het invullen of vervangen van een deel van een foto of afbeelding zodanig dat de wijziging niet kan worden gedetecteerd. Voorbeelden zijn het herstellen van een kapotte hoek van een historische foto, of het verwijderen van een ex-echtgenoot van een familiefoto. De methode van Hays en Efros is verbazend simpel. Hun algoritme zoekt in een enorme database naar soortgelijke afbeeldingen en gebruikt daarvan beeldfragmenten om de afbeelding te voltooien. Ongetwijfeld zou een dergelijke methode niet werken op een kleine database, maar zodra de database groot genoeg is – de auteurs hadden miljoenen foto's – dan overtreft de methode complexere methoden, bijvoorbeeld methoden die de ontbrekende delen van een afbeelding door de analyse van kleuren en textuur proberen te extrapoleren uit de bekende delen.

Een ander mooi voorbeeld van de kracht van grote gegevens is het vraag-en-antwoord systeem van Susan Dumais en collega's (2002). Vraag-en-antwoord systemen beantwoorden vragen zoals: “Wanneer werd Vincent van Gogh geboren?”, “Waar schreef Anne Frank haar dagboek?”, of “Wie was de eerste Nederlandse koning?” Zulke systemen maken veelal gebruik van uitgebreide taalkundige kennis, zoals het ontleden van zinnen, het modelleren van personen, plaatsen en tijden, anaforenresolutie, het gebruik van synoniemen, enz. De aanpak van Dumais en haar collega's beperkt zich tot het herschrijven van de vragen naar simpele beweringen, en het gebruik van een internetzoekmachine waarmee documenten met die exacte bewering wordt gevonden, waarna de tekst volgend op de bewering wordt

verzameld. De redenering is als volgt: Als de gegevens groot genoeg zijn, is er altijd wel een bewering te vinden die exact overeen komt met de gestelde vraag, bijvoorbeeld: “Vincent van Gogh werd geboren… ”, “Anne Frank schreef haar dagboek...” en “De eerste Nederlandse koning is...”. Op de plaats van de puntjes vindt het systeem vervolgens het antwoord, met minimaal gebruik van taalkundige kennis. Ook deze aanpak zou nooit werken als er geen grootschalige gegevensverzamelingen beschikbaar zouden zijn. Hoe groot moeten de gegevens zijn voordat zo'n aanpak succesvol is? Het werk van Arjen Hoekstra (2006) laat zien dat ook de hoeveelheid Nederlandse pagina's voldoende groot is voor een “grote gegevens”-aanpak.

Iedereen kan “Google zijn”

Hoe groot is groot? Torsten Brants en collega's van Google trainden hun statistische taalmodellen met behulp van een biljoen woorden. Zelfs als een webpagina gemiddeld 1000 woorden bevat, dan komt dat nog steeds neer op een miljard webpagina's, vele terabytes aan gegevens. Is dat haalbaar voor onderzoekers die niet bij Google in dienst zijn? Het antwoord hierop is: Ja hoor, dat was jaren terug al haalbaar. Eind 2008 kocht de Universiteit Twente, gesponsord door Yahoo, een eerste Hadoop cluster bestaande uit 16 machines, ongeveer € 1000,- per stuk, die gezamenlijk terabytes aan gegevens op kunnen slaan. Bovendien gaf Carnegie Mellon University begin 2009 de ClueWeb09 webcollectie vrij, een gegevens-collectie bestaande uit een miljard webpagina's. Gecombineerd met de gegevens die op dat moment al aan de Universiteit Twente beschikbaar waren, was dat genoeg voor een biljoen woorden. Medio 2009 had de universiteit de dus de kennis, de infrastructuur, en de gegevens in handen om ook een biljoen woorden te analyseren. Wat Google vandaag publiceert, kan men binnen 3 a 4 jaar dus repliceren. Wat vandaag grote gegevens zijn, zijn morgen gewone gegevens. Iedereen kan Google zijn.

Makkelijker zoeken in grote gegevens

De experimenten met ClueWeb09 laten interessante gevallen zien waarbij simpele modellen net zo goed werken als complexere modellen. De literatuur van het onderzoek naar zoekalgoritmen, in het Engels *Information Retrieval* genoemd, kent een aantal standaard heuristieken en ordeningsprincipes. De *tf.idf* weging is bijvoorbeeld een bekende heuristiek waarbij het belang van een document voor een zoekvraag wordt berekend uit de *term frequency*, of *tf* waarde van een term: het aantal voorkomens van de term in het document en de *inverse document frequency* of *idf* waarde van een term: het aantal documenten waarin de term voorkomt. De intuïtie is dat termen die in weinig documenten voorkomen (met een hoge *idf*) belangrijker zijn dan termen die in alle documenten voorkomen, en dat documenten met veel voorkomens van de termen (met een hoge *tf*) belangrijker zijn dan documenten met weinig voorkomens van de termen. De *tf.idf* heuristiek kan ook verklaard worden door het gebruik *smoothing* van statistische taalmodellen.

Uit de ClueWeb09 experimenten blijkt dat *smoothing*, of het gebruik van de *idf* component in *tf.idf*, niet langer noodzakelijk is voor het met hoge precisie vinden van documenten op het web. Simpele modellen – taalmodellen zonder *smoothing*, of een weging met enkel de *tf* component – werken net zo goed als de gegevens maar groot genoeg zijn (Hiemstra & Hauff 2011). Zoeken wordt dus gemakkelijker als de gegevens waarin gezocht wordt groter zijn.

Conclusie

In de wereld van “grote gegevens” troeven eenvoudige modellen de complexe modellen af. Toch zijn veel onderzoekers nog altijd in de weer om betere, complexere modellen te bedenken, om dan met weinig gegevens aan te tonen dat die modellen een verbetering opleveren. De in dit artikel beschreven ervaringen met grote gegevens suggereren het volgende advies aan onderzoekers: Misschien is het goed om even pas op de plaats te maken, en de komende tijd te besteden aan het verzamelen van *grote gegevens* in plaats van aan het ontwikkelen van nieuwe, complexere methoden.

Literatuur

Michele Banko and Eric Brill (2001). Scaling to very large corpora for natural language disambiguation. In: *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics (ACL)*.

Thorsten Brants, Ashok Popat, Peng Xu, Franz Och, Jeffrey Dean (2007). Large Language Models in Machine Translation. In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Susan Dumais, Michele Banko, Eric Brill, Jimmy Lin, and Andrew Ng (2002). Web Question Answering: Is more always better? In: *Proceedings of the 25th ACM Conference on Research and Development in Information Retrieval (SIGIR)*.

Alon Halevy, Peter Norvig, and Fernando Pereira (2009). The unreasonable effectiveness of data. *IEEE Intelligent Systems, 24*(2).

James Hays and Alexei Efros. Scene Completion Using Millions of Photographs (2007) *ACM Transactions on Graphics (SIGGRAPH)* 26(3).

Djoerd Hiemstra and Claudia Hauff (2011). MapReduce for Experimental Search. In *Proceedings of the 19th Text Retrieval Conference (TREC), NIST Special Publications*.

Arjen Hoekstra, Djoerd Hiemstra, Paul van der Vet and Theo Huibers (2006). Question Answering for Dutch: Simple does it. In: *Proceedings of the 18th BeNeLux Conference on Artificial Intelligence (BNAIC)*.