Nederlandse Samenvatting

In dit proefschrift staan twee Nederlandse samenvattingen. De eerste samenvatting is geschreven voor mensen zonder achtergrond in de informatica. Deze zal op een toegankelijke manier uitleggen wat er voor dit proefschrift onderzocht is en wat de resultaten zijn. Hierbij wordt vooral gefocust op de context van het onderzoek en minder op de technische details. De tweede samenvatting is geschreven voor mensen met een zekere achtergrondkennis binnen de informatica. Hier wordt de context als bekend beschouwd en ligt de nadruk op de gebruikte technieken.

Samenvatting voor Leken

Computers spelen een steeds grotere rol in ons dagelijks leven. Hun enorme opslagvermogen biedt veel mogelijkheden. Hoe kunnen we de opgeslagen data nog beter in ons voordeel gebruiken? Dit kan door naar interessante patronen in deze databestanden te zoeken. *Datamining* (het delven van data) is de tak binnen de informatica die zich hiermee bezig houdt.

Wat een interessant patroon is, hangt onder andere af van de gebruikersvraag. Verschillende vragen worden beantwoord door verschillende dataminingprogramma's. Neem een databestand met teksten van boeken. Een gebruiker wil de boeken die over hetzelfde onderwerp gaan bij elkaar verzamelen. Deze toepassing noemen we *clusteren*. Nog een voorbeeld: een databestand bevat eigenschappen van eiwitten waarbij voor een deel van deze eiwitten bekend is in hoeverre ze gebruikt kunnen worden voor de ontwikkeling van medicijnen. Onderzoekers naar geneesmiddelen willen de eigenschappen van eiwitten die geschikt zijn voor het maken van medicijnen vergelijken met de eigenschappen van eiwitten die niet geschikt zijn. Dit kan helpen bepalen of een onbekend eiwit een goede kans heeft aan de basis te staan voor de ontwikkeling van een medicijn. Dit staat bekend als *classificatie*. Weer een derde databestand houdt van klanten in een supermarkt bij wat ze per bezoek kopen. Voor de winkelier is het onder meer interessant om te weten welke artikelen vaak tegelijk gekocht worden om zijn winkel beter in te delen. Deze techniek staat bekend als *associatie analyse*.

Naast de gebruikersvraag, is het soort data van belang bij de ontwikkeling van een dataminingprogramma. Neem het eerder genoemde clusteren van boeken. Het is voor een computer niet makkelijk om te bepalen in hoeverre twee boeken over hetzelfde onderwerp gaan. Een mogelijke oplossing is om te kijken naar hoe vaak woorden in boeken voorkomen. Twee boeken die veel woorden gemeenschappelijk hebben, hebben een grotere kans over hetzelfde onderwerp te gaan dan twee boeken die relatief weinig woorden gemeenschappelijk hebben. In dit geval is het soort informatie *inhoudelijk* van aard; de boeken worden geclusterd op basis van hun inhoud.

Informatie kan echter ook *relationeel* van aard zijn. Een sociaal netwerk zoals Facebook kan je zien als een groot databestand. Hier is de belangrijkste informatie of er een relatie bestaat tussen twee mensen. Clusteren betekent hier dus het verdelen van gebruikers in groepen 'vrienden'. Het programma probeert clusters te maken waarbij er relatief veel verbindingen zijn tussen mensen in hetzelfde cluster, maar relatief weinig verbindingen van het ene cluster naar het andere. In dit geval is de informatie relationeel; de gebruikers worden geclusterd op basis van hun onderlinge relaties.

Een clusterprogramma dat werkt op basis van inhoudelijke informatie werkt heel anders dan een clusterprogramma dat werkt op basis van relationele informatie. In de loop der tijd zijn er voor beide soorten informatie veel verschillende clusterprogrammas ontwikkeld. Het kan echter zijn dat een databestand bestaat uit beide soorten informatie. Dit betekent dat een clusterprogramma dat werkt op basis van inhoudelijke informatie niets kan met de relationele informatie die ook beschikbaar is en vice versa.

Een voorbeeld ter verduidelijking: in een databestand met wetenschappelijke artikelen is per artikel een korte samenvatting van de inhoud bekend en ook is bekend welke artikelen door welk ander artikel geciteerd worden. De samenvattingen zijn de inhoudelijke informatie en de citaties zijn de relationele informatie. Stel dat er in de samenvatting van een bepaald artikel voornamelijk woorden staan die te maken hebben met het ene onderwerp, maar dit artikel citeert voornamelijk artikelen die gaan over het andere onderwerp. Bij welk onderwerp moet je dit artikel dan indelen? Een clusterprogramma dat slechts naar n soort informatie (inhoudelijk of relationeel) kijkt, zal makkelijk tot een oordeel komen, maar er bestaat dan een redelijke kans dat dit niet het juiste onderwerp is.

De methode die in dit proefschrift wordt beschreven, probeert beide soorten informatie te benutten om zo tot een beter resultaat te komen. Dit betekent dat in het ene geval de inhoudelijke informatie de doorslag krijgt en in het andere geval de relationele informatie, afhankelijk van welke het meest bepalend is. De methode is algemeen toepasbaar in dataminingprogrammas die ontworpen zijn voor inhoudelijke informatie. Dit betekent dat je de methode eigenlijk kan zien als een soort plugin die relationele informatie toevoegt in een bestaand dataminingprogramma dat zelf alleen kijkt naar de inhoud. De methode is

getest op diverse clusterprogrammas en op een classificatieprogramma. In alle gevallen kwam het tot betere resultaten dan zonder het gebruik van deze plugin. Er kan dus geconcludeerd worden dat het met deze methode mogelijk is om de resultaten van bestaande dataminingprogrammas te verbeteren.

Samenvatting voor Informatici

Het ontwerp van een dataminingalgoritme hangt van veel factoren af. Een belangrijke factor is het soort data dat er gebruikt wordt. Er zijn heel veel verschillende soorten en voor elk is er een grote verscheidenheid aan dataminingalgoritmes ontworpen. Wanneer een databestand echter meerdere soorten data bevat, zal de gebruiker moeten kiezen welke hij wel en vooral welke hij niet wil benutten. Een alternatief is het ontwikkelen van een nieuw dataminingalgoritme voor deze combinatie van datatypen. Het aantal combinaties van verschillende datatypen in een databestand is echter exponentieel in verhouding tot het aantal soorten data en het is derhalve ondoenlijk om voor elke combinatie aparte dataminingalgoritmes te ontwikkelen.

In dit proefschrift onderscheiden wij twee belangrijke categorien informatie: inhoudelijke en relationele. Inhoudelijke informatie zegt iets over hoe een element eruit ziet en relationele informatie zegt iets over hoe een element zich verhoudt tot een ander element. In een geannoteerde graaf kan je de annotatie van de knoop beschouwen als de inhoudelijke informatie en de takken tussen de knopen als de relationele informatie. De methode die wij in dit proefschrift presenteren, voegt indirect relationele informatie toe aan een algoritme dat is ontworpen voor inhoudelijke informatie. Zo hoeft er dus niet voor elke combinatie van een inhoudelijk en een relationeel datatype in hetzelfde databestand een nieuw dataminingalgoritme ontworpen te worden.

De methode die in dit proefschrift gepresenteerd wordt, is in principe toepasbaar op elk dataminingalgoritme dat gebruik maakt van een afstands- of gelijkaardigheidsmaat op basis van inhoudelijke informatie. Onze methode kan dan gezien worden als een plugin die deze maat vervangt voor een die ook de relationele informatie in ogenschouw neemt. Dit wordt gedaan door voor de afstand of gelijkaardigheid tussen twee knopen ook de annotaties van de buren van deze knopen te gebruiken. Zo wordt de relationele informatie dus indirect gebruikt om een nieuwe maat te genereren.

Een goede vergelijkingsmaat is belangrijk voor een goed resultaat. In de praktijk is het echter heel moeilijk om een goede maat te ontwerpen. Het kan dus zeer goed voorkomen dat zon maat in sommige gevallen een verkeerde uitkomst genereert wat er vervolgens voor kan zorgen dat een element in het verkeerde cluster wordt ingedeeld. Onze methode zorgt er voor dat een dataminingalgoritme robuuster beter bestand is tegen dit soort fouten.

Als nu de inhoud van een knoop niet zo erg lijkt op de inhoud van knopen uit zijn eigen categorie, is het voor een vergelijkingsmaat moeilijk om de juiste

waarde voor die knoop in verhouding tot andere knopen te bepalen. Hierdoor wordt deze knoop waarschijnlijk verkeerd ingedeeld. In veel databestanden met relationele informatie zijn het vooral knopen die in dezelfde categorie horen die met elkaar verbonden zijn. Dit betekent dat over het algemeen de buren van een knoop voor het grootste gedeelte uit dezelfde categorie komen. De kans dat alle inhouden van de buren sterk afwijken van de gemiddelde inhoud van knopen uit die categorie is zeer klein. In dit geval geven de buren van deze knoop een betere representatie van de categorie van deze knoop en dus zorgt het feit dat onze methode gebruik maakt van deze buren ervoor dat de resultaten beter zullen zijn.

Onze methode is getest voor diverse clusteringalgoritmes en voor een classificatiealgoritme. In alle gevallen waren de resultaten beter dan het originele algoritme. Hieruit kan geconcludeerd worden dat deze methode in staat is om bestaande dataminingalgoritmes te verbeteren.

Het proefschrift bestaat uit vier delen. Het eerste deel behandelt de theoretische basis. In hoofdstuk 2 wordt de stand van zaken in data mining, voor zover relevant voor dit proefschrift, besproken. Hoofdstuk 3 beschrijft uitvoerig hoe de methode werkt. Ten slotte geeft hoofdstuk 4 een intutieve motivatie waarom onze methode voor betere resultaten kan zorgen.

Het tweede deel beschrijft de experimenten die gedaan zijn om de praktische toepasbaarheid aan te tonen. In hoofdstuk 5 wordt de methode toegepast op agglomerative hierarchical clustering. Hoofdstuk 6 wordt dit vervolgens gedaan voor k-means en k-medoids. Dit hoofdstuk laat tevens zien waarom deze methode niet direct kan worden toegepast op k-means en wat er kan worden gedaan om onze methode hier toch te gebruiken. In hoofdstuk 7 wordt deze methode toegepast op KNN-classification.

Het derde deel analyseert de werking van onze methode. In hoofdstuk 8 worden de effecten van ruis op de inhoud en ruis op de relaties met elkaar vergeleken. Hier wordt aangetoond dat de *homophily* in een geannoteerde graaf kan compenseren voor de *content variability*. Hoofdstuk 9 bespreekt welke varianten op onze methode mogelijk zijn en hoe deze ten opzichte van elkaar presteren.

Het vierde deel presenteert in hoofdstuk 10 tot slot de conclusies die uit alle onderzoeken kunnen worden getrokken.