Mining Structured Data

Nijssen, Siegfried Gerardus Remius

Samenvatting

In het laatste decennium is de hoeveelheid data enorm toegenomen, zowel in bedrijven als op wetenschappelijk gebied. Als gevolg hiervan is er een toenemende behoefte aan algoritmen die gegevens kunnen analyseren. Eén van de onderzoeksgebieden die zich bezig houdt met het analyseren van gegevens met behulp van de computer is *data mining*. Dit onderzoeksgebied ontwikkelt algoritmen voor het vinden van verbanden in grote hoeveelheden data.

De meeste data mining methoden die tot nu toe ontwikkeld zijn, gaan uit van data die op te slaan zijn in één tabel bestaande uit rijen en kolommen gevuld met cijfers. Voor veel toepassingen is dit een acceptabele aanname en is het relatief eenvoudig één tabel op te bouwen. In sommige gevallen is dit echter minder voordehandliggend, en is het onduidelijk waaruit de tabel zou moeten bestaan: voor het bouwen van de tabel is in zo'n geval ook een algoritme nodig.

Een voorbeeld van een toepassing waarbij dit van bijzonder belang is, is het zoeken naar eigenschappen van moleculen die kunnen leiden tot giftigheid of kankerverwekkendheid. Er zijn grote hoeveelheden moleculaire databanken en er kan op allerlei manieren naar deze data gekeken worden. Natuurlijk zijn moleculen oorspronkelijk 3 dimensionale structuren, maar een '2 dimensionale' representatie, waarin alleen rekening wordt gehouden met atoombindingen, kan soms tot beter inzicht leiden. Voor toepassing van de meeste data mining methoden is het noodzakelijk de moleculen eerst terug te voeren tot een vast aantal eigenschappen die in cijfers uit te drukken zijn. Het aantal manieren om dat te doen is welhaast oneindig, en het is vaak op voorhand onduidelijk welke representatie de voorkeur verdient.

In dit proefschrift bestuderen we daarom algoritmen die, onder andere, kunnen helpen bij het vinden van de juiste representatie voor structuren, zoals moleculen. Op aanwijzingen van een expert is de taak van de algoritmen in een grote ruimte te zoeken naar interessante patronen in structuren, en op die manier verder inzicht te verkrijgen in de gestructureerde data. De patronen kunnen dan gebruikt worden om een zinvolle representatie voor structuren te verkrijgen.

Daartoe beginnen we met een overzicht van recent onderzoek in 'inductive databases'. Alhoewel er geen alom erkende definitie is voor wat 'inductive databases' eigenlijk zijn, is één mogelijke definitie gebaseerd op de analogie tussen data mining en data querying. In de beginjaren van databasetechnologie was het gebruikelijk om een specifiek databasesysteem te schrijven voor elke toepassing. Tegenwoordig wordt meestal gebruik gemaakt van een algemeen systeem. Aan zo'n algemeen systeem is een programmeertaal verbonden waarin de toegang tot de database voor specifieke gevallen bechreven kan worden (er kunnen 'queries' geschreven worden in deze taal). Het idee achter inductive databases is dat het wellicht ook mogelijk is om het zoeken naar patronen in data op zo'n manier op te lossen. Er wordt dan

een relatief algemeen systeem ontwikkeld, dat vervolgens door onderzoekers of databasebeheerders gebruikt kan worden om specifiekere vragen te beantwoorden.

In dit proefschrift beperken we ons tot declaratieve queries, die erg op traditionele database queries lijken: de gebruiker kan een verzameling vereisten ('constraints') aan de patronen specificeren waarin zij geïnteresseerd is, en de taak van het algoritme is om alle patronen te vinden die aan deze eisen voldoen. Een mogelijke constraint, die in het laatste hoofdstuk uitvoerig aan bod komt, is de minimum correlation constraint, die verlangt dat een patroon een voldoende hoge waarde haalt in een χ^2 statistische test. We laten zien dat deze constraint nauw verbonden is met de minimum support constraint, welke inhoudt dat we alleen patronen willen vinden die in tenminste *minsup* voorbeelden in een databank voorkomen, voor een voorafgegeven, door de gebruiker gespecificeerde, grenswaarde *minsup*.

Aangezien er voor simpele, tabelvormige databanken veel onderzoek gedaan is naar efficiënte algoritmen voor het vinden van patronen met hoge support, geven we een overzicht van dit onderzoek, voor zover dat van belang is voor ons werk.

Het grootste deel van dit proefschrift bestaat vervolgens uit een studie van algoritmen voor het ontdekken van patronen in databanken die niet eenvoudig in één tabel op te slaan zijn. In eerste instantie bestuderen we het gebruik van eerste orde logica als representatie voor data en patronen. Aangezien eerste orde logica zeer expressief is, is het algoritme dat we hier ontwikkelen zeer algemeen toepasbaar. Ons algoritme bestaat uit enkele uitbreidingen van een bestaande methode: allereerst ontwikkelen we een nieuwe, algemene relatie tussen patronen en data, die het mogelijk maakt naar langere, intuïtief beter begrijpbare patronen te zoeken. Vervolgens ontwikkelen we algoritmen om op een efficiënte manier naar deze patronen te zoeken. Uit experimenten blijkt dat het resulterende algoritme efficiënter is dan andere algemene methoden, maar minder efficiënt dan methoden die voor specifiekere structuren ontwikkeld zijn.

In de volgende hoofdstukken bestuderen we daarom specifiekere methoden. Voor zover mogelijk is ons uitgangspunt daarbij de ontwikkeling van zowel theoretisch als practisch efficiënte methoden. We kijken daarom eerst naar boomstructuren. Voor de verwerking van boomstructuren zijn al efficiënte algoritmen bekend in de literatuur, zowel vanuit theoretisch als practisch oogpunt. We hebben deze algoritmen als uitgangspunt genomen voor de implementatie van een nieuw data mining algoritme, dat een zoekruimte ook theoretisch zeer efficiënt af kan zoeken. Concreet bestaat onze bijdrage uit een algoritme dat ongeordende bomen in $O(1)$ tijd per boom kan opsommen, en een incrementeel polynomiaal algoritme om de relatie tussen een patroon en de data te berekenen. Uit experimenten blijkt dat onze methode in veel gevallen vergelijkbaar presteert met andere algoritmen, maar dat ze robuuster is: het algoritme slaagt in sommige berekeningen waar andere falen.

Na boomstructuren nemen we graafstructuren onder de loep. Grafen zijn bijzonder interessant, omdat ze gebruikt kunnen worden om op een voor de hand liggende manier naar patronen in moleculaire databanken te zoeken. Eerst breiden we onze methode voor de analyse van boomstructuren uit om te zoeken naar graafpatronen. Het idee hierbij is dat we de efficiënte methoden voor boomstructuren kunnen gebruiken om een deel van de graafpatronen efficiënter te vinden. Verder blijken sommige constraints op de structuur van patronen gemakkelijker in het zoekproces op te nemen bij deze zoekstrategie. Onze methode vergelijken we vervolgens met andere graafmining algoritmen. In tegenstelling tot andere onderzoeksgroepen, proberen we deze vergelijking zo diepgaand mogelijk uit te voeren, om tot een

werkelijk inzicht te komen in de factoren die de efficiëntie van dit soort algoritmen bepalen. Uit dit onderzoek blijkt dat de opsommingsmethode die gebruikt wordt om de zoekruimte te doorlopen van ondergeschikt practisch belang is, en dat de invloed van implementatiekeuzes bijzonder groot kan zijn. Hierdoor worden de resultaten in eerdere publicaties ook in een nieuw daglicht geplaatst.

In het laatste hoofdstuk besteden we tenslotte ook aandacht aan de toepassing van graafminingmethoden voor het analyseren van moleculaire databanken. We bespreken hoe verzamelingen van molecuulfragmenten verder gereduceerd kunnen worden in de zoektocht naar fragmenten die werkelijk practisch interessant zijn.