



UNIVERSITEIT VAN AMSTERDAM

---

# Hoe krijg je je motie aangenomen

het bepalen van succes criteria voor motie classificatie

---

AFSTUDEERPROJECT BACHELOR AI

*Auteur:*

N. DAEMS  
FNWI Kunstmatige Intelligentie  
Kruislaan 404  
1098 SM Amsterdam

*Begeleider:*

Maarten MARX  
Informatics Institute  
Kruislaan 403  
1098 SJ Amsterdam

27 juni 2008

## Samenvatting

In dit onderzoek wordt gekeken of het mogelijk is om te voorspellen hoe politieke partijen zullen stemmen op een motie en of er criteria kunnen worden gevonden die de accuratesse van het voorspellen kunnen verhogen. Daarnaast wordt er gekeken of het stemgedrag van de verschillende politieke partijen inzichtelijk kan worden gemaakt. Als uitgangspunt dient een corpus van ruim 27.000 motie bestanden in XML formaat.

Gekeken wordt of op basis van de informatie die in een motie aanwezig is, het mogelijk is om te voorspellen hoe politieke partijen op een motie zullen stemmen. De motie bestanden bevatten naast de gebruikelijke tekst zoals titel, datum, indiener etc. ook trefwoorden die door de Dienst Informatievoorziening (DIV) zijn toegekend en worden gebruikt voor inhoudelijke ontsluiting van parlementaire documenten. Onderzocht wordt of de gebruikte trefwoorden kunnen bijdragen bij het classificeren van de moties. Voor het classificeren is gebruik gemaakt van het C4.5 Decision tree learner algoritme [Quinlan, 1993].

Uit het onderzoek blijkt dat voor het voorspellen van stemgedrag van politieke partijen een hoge mate van nauwkeurigheid kan worden bereikt (gemiddeld 81,85 procent / SD 2,8) en dat het gebruik van trefwoorden niet bijdraagt aan de nauwkeurigheid, maar wel meer inzicht kan geven aan de manier waarop de voorspelling tot stand is gekomen.

## Inhoudsopgave

<b>1</b>	<b>Inleiding</b>	<b>3</b>
<b>2</b>	<b>Beschrijving data</b>	<b>4</b>
<b>3</b>	<b>Methode</b>	<b>10</b>
3.1	C4.5 Decision tree algoritme . . . . .	10
3.2	Onderzoek . . . . .	12
<b>4</b>	<b>Resultaten</b>	<b>15</b>
<b>5</b>	<b>Discussie</b>	<b>15</b>
<b>6</b>	<b>Vervolgonderzoek</b>	<b>17</b>
<b>7</b>	<b>Bijlagen</b>	<b>18</b>
7.1	Voorbeeld motie XML bestand . . . . .	18
7.2	Kabinetten in de periode 1972 - 2008 . . . . .	19

# 1 Inleiding

Nu data opslag steeds voordeliger wordt zie je dat er ook steeds meer digitale informatie beschikbaar komt. Niet alleen nieuwe informatie wordt opgeslagen en gearchiveerd maar ook oude nog niet digitale informatie wordt gedigitaliseerd. Een goed voorbeeld hiervan is het Google Library Project<sup>1</sup> waarvoor complete collecties van bibliotheken worden gescand en gedigitaliseerd. Inmiddels maken de collecties van de meeste grote Amerikaanse universiteiten deel uit van dit project en volgen langzamerhand ook de Europese universiteiten.

Ook het internet kan worden gezien als een enorm digitaal archief met informatie over nagenoeg elk onderwerp dat de mensheid bezig houdt of heeft gehouden. In deze enorme verzameling van digitale informatie zit veel informatie verborgen. Het zichtbaar maken en het vinden van patronen in deze digitale hooiberg is het werkterrein van de data mining.

Bij data mining wordt geprobeerd om middels een automatisch of semi-automatisch proces patronen te vinden in de data. De patronen die in de data worden gevonden moeten van betekenis zijn en moeten kennis of inzicht geven. Met kennis of inzicht wordt hier bedoeld dat de gevonden patronen het mogelijk moeten maken om voorspellingen of uitspraken te kunnen doen over nieuwe niet eerder geziene data.

Het PoliticalMashup<sup>2</sup> programma richt zicht op een gedeelte van deze enorme berg aan een digitale informatie. Het PoliticalMashup programma gestart in maart 2008, probeert in het enorme woud van politiek gerelateerde informatie verbanden te vinden. Hiervoor worden meerdere bronnen met politieke informatie gecombineerd. Politieke informatie wordt op deze wijze inzichtelijker en ook toegankelijker gemaakt. Het PoliticalMashup programma haalt informatie uit verschillende bronnen waaronder bijvoorbeeld online media, programma's van politieke partijen, blogs, fora en digitale Kamerstukken zoals handelingen, moties en Kamervragen. Het combineren van deze verschillende bronnen maakt het mogelijk om vragen te stellen en te beantwoorden, die niet mogelijk zouden zijn wanneer de verschillende bronnen afzonderlijk zouden worden geraadpleegd.

Inmiddels staan de eerste toepassingen van het PoliticalMashup programma online. Hoewel nog in testfase, is nu al goed te zien dat het combineren van de verschillende bronnen leidt tot een schat aan data. In deze enorme databrij aan politieke informatie bevindt zich veel impliciete informatie die door het project inzichtelijk wordt gemaakt en wordt gecombineerd tot nieuwe toepassingen. Zo wordt er gewerkt aan een toepassing die Kamervragen linkt aan artikelen in de media om zo in kaart te brengen hoeveel tijd er tussen het stellen van de Kamervraag en het verschijnen van het artikel

---

<sup>1</sup><http://books.google.com/googlebooks/library.htm>

<sup>2</sup><http://www.PoliticalMashup.nl>

in de media zitten. Op deze manier kan in kaart worden gebracht hoe snel Kamerleden reageren op nieuws uit de media, of zoals door het programma zelf omschreven, de *politieke hijgerigheid* in beeld worden gebracht.

In het kader van het PoliticalMashup project is er de beschikking over en grote verzameling aan digitale Kamerstukken. Binnen deze verzameling bevindt zich ook een corpus van motie bestanden. Moties zijn uitspraken van de Tweede of Eerste kamer, die door één of twee Kamerleden of een commissie worden voorgesteld. Vaak worden moties door Tweede Kamerleden gebruikt voor het vastleggen van een conclusie van een debat of een actie punt voor een minister of staatssecretaris. Moties komen veel voor bij het bespreken van regeringsnota's en -notities in de Tweede Kamer. Moties geven dus een goed beeld van wat er zich in de Tweede Kamer afspeelt en welke onderwerpen en thema's er in de Tweede Kamer leven. Een andere belangrijke bijkomstigheid van moties is dat de verschillende politieke fracties moeten stemmen op de motie alvorens deze wordt aangenomen of afgewezen. Uit een motie is dus af te leiden hoe een politieke partij over een bepaald onderwerp denkt.

Voor het PoliticalMashup programma zou het interessant kunnen zijn om in de motie corpus verbanden te vinden, die meer zouden kunnen zeggen over de manier waarop politieke partijen op moties over bepaalde onderwerpen stemmen en hoe politieke partijen zich tot elkaar verhouden rondom deze onderwerpen/thema's. Voor het vinden van deze verbanden in het motie corpus, was het noodzakelijk om eerst het corpus in kaart te brengen. Het corpus is nog niet eerder beschreven en over de inhoud was tot op heden dan ook nog maar weinig bekend. In sectie 2 wordt dan ook eerst het corpus beschreven voor zover dit noodzakelijk is voor het onderzoek, vervolgens wordt in sectie 3 beschreven hoe het onderzoek naar het vinden van succes criteria voor het voorspellen van moties is opgezet en uitgevoerd. In sectie 4 worden de resultaten van het onderzoek besproken en in sectie 5 worden dan de gevonden resultaten bediscussieerd. Afsluitend wordt in sectie 6 kort in gegaan op mogelijk vervolgonderzoek.

## 2 Beschrijving data

De data die wordt gebruikt voor dit onderzoek komt voort uit het PoliticalMashup programma en is beschikbaar gesteld door de Dienst Informatievoorziening (DIV) van de Tweede Kamer der Staten-Generaal. De data bestaat uit een corpus van 27.946 motie bestanden, allemaal in XML formaat (Extensible Mark-up Language). De data is niet voorzien van een DTD model (Document Type Definition) of Schema, wat de validatie van de bestanden erg moeilijk, zo niet onmogelijk maakt. Naast het ontbreken van een DTD model, is de structuur van de XML bestanden niet altijd even consequent toegepast, een voorbeeld hiervan is de `< TEKST >` tag, waar-

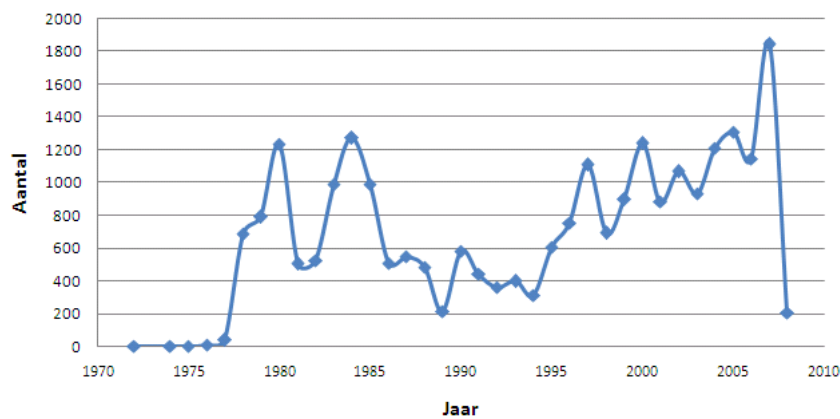
binnen de tekst van de motie staat. Soms wordt deze voorzien van nieuwe tags waarbinnen de tekst de ene keer als platte tekst en de andere keer als HTML tekst wordt opgemaakt. Een ander probleem is dat de namen van de politieke partijen en personen niet zijn genormaliseerd. Zo komen er voor de voormalige Centruumpartij twaalf verschillende spellingsvormen voor, waarvan er een aantal berusten op schrijffouten/typefouten. Ook worden er voor opsommingen in de data verschillende conventies gebruikt. Soms worden partijnamen gescheiden door komma's, dan weer door slashes, punt komma's of alleen spaties en in sommige gevallen zelfs een combinatie van deze mogelijkheden. Ook komen partijnamen soms dubbel voor in de uitslag. Dit alles maakt de extractie van de informatie uit de bestanden gecompliceerd. Een voorbeeld van een XML motie bestand is als bijlage I toegevoegd.

Van de 27.946 motie bestanden zijn er 2086 gelabeld met de status *gewijzigd*. Moties die het label *gewijzigd* dragen komen dubbel voor, en blijken altijd eerdere edities te zijn van andere motie bestanden. Daarnaast zijn er 1098 motie bestanden waarvan de uitslag niet bekend is. Voor de analyses zijn zowel de bestanden met het label *gewijzigd* als de bestanden zonder uitslag, weggelaten. Dit brengt de totale motie corpus terug tot 24.762 bestanden.

De oudste motie uit de corpus dateert van 12 juni 1972, de meest recente motie dateert van 27 maart 2008. In figuur 1 is weergegeven hoe het aantal moties is verdeeld over de tijd. Zoals te zien is in de figuur, is er van de periode 1972 tot 1977 weinig data beschikbaar, het vermoeden bestaat dan ook, dat de data voor deze periode incompleet is. Wat verder opvalt in de data zijn de twee grote pieken in 1980 en 1984 met daar tussenliggend een diep dal. De eerste piek valt samen met het kabinet van Agt I, het dal wordt gevormd door de kabinetten van Agt II en van Agt III. De kabinetten van Agt II en III hebben bij elkaar iets langer dan een jaar geduurd en worden gekenmerkt door een onrustige periode, waarin het kabinet nauwelijks aan regeren toe kwam en wat een mogelijke verklaring zou kunnen zijn voor de afname in het aantal moties. Wanneer je de verschillende kabinetten afzet tegen de data uit figuur 1 dan blijkt dat de verschillende regeerperiodes redelijk samen vallen met de pieken in de data. Een regeerperiode begint met weinig moties, in het midden van de periode wordt dan gepiekt om vervolgens aan het einde wanneer de nieuwe verkiezingen inzicht zijn weer af te nemen. Dit is ook wat je zou verwachten, immers bij aanvang van een nieuwe regering is er nog weinig rede om een motie in te dienen, gedurende de regeerperiode zullen er meer redenen zijn en nemen dus het aantal moties toe om vervolgens tegen het einde wanneer iedereen zich aan het opmaken is voor de komende verkiezingen weer af te nemen. In bijlage II is een overzicht gegeven van de verschillende kabinetsperiodes in de periode 1972 – 2008.

In de motie data wordt voor indienen van een motie gesproken over *indiener* en *mede-indiener*. Daar er in sommige bestanden wordt gesproken over meerdere indieners, is er voor gekozen om de indieners van de motie

als de initiatief nemers van de motie te bestempelen. Een politieke partij wordt dan ook alleen meegeteld als zijnde *indiener* van de motie, als zij de initiatiefnemer is van de motie en dus niet wanneer zij de *mede-indiener* is van de motie. In figuur 2 is weergegeven hoeveel moties er in de periode 1972 tot en met 2008 door een politieke partij in totaal zijn ingediend. Voor de partijen die van 1972 tot en met 2008 in de regering hebben gezeten of nog zitten, is in het rood aangegeven hoeveel van de ingediende moties zijn ingediend door de partij gedurende regeringsdeelname. Te zien is dat de PvdA in de gestelde periode de meeste moties heeft in gediend gevolgd door het CDA. Het CDA heeft daarentegen de meeste moties ingediend gedurende regeringsdeelname. De data is niet genormaliseerd, er worden alleen totalen weergegeven over de gehele periode.

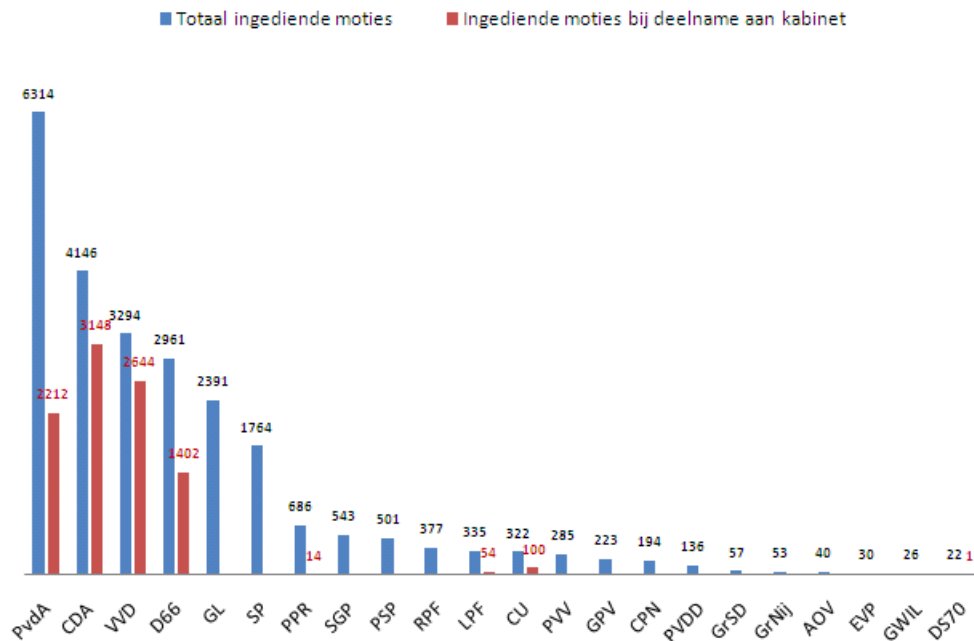


**Figuur 1:** Totaal aantal ingediende moties op jaar basis.

Voor de uitslag van een motie worden 16 verschillende categorieën toegepast. In tabel 1 is een overzicht gegeven van de verschillende uitslag categorieën en de daarbij behorende aantallen.

Zoals is te zien in bijlage I zijn de motie XML bestanden voorzien van trefwoorden. De trefwoorden worden toegekend door de Dienst Informatievoorziening en komen uit de Parlementsthesaurus<sup>3</sup>. De Parlementsthesaurus is opgezet en wordt onderhouden door de Dienst Informatievoorziening. Momenteel bestaat de thesaurus uit 10.109 termen. De thesaurus richt zich op alle voor de politiek relevante onderwerpen en omvat alle voor de Tweede Kamer relevante beleidsterreinen. Het doel van de thesaurus is het mogelijk maken van inhoudelijke ontsluiting van documenten door hieraan trefwoorden toe te kennen uit de thesaurus. Uit de verzameling van 10.109 thesaurus trefwoorden, komen 2929 trefwoorden voor in de labels van de motie

<sup>3</sup>De Parlementsthesaurus is beschikbaar gesteld aan het PoliticalMashup onderzoeksprogramma en is niet vrij toegankelijk.



**Figuur 2:** Totaal aantal moties per partij in de periode 1972 – 2008 (blauw). In rood is het totaal aantal moties bij regeringsdeelname weergegeven. Partijen met minder dan 20 ingediende moties zijn weggelaten.

bestanden. Niet alle motie XML bestanden zijn voorzien van trefwoorden. Van de 24.762 motie bestanden in de motie corpus, zijn er 24.698 voorzien van trefwoorden. Gemiddeld wordt een motie bestand beschreven door 3,75 ( $SD = 1,57$ ) trefwoorden. In tabel 2 is voor de vijfenveertig meest gebruikte trefwoorden weergegeven hoe vaak het betreffende trefwoord in de motie corpus wordt gebruikt als label. Ook is voor elk trefwoord de collocatie [Sinclair, 1996] weergegeven. De collocatie geeft aan hoe vaak een trefwoord voorkomt met verschillende andere trefwoorden. Een collocatie van 6 betekend hier dus dat een trefwoord in het moties corpus in totaal met zes andere trefwoorden voorkomt. Voor Wetgeving geldt bijvoorbeeld een count van 1458 en een collocatie van 1168. De term wetgeving komt dus in 1458 moties voor en wordt met 1168 verschillende andere trefwoorden gebruikt, voorbeelden van trefwoorden waarmee wetgeving voorkomt zijn: EG richtlijnen en Gemeente.

Stemuitslag	Count
Aangehouden	227
Aangenomen	5190
Aangenomen bij handopsteken	150
Aangenomen bij zitten en opstaan	1253
Aangenomen met algemene stemmen	3444
Aangenomen met hoofdelijke stemming	62
Aangenomen zonder stemming	36
Afgevoerd van lijst met werkzaamheden	601
Ingetrokken	2265
Niet aangenomen wegens staken van stemmen	11
Vervallen	1022
Verworpen	8381
Verworpen bij handopsteken	249
Verworpen bij zitten en opstaan	1795
Verworpen met algemene stemmen	2
Verworpen met hoofdelijke stemming	75

**Tabel 1:** Uitslag categorieën gebruikt in motie data.

Trefwoord	Count	Collocatie
wetgeving	1458	1168
subsidies	1014	804
toegepast onderzoek	831	1001
kostenbeheersing	777	635
rijksbegrotingen	761	743
eu	652	584
gemeenten	646	613
regeringsbeleid	621	683
regeringsnota's	619	763
jeugdigen	611	541
pkb's	596	298
overheidsuitgaven	582	598
milieubeheer	577	529
toelating	577	412
prijzen	508	506
infrastructuur	490	334
spoorlijnen	480	197
openbaar vervoer	476	273
werkgelegenheid	462	172
wegen	454	209
ondernemingen	428	518
ontwikkelingssamenwerking	407	317
gezondheidszorg	404	423
verdragen	393	455
werkloosheidsbestrijding	392	337
vluchtelingen	389	307
natuurbehoud	378	305
onderwijs	371	379
structurele ontwikkeling	370	447
ruimtelijke ordening	370	304
eg	362	405
vreemdelingen	354	285
commissies van advies	353	519
minderheden	352	334
regionale ontwikkeling	343	314
beleidsonderzoek	335	511
gehandicapten	331	283
besteedbaar inkomen	317	233
vrouwen	315	359
experimenten	311	412
uitwijzingen	305	234
decentralisatie	302	305
economische ontwikkeling	298	298
provincies	285	327

**Tabel 2:** Overzicht van de 45 meest gebruikt thesaurus termen die voor het labelen van de moties worden gebruikt.

### 3 Methode

Het doel van dit onderzoek is het vinden van succes criteria voor de classificatie van motie bestanden, en om dit classificatie proces inzichtelijk te maken. Met inzichtelijk maken wordt hier bedoeld dat de uitkomst van het classificatie proces niet het doel op zich is, minstens even belangrijk is het om inzicht te krijgen in hoe de classificatie tot stand is gekomen. Deze aanname sluit *blackbox* classificatie algoritmes zoals bijvoorbeeld *neurale netwerken* en *multi-dimensionale cluster algoritmes* uit. Het classificatie algoritme moet dus naast het vinden van een voorspelling (*prediction*) ook instaat zijn om de voorspelling op een gestructureerde manier weer te geven zodat over de voorspelling kan worden geredeneerd of zodat de voorspelling kan dienen als basis voor het doen van voorspellingen over nieuwe data. Hieronder is een voorbeeld gegeven van hoe zo'n voorspelling eruit zou kunnen zien.

```
If indiener = GL and trefwoord = Milieu
    then CU vote = true;

If indiener = D66 and trefwoord = euthanasie
    then CU vote = false
Otherwise,
if indiener = CDA and trefwoord = euthanasie
    then CU vote = true
```

In het voorgaande voorbeeld kun je zien dat als GroenLinks de indiener is van een motie en als de motie is voorzien van het trefwoord Milieu, dat de ChristenUnie voor de motie zal stemmen. Wanneer echter D66 de indiener is van de motie en de motie is voorzien van het trefwoord euthanasie, dan zal de ChristenUnie tegen stemmen. Dit in tegenstelling tot een motie die is voorzien van hetzelfde trefwoord euthanasie, maar ingediend door het CDA. Deze gestructureerde vorm van classificatie geeft veel meer informatie, dan alleen het toewijzen van klassen. De structurele beschrijving hoeft echter niet perse op regels te zijn gebaseerd, een andere mogelijkheid voor het weergeven van de classificatie structuur is door gebruik te maken van *Decision trees* (beslisbomen). Hoewel het mogelijk is, om uit een Decision tree een verzameling regels te extraheren die onafhankelijk van elkaar zijn, levert dit over het algemeen onnodig complexe regels op [Witten and Frank, 2005]. Voor dit onderzoek is er voor gekozen om als classificatie algoritme gebruik te maken van het C4.5 Decision tree algoritme van Quinlan [Quinlan, 1993].

#### 3.1 C4.5 Decision tree algoritme

Het C4.5 Decision tree algoritme kan worden gezien als de standaard voor Decision tree algoritmes. Het is het meest gebruikte en beschreven algoritme voor Decision trees [Deshpande and Karypis, 2002]. Het algoritme is

ontwikkeld door Quinlan [Quinlan, 1993] en is een verbetering van het ID3 algoritme [Quinlan, 1986] dat eerder ook door Quinlan is ontwikkeld. Als input heeft het algoritme een verzameling van voorbeelden (training set) nodig. De voorbeelden, ook wel *instances* genoemd, bestaan uit een verzameling van attributen die vooraf zijn bepaald/gekozen. Om een concept uit voorbeelden te kunnen leren, is het dus noodzakelijk dat een voorbeeld kan worden uitgeschreven als een lijst van attributen. Het C4.5 algoritme maakt gebruik van het concept *information gain* uit de informatie theorie [Shannon, 1948] om een attribuut te selecteren uit de voorbeelden, waarop de splitsing in tree zal gaan plaats vinden. Hiervoor kijkt het algoritme voor elk attribuut wat de hoogte is van de *information gain* en selecteert vervolgens het attribuut met de hoogste waarde. De information gain kan worden berekend met behulp van formule 1 waarmee de information waarde voor een attribuut in bits kan worden berekend.

$$I(P(v_1), \dots, P(v_n)) = \sum_{i=1}^n -P(v_i) \log_2 P(v_i) \quad (1)$$

Wanneer een attribuut is geselecteerd, dan worden de voorbeelden gesplit op het geselecteerde attribuut en het proces herhaalt zich vervolgens op de subsets die zijn ontstaan na de split. Het hele proces gaat net zolang door totdat verdere opsplitsing niet meer mogelijk is of niet meer bijdraagt aan een betere uitkomst (information gain = 0). Een dergelijk eindpunt wordt ook wel een *leaf* genoemd. In pseudo code ziet het C4.5 algoritme er als volgt uit:

```

For each attribute a
  Find the normalized information gain
  from splitting on a
Let a_best be the attribute with the highest
normalized information gain

Create a decision node node that splits on a_best
recur on the sublists obtained by splitting on
a_best and add those nodes as children of node

```

Een Decision tree die na training alle voorbeelden goed kan classificeren is niet altijd beter dan een kleinere Decision tree die na training niet alle voorbeelden goed kan classificeren. Dit fenomeen wordt *overfitting* genoemd en is het gevolg van een te specifiek model. Overfitting resulteert dus in een model dat niet goed kan generaliseren en dus niet goed met nieuwe niet eerder geziene data kan omgaan. C4.5 maakt gebruik van *pruning* om overfitting tegen te gaan, dit is dan ook een van de verschillen met zijn voorganger ID3. Het pruning mechanisme is gebaseerd op een heuristische

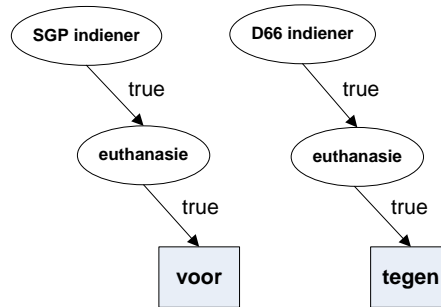
methode en gebruikt de trainingsdata voor het schatten van de fout (error rate). Hoewel de statistische onderbouwing zwak is, blijkt in de praktijk dat het algoritme met pruning tot betere resultaten leidt dan wanneer pruning wordt weggelaten [Witten and Frank, 2005]. Het pruning algoritme van C4.5 maakt een schatting van de error rate van elke subtree waaruit de volledige tree bestaat. Hiervoor wordt voor elke node in de tree de error rate geschat door aan te nemen dat de „echte” klasse van de node de klasse is die het meest is vertegenwoordigd in de node, beginnend onderaan de boom. Wanneer nu blijkt dat de schatting aangeeft dat de tree accurater (lagere error rate) zal zijn als de child nodes van node  $n$  worden weggelaten, dan zal in de uiteindelijk Decision tree node  $n$  worden weggelaten en worden vervangen door een leaf. Dit leidt tot compactere trees die beter kunnen generaliseren.

### 3.2 Onderzoek

Voor het voorspellen of een motie al dan niet wordt aangenomen is gekeken of het mogelijk is om te voorspellen hoe een politieke partij  $X$  zal stemmen op een motie  $Y$ . Zoals in sectie 2 is aangegeven zijn alle moties voorzien van een uitslag. Naast de uitslag is ook in elke motie aangegeven, hoe de verschillende fracties op de motie hebben gestemd. Een fractie kan voor, tegen of verdeeld hebben gestemd. In het laatste geval bestaat er dus onenigheid binnen de fractie, immers de fractie heeft niet unaniem gestemd. Omdat er in de stemmingsuitslag geen zetel aantallen staan vermeld en een motie pas wordt aangenomen bij een meerderheid van stemmen, is ervoor gekozen om niet te voorspellen of een motie wordt aangenomen en of hiervoor succes criteria uit de motie corpus kunnen worden geëxtraheerd, maar of het stemgedrag van de verschillende politieke partijen kan worden voorspeld en of hiervoor succes criteria kunnen worden gevonden in het motie corpus.

Om meer inzicht te krijgen in het stemgedrag van de verschillende politieke partijen is gekeken of de trefwoorden die aan motie bestanden zijn toegevoegd door de Dienst informatievoorziening, kunnen bijdragen aan een meer beschrijvende Decision tree. Het achterliggende idee hierbij is dat de trefwoorden een inhoudelijke weergave geven van de inhoud van de motie. Een motie die voorzien is van de term *euthanasie* en *wetgeving* zal dus over deze twee termen gaan. Wanneer echter alleen zou worden gekeken naar het stemgedrag van een politieke partij rondom moties met de twee eerder genoemde termen, dan zal een Decision tree algoritme tegen problemen aanlopen wanneer de moties betreffende deze twee termen mooi zouden zijn verdeeld in twee subsets, waarvan de eerste subset bestaat uit moties *pro euthanasie* en tweede subset bestaande uit moties *contra euthanasie*. Immers de trefwoorden die gebruikt zijn zeggen wel iets over de inhoud van de motie, maar niet over de context, in dit geval dus *pro* of *contra* euthanasie. Door nu naast trefwoorden ook te kijken naar welke politieke partijen de in-

dieners zijn van de motie en welke politieke partijen mede-indieners zijn van de motie zou een Decision tree algoritme wel een scheiding moeten kunnen aanbrengen tussen de twee subsets. In figuur 3 is dit grafisch weergegeven.



**Figuur 3:** Stukje van een Decision tree waarin attribuut indiener een scheiding tussen subsets mogelijk maakt.

In de motie XML files zit veel informatie die zou kunnen worden gebruikt voor het classificeren van de motie data. Zo is elke motie XML file voorzien van een titel, een citeer titel, vergaderjaar, status, tekst van de motie en nog veel meer (meta-)data. Voor dit onderzoek is uit de mogelijke attributen een selectie gemaakt. Uiteindelijk zijn er zes variabelen meegenomen die zijn vertaald naar attributen. Als eerste is er gekeken of een partij op het moment van de motie al dan niet in de regering zit. Dit is vertaald naar een attribuut  $\text{inRegering} = \{\text{true}, \text{false}\}$  waarbij true staat voor in de regering en false voor niet in de regering. De keuze voor het attribuut regering berust op het idee dat wanneer een partij in de regering zit dit een goede voorspeller is voor een *tegen stem*, immers moties zijn vaak tegen regeringsbeleid. Als tweede variabele is ervoor gekozen om te kijken of een partij de indiener is van de motie, dit zou een goede voorspeller moeten zijn voor een *voor stem*. Dit is vertaald naar een attribuut  $\text{isIndiener} = \{\text{true}, \text{false}\}$ . Vervolgens is gekeken of de partij mede-indiener is van de motie, dit is vertaald naar een attribuut  $\text{isMedeIndiener} = \{\text{true}, \text{false}\}$ . Als vierde variabele is de datum van de motie meegenomen. Als vijfde variabele is gekeken of bepaalde trefwoorden in de motie voorkomen. Hiervoor is een selectie gemaakt van de 45 meest voorkomende trefwoorden in het motie corpus. Een overzicht van de trefwoorden is te vinden in tabel 2. Voor elk trefwoord is een attribuut aangemaakt met daaraan gekoppeld een boolean value die codeert voor het al dan niet aanwezig zijn van het trefwoord. In totaal zijn er dus 45 attributen voor de trefwoorden. Op dezelfde manier is voor elke motie gecodeerd welke andere politieke partijen indiener of mede-indiener zijn van een motie. Als laatste is de classificatie variabele vertaald naar een attribuut. Partijen kunnen voor, tegen of verdeeld stemmen op een motie, dit is vertaald naar het attribuut  $\text{Vote} = \{\text{voor}, \text{tegen}, \text{verdeeld}\}$ . In listing 1 is weergegeven

**Tabel 3:** Verdeling training sets voor partijen

<b>Partij</b>	<b>Totaal</b>	<b>Stem</b>		
		<i>Voor</i>	<i>Tegen</i>	<i>Verdeeld</i>
CDA	20650	10240	10198	212
CU	7143	2992	4045	6
VVD	20679	11315	9200	164
PvdA	20643	6787	13671	185
D66	20655	6857	13648	150
LPF	4367	2323	2021	23

hoe dit er in ARFF (Attribute Relational File Format) notatie uitziet.

**Listing 1:** Voorbeeld van een ARFF bestand

```
@relation
@attribute inRegering {false,true}
@attribute isIndiener {false,true}
@attribute vote {tegen,voor,verdeeld}
@attribute datum date[yyyy]
#codeert voor mede-indieners
@attribute PvdA {false,true}
...
...
#codeert voor aanwezigheid trefwoord
@attribute BELEIDSONDERZOEK {false,true}
...
...
@data
true, false, tegen, 2000, false .....
```

Het onderzoek heeft zich in eerste instantie gericht op alle politieke partijen die vanaf 1972 tot 2008 in de regering hebben gezeten of nog zitten. Voor de partijen: KVP, ARP, CHU, DS'70 en PPR waren er te weinig moties om zinvol onderzoek te doen. Het onderzoek heeft zich dan ook uiteindelijk geconcentreerd op de partijen: CDA, CU, VVD, PvdA, D66 en LPF. In tabel 3 is weergegeven hoe de training sets voor de verschillende partijen zijn verdeeld. Bij het testen zijn de training sets in tien random subsets verdeeld, vervolgens is steeds een voor een, een set gekozen als test set en de overige negen als training set (10 times cross validation), het uiteindelijke resultaat is dan het gemiddelde over de tien trials. De resultaten zijn weergegeven in sectie 4.

## 4 Resultaten

In tabel 4 zijn de resultaten weergegeven voor het trainen met het C4.5 algoritme op de verschillende training sets uit tabel 3. In de linker kolom staat de naam van de partij waarover de training set gaat, het getal achter de naam van de partij geeft aan hoeveel verschillende trefwoorden er zijn gebruikt in de training set, dit kan variëren van 0, 20 tot 45 trefwoorden. Voor elke partij is ook een analyse uitgevoerd waarbij het classificeren alleen is gedaan op basis van het feit of een partij indiener of mede-indiener is van een motie. Dit is aangeven met een  $X$ . De  $N$  staat voor het totaal aantal items in de training set.  $Attr$  staat voor het aantal attributen dat is gebruikt.  $Leaves$  geeft het aantal leaves aan van de geleerde tree en  $Size$  de grote van de tree.  $Goed$  is het aantal goed geclassificeerde moties en  $fout$  het aantal fout geclassificeerde moties.  $Kappa$   $\kappa$  staat voor Cohen's  $\kappa$  coëfficiënt.

Uit de resultaten in tabel 4 valt af te lezen dat voor CDA en VVD de attributen `isIndiener` en `isMedeindiener` de belangrijkste voorspellers zijn. Het toevoegen van nieuwe attributen levert nauwelijks winst op. Voor het CDA geldt een stijging van 2,3% en voor de VVD geldt slechts een stijging van 1,2%. Het lijkt erop dat het CDA en de VVD alleen voor stemmen op een motie wanneer ze zelf indiener en of mede-indiener zijn van de motie. Voor de andere politieke partijen blijkt er echter wel winst te behalen door meer attributen te gebruiken. Uit de resultaten blijkt dat wanneer er rekening wordt gehouden met wie de indieners of mede-indieners zijn van de motie dat deze informatie sterk kan bijdragen aan het voorspellen van het stemgedrag. Voor de PvdA geldt een toename van 6%, voor de LPF een toename van 12,7%, voor D66 een toename van 14% en voor de ChristenUnie zelfs een toename van 17,8%. Verder blijkt uit de resultaten dat het toevoegen van de trefwoorden niet leidt tot betere resultaten voor het classificeren.

## 5 Discussie

Zoals uit de resultaten blijkt kan informatie over wie de indieners of mede-indieners van moties zijn, van waarde zijn voor het voorspellen van hoe een partij op een motie zal gaan stemmen. Ook blijkt dat de trefwoorden niet bijdragen aan een betere voorspeller. Wanneer we echter naar de structuren kijken van de Decision trees, dan blijkt dat de trefwoorden wel degelijk bijdragen aan de leesbaarheid van de trees en de trees van context voorzien. In listing 2 is een stukje uit de Decision tree gegeven voor de training set van de ChristenUnie. Hieruit blijkt dat de ChristenUnie wanneer zij zelf niet de indiener en of mede-indiener zijn van een motie, zij plichtsgetrouw het CDA volgen, echter wanneer het een motie betreft met het trefwoord `Vluchtelingen` dan volgt de ChristenUnie niet langer het CDA maar D66.

**Tabel 4:** Resultaten trainen en classificeren met C4.5 op motie data

Partij	N	Attr	Leaves	Size	Goed	%Goed	Fout	%Fout	$\kappa$ Kappa
CDA 0	20650	51	26	51	17372	84,1259	3278	15,8741	0,6857
CDA2 0	20650	71	81	161	17418	84,3487	3232	15,6513	0,6901
CDA 45	20650	96	123	245	17358	84,0581	3292	15,9419	0,6844
CDA X	20650	3	3	5	16899	81,8354	3751	18,1646	0,6401
CU 0	7143	51	26	51	5523	77,3205	1620	22,6795	0,5402
CU2 0	7143	71	57	113	5458	76,4105	1685	23,5895	0,5231
CU 45	7143	96	81	161	5466	76,5225	1677	23,4775	0,5239
CU X	7143	3	3	5	4249	59,4848	2894	40,5152	0,257
D66 0	20655	51	45	89	16530	80,029	4125	19,971	0,5592
D66 20	20655	71	132	263	16445	79,6175	4210	20,3825	0,5484
D66 45	20655	96	191	381	16486	79,816	4169	20,184	0,5498
D66 X	20655	3	1	1	13648	66,076	7007	33,924	0
Lpf 0	4367	51	12	23	3561	81,5434	806	18,4566	0,6334
Lpf 20	4367	71	20	39	3559	81,4976	808	18,5024	0,6327
Lpf 45	4367	96	36	71	3558	81,4747	809	18,5253	0,6321
Lpf X	4367	3	3	5	3007	68,8573	1360	31,1427	0,4024
PvdA 0	20643	51	35	69	17204	83,3406	3439	16,6594	0,6205
PvdA 20	20643	71	82	163	17196	83,3018	3447	16,6982	0,628
PvdA 45	20643	96	153	305	17203	83,3358	3440	16,664	0,6317
PvdA X	20643	3	3	5	15960	77,3143	4683	22,6857	0,5648
VVD 0	20679	51	23	45	17517	84,7091	3162	15,2909	0,686
VVD 20	20679	71	69	137	17517	84,7091	3162	15,2909	0,6871
VVD 45	20679	96	113	225	17488	84,5689	3191	15,4311	0,6846
VVD X	20679	3	3	5	17272	83,5243	3407	16,4757	0,6588

**Listing 2:** Stukje uit Decision tree ChristenUnie

```

CU medeindiener = false
|
|   CU isindiener = false
|   |
|   |   CDA = false
|   |   |
|   |   |   inregering = false
|   |   |   |
|   |   |   |   PvdA = false
|   |   |   |   |
|   |   |   |   |   SGP = false: tegen (1097.0/103.0)
|   |   |   |   |   SGP = true
|   |   |   |   |   |
|   |   |   |   |   |   JEUGDIGEN = false
|   |   |   |   |   |   |
|   |   |   |   |   |   |   PVV = false: tegen (25.0/7.0)
|   |   |   |   |   |   |   PVV = true: voor (2.0)
|   |   |   |   |   |   |   JEUGDIGEN = true: voor (3.0)
|   |   |   |   |   PvdA = true
|   |   |   |   |   |
|   |   |   |   |   |   SGP = false
|   |   |   |   |   |   |
|   |   |   |   |   |   |   GL = true: tegen (32.0/11.0)
|   |   |   |   |   |   |   GL = false
|   |   |   |   |   |   |   |
|   |   |   |   |   |   |   |   NATUURBEHOUD = false
|   |   |   |   |   |   |   |   |
|   |   |   |   |   |   |   |   |   VROUWEN = false: voor (45.0/14.0)
|   |   |   |   |   |   |   |   |   VROUWEN = true: tegen (2.0)
|   |   |   |   |   |   |   |   |   NATUURBEHOUD = true: tegen (2.0)
|   |   |   |   |   |   |   SGP = true: voor (3.0)
|   |   |   |   |   |   VLUCHTELINGEN = true: voor (15.0)
|   |   |   CDA = true
|   |   |   |
|   |   |   |   VLUCHTELINGEN = false: voor (1527.0/244.0)
|   |   |   |   VLUCHTELINGEN = true
|   |   |   |   |
|   |   |   |   |   D66 = false: tegen (7.0/2.0)
|   |   |   |   |   D66 = true: voor (5.0)
|   |   |   CU isindiener = true: voor (259.0/3.0)
CU medeindiener = true: voor (912.0/5.0)

```

Voor dit onderzoek zijn de trefwoorden geselecteerd op basis van hun voorkomen. De 45 meest voorkomende trefwoorden zijn gebruikt voor het trainen. Het zou echter heel goed mogelijk kunnen zijn dat er betere trefwoorden kunnen worden geselecteerd, die wel tot een betere classificatie leiden. Een mogelijk andere verklaring voor het niet bijdragen van de trefwoorden zou kunnen zijn de manier waarop ze aan de moties worden toege-

kend. Trefwoorden worden handmatig toegekend. Zo bevinden zich moties in het corpus die over hetzelfde onderwerp gaan, maar van verschillende trefwoorden zijn voorzien. Een voorbeeld zijn de moties die gaan over de afschaffing van accijnzen op frisdrank. Sinds 1993 mogen er niet langer accijns op frisdranken worden geheven. De oplossing van de toenmalige regering was om accijnzen op frisdranken niet langer accijnzen te noemen maar verbruiksbelasting. Moties die hierover gaan worden op drie verschillende manieren voorzien van trefwoorden. Soms worden moties voorzien de trefwoorden *Frisdrank*, *Accijns* soms alleen van het trefwoord *Accijns op frisdranken* en soms van de trefwoorden *Frisdranken*, *Verbruiksbelastingen*. Dit zijn dus drie verschillende manieren van taggen voor hetzelfde type van moties. Of dit dit probleem bij meer trefwoorden speelt is is niet verder in kaart gebracht.

## 6 Vervolgonderzoek

Uit de gevonden resultaten blijkt dat het goed mogelijk is om te voorspellen hoe een politieke partij op een motie zal gaan stemmen. Voor vervolg onderzoek zou het interessant kunnen zijn om te kijken of het mogelijk is om de individuele voorspellingen van de partijen zijn te combineren zodat kan worden voorspeld of een motie al dan niet zal worden aangenomen. Ook zou het interessant kunnen zijn om te kijken of het mogelijk is om op basis van de inhoud van de moties de juiste trefwoorden zijn te voorspellen, dit kan ook inzicht geven in de kwaliteit van de manier waarop de moties zij getagd of zelfs het begin zijn van een automatisch tagging mechanisme. Als laatste zou het interessant kunnen zijn om niet de trefwoorden van de moties te gebruiken, maar de teksten van de moties. Dit geeft een legio aan mogelijkheden voor linguïstische en statistische technieken.

## 7 Bijlagen

### 7.1 Voorbeeld motie XML bestand

```
<?xml version="1.0" encoding="ISO-8859-1" ?>
<document id="6289668" template="Doc_Kamerstukken_Moties"
  templateID="333" timestamp="03-04-2008_09:56:06">
<hiddendatum>2008.02.14</hiddendatum>
<hiddentitel>| Informatie- en communicatietechnologie (ICT)|</hiddentitel>
<hiddenkamerstuknr>|26643|</hiddenkamerstuknr>
<hiddenkamerstukvolgnr>|0115|</hiddenkamerstukvolgnr>
<hiddenekltr/>
<hiddenhoofdstuk/>
<hiddenraad/>
<status><alt omschrijving="status">Definitief</alt></status>
<kamerstukmoties><alt omschrijving="kamer_(kamerstukken_moties)">
  Tweede</alt></kamerstukmoties>
<kamerstuklodytem><kamerstuknummer>26643</kamerstuknummer>
<kamerstukvolgnummer>0115</kamerstukvolgnummer>
<ek/>
<hoofdstuk/>
<raad/>
<herdruk/>
<wijziging/>
<link/>
</item>
</kamerstuklodytem>
<rijkswet></rijkswet>
<titel>
  <item>Informatie- en communicatietechnologie (ICT)</item>
</titel>
<citeertitel>
  Meer eisen ten aanzien van open standaarden oplossingen
</citeertitel>
<hiddensoort>Motie</hiddensoort>
<soort/>
<datum>2008.02.14</datum>
<indienerlodytem><item><naam>M.L. Vos</naam>
<partij>PvdA</partij>
<departement>Lid Tweede Kamer</departement>
</item>
</indienerlodytem>
<medeindienerlodytem></medeindienerlodytem>
<bronlodytem><item><datum>2008.02.14</datum>
<actie>
<alt omschrijving="bron_actie_(kamerstukken_moties)">Ingediend</alt>
</actie><soortbronmoties>
<alt omschrijving="bron_soort_(kamerstukken_moties)"></alt>
</soortbronmoties>
<vergaderjaar>2007 - 2008</vergaderjaar>
<editie>054</editie>
<nummer/>
<startpagina>3905</startpagina>
<eindpagina/>
</item>
<item><datum>2008.02.14</datum>
<actie><alt omschrijving="bron_actie_(kamerstukken_moties)">Stemming</alt>
</actie><soortbronmoties><alt omschrijving="bron_soort_(kamerstukken_moties)">
  Handelingen II</alt></soortbronmoties>
<vergaderjaar>2007 - 2008</vergaderjaar>
<editie>054</editie>
<nummer/>
<startpagina>3949</startpagina>
<eindpagina/>
</item>
</bronlodytem>
<stemminglodytem><item><datum>2008.02.14</datum>
<uitslag><alt omschrijving="stemming_uitslag">Aangenomen</alt></uitslag>
</item>
</stemminglodytem>
<stemverdelinglodytem><item><stem>voor</stem>
<fractie>CDA, CU, D66, GL, LVER, PvdA, PVDD, SGP, SP, VVD</fractie>
</item>
<item><stem>tegen</stem>
<fractie>PVV</fractie>
</item>
</stemverdelinglodytem>
<tekstxml><p>De Kamer,</p>
<p>gehoord de beraadslaging,</p>
<p>overwegende, dat bij de selectie van de leverancier van GOUD 1.0 de
voornemens zoals verwoord in de nota Nederland Open in Verbinding
nadrukkelijker kunnen worden geborgd door meer knock-outeisen over open
```

standaarden te stellen;

verzoekt de regering om in de lijst van eisen en wensen waaraan de bieding moet voldoen, meer wensen ten aanzien van open standaarden oplossingen om te zetten naar eisen;

en gaat over tot de orde van de dag.

```

</tekstxml>
<trefwoordlod>
<item>
  <trefwoord>COLLECTIEVE SECTOR</trefwoord>
</item>
<item>
  <trefwoord>SOFTWARE</trefwoord>
</item>
</trefwoordlod>
<relatielod></relatielod>
<extrainfo/>
<aantekening/>
<starrecordnummer/>
<informatiedossierlod></informatiedossierlod>
</document>

```

## 7.2 Kabinetten in de periode 1972 - 2008

Kabinetten in de periode 1972 - 2008					
Naam	Partijen	Van	Tot	Dagen	
Biesheuvel I	KVP, VVD, ARP, CHU, DS'70	6 juli 1971	9 augustus 1972	400	
Biesheuvel II	KVP, VVD, ARP, CHU	9 augustus 1972	11 mei 1973	275	
Den Uyl	PvdA, KVP, ARP, PPR, D66	11 mei 1973	19 december 1977	1683	
Van Agt I	CDA, VVD	19 december 1977	11 september 1981	1362	
Van Agt II	CDA, PvdA, D66	11 september 1981	29 mei 1982	260	
Van Agt III	CDA, D66	29 mei 1982	4 november 1982	159	
Lubbers I	CDA, VVD	4 november 1982	14 juli 1986	1348	
Lubbers II	CDA, VVD	14 juli 1986	7 november 1989	1212	
Lubbers III	CDA, PvdA	7 november 1989	22 augustus 1994	1749	
Kok I (Paars I)	PvdA, VVD, D66	22 augustus 1994	3 augustus 1998	1442	
Kok II (Paars II)	PvdA, VVD, D66	3 augustus 1998	22 juli 2002	1449	
Balkenende I	CDA, LPF, VVD	22 juli 2002	27 mei 2003	309	
Balkenende II	CDA, VVD, D66	27 mei 2003	7 juli 2006	1137	
Balkenende III	CDA, VVD	7 juli 2006	22 februari 2007	230	
Balkenende IV	CDA, PvdA, CU	22 februari 2007			

## Referenties

- [Deshpande and Karypis, 2002] Deshpande, M. and Karypis, G. (2002). Using conjunction of attribute values for classification.
- [Quinlan, 1986] Quinlan, J. R. (1986). Induction of decision trees. *Machine Learning*, 1:81–106.
- [Quinlan, 1993] Quinlan, J. R. (1993). *C4.5: Programs for machine learning*. Morgan Kaufmann.
- [Shannon, 1948] Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal*, 27:379–423.
- [Sinclair, 1996] Sinclair, J. (1996). The search for units of meaning. *Textus*, IX:75–106.
- [Witten and Frank, 2005] Witten, I. H. and Frank, E. (2005). *Data Mining: practical machine learning tools and techniques*. Morgan Kaufmann, San Francisco, 2nd edition.